Image and Vision Computing 159 (2025) 105576

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

LCNet: Lightweight real-time image classification network based on efficient multipath dynamic attention mechanism and dynamic threshold convolution

Xiaoxia Yang [a], Zhishuai Zheng [b], Huanqi Zheng [c], Zhedong Ge [b], Xiaotong Liu [b], Bei Zhang [d], Jinyang Lv [e,*]

[a] School of New Generation Information Technology, Shandong Polytechnic, Jinan, 250104, Shandong, China
[b] School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan, 250101, Shandong, China
[c] Shandong Institute for Product Quality Inspection, Jinan, 250102, Shandong, China
[d] Wood Value promotion and sustainable Development Center, Haidian, 100036, Beijing, China
[e] College Computer Application, Guilin University of Technology at Nanning, Nanning, 532100, Guangxi, China

ARTICLE INFO

Keywords:
Visual classification
Multipath dynamic attention mechanism
Dynamic threshold convolution
Star operation
Multipath architecture

ABSTRACT

Hybrid architectures that integrate convolutional neural networks (CNNs) with Transformers can comprehensively extract both local and global image features, exhibiting impressive performance in image classification. However, their large parameter sizes and high computational demands hinder deployment on low-resource devices. To address this limitation, we propose a dual-branch classification network based on a pyramid architecture, termed LCNet. First, we introduce a dynamic threshold convolution module that adaptively adjusts convolutional parameters based on the input, thereby improving the efficiency of feature extraction. Second, we design a multi-path dynamic attention mechanism that optimizes attention weights to capture salient information and enhance the significance of key features. Third, a star-shaped connection is adopted to enable efficient information fusion between the two branches in a high-dimensional implicit feature space. LCNet is evaluated on four public datasets and one wood dataset (Tiny-ImageNet, Mini-ImageNet, CIFAR100, CIFAR10, and Micro-CT) using recognition accuracy and inference efficiency as metrics. The results show that LCNet achieves a maximum accuracy of 99.50% with an inference time of only 0.0072 s per image, outperforming other state-of-the-art (SOTA) models. Extensive experiments demonstrate that LCNet is more competitive than existing neural networks and can be effectively deployed on low-performance computing devices. This broadens the applicability of image classification techniques, aligns with the trend of edge computing, reduces reliance on cloud servers, and enhances both real-time processing and data privacy.

1. Introduction

Image classification aims to lay the foundation for subsequent statistical analysis and pattern recognition by identifying image labels and plays an essential role in vast fields, such as medical image analysis, autonomous driving, facial recognition, and video surveillance [1–3]. Since the advent of Convolutional Neural Networks (CNNs), existing image classification models have always been improved based on them. However, the limited receptive field of CNN-based models constrains their ability to learn long-distance dependencies in images, which affects the model′s extraction of comprehensive shapes, textures, and high-level semantic information of images. This deficiency significantly limits their application in dense and high-precision tasks, especially in traffic monitoring image recognition and CT image analysis. To address this issue, recent methods have been proposed to use large convolutional kernels, dilated convolutions, and feature pyramids to expand the receptive field [4–6]. Another approach involves integrating attention mechanisms into the CNN architecture, a concept derived from the field of Natural Language Processing (NLP), which aims to model the global interaction of all pixels in feature maps, although this increases memory consumption and computational overhead [7,8]. At the same time, researchers have also tried to use full convolution and explored models that only use attention mechanisms, such as Transformers. This model is specifically designed for modeling long-distance dependencies from sequence to sequence and capturing relationships at any position.

Unlike previous CNN-based image recognition approaches, Transformers not only perform well in global contexts modeling but also achieve satisfactory results in downstream tasks after extensive pre-training [9–12]. Initially employed for image recognition, the performance of Vision Transformers (ViT) is comparable to CNNs [13]. ViT conducts image recognition by taking two-dimensional image blocks

* Corresponding author.
E-mail address: lvjinyang@glut.edu.cn (J. Lv).

https://doi.org/10.1016/j.imavis.2025.105576
Received 9 January 2025; Received in revised form 10 April 2025; Accepted 29 April 2025
Available online 13 May 2025

embedded with positional encodings as inputs. However, compared with CNNs, ViT has several disadvantages: firstly, it has a larger parameter count which leads to suboptimal training performance on smaller datasets, necessitating to be pre-trained on large datasets; secondly, its computation process is more complex, requiring significant memory resources. Additionally, when Transformers are applied to image classification, the two-dimensional images are segmented and input as one-dimensional sequences into the model, thus breaking the connections between local structures and focusing only on the global context at all stages. Consequently, only relying on Transformers can lead to significant loss of local information and lower classification accuracy due to insufficient training of parameters on small and medium-sized datasets.

In addition, Transformers have shortcomings in capturing fine-grained spatial details. On the contrary, CNNs can effectively extract low-level visual features to compensate for these deficiencies in detail. Based on this, some approaches of fusing CNNs with Transformers have emerged as promising alternatives for image classification. For instance, some models employ CNNs in shallow layers and Transformers in deeper layers [14–16], while others integrate CNNs and Transformers in their architecture modules [17–19]. Moreover, such hybrid models have been deployed in various practical scenarios, such as in medical CT image classification, where VTCNet [20], MSFNet-2SE [21], and PMANet [22] have all achieved satisfactory results. In the field of wood microscopic image recognition, models such as WoodGLNet [23] and TimberIDNet [24] have achieved recognition accuracy of exceeding 99%. These studies demonstrate that CNN-Transformer hybrid models effectively combine the strengths of both, achieving significant success in various vision tasks.

Motivated by the above analysis, this paper introduces a lightweight real-time image recognition model based on CNN and Transformer, called LCNet, which is composed of two key components: Multi-path Dynamic Attention Mechanism (MDAM) and Dynamic Threshold Convolution (DTConv). Specifically, the MDAM extracts multi-scale global features from images, while DTConv dynamically adjusts convolution parameters based on input data, swiftly and accurately capturing local features of images. As illustrated in Fig. 1, the x-axis represents the computational complexity (in GFLOPs, G). The y-axis denotes the recognition accuracy achieved on the CIFAR-10 benchmark dataset. Each colored circle corresponds to a different classification model or its variant, with the diameter of the circle indicating the number of parameters. A gray reference circle in the bottom-right corner indicates the scale for parameter size (in millions, M). This visualization enables a more intuitive comparison of performance across different models. LCNet significantly outperforms other SOTA vision Transformers and convolutional networks in various datasets. LCNet achieves a balance between model performance, size, and inference speed, enhancing efficiency without sacrificing capability. Notably, when applied to five different datasets, LCNet's Top-1 accuracy is 62.75%, 66.62%, 70.24%, 95.82%, and 99.75%, respectively.

The principal contributions of this study are as follows:

(1) We propose an efficient dual-path fusion network, which consists of the DTConv module to extract local image information and the MDAM module to capture global image information, and the efficient fusion of information flows of the two modules is realized by a star operation connection method. DTConv dynamically adjusts its convolution kernels for fusion processing according to the local features of the input image, improves the efficiency of convolution parameter update to effectively extract the local image features. MDAM employs dynamic modeling and attention correction strategies to select k tokens based on overall image correlations, reducing image redundancy, decreasing the number of model parameters, and enhancing recognition efficiency. The star operation method accelerates the fusion of multi-path information flows through its powerful high-dimensional nonlinear feature space representation capability, achieving compression and acceleration of model training.

(2) We develop a dynamic threshold convolution branch to extract local image feature information. This branch initially utilizes a multi-scale convolution module (MSMod1) to capture local features of different scales, and fuses these features in a concat way to output image information containing richer features, thereby enhancing the model's capacity to comprehend images. Subsequently, it generates dynamic weights based on the fused features, and dynamically selects the number of convolutions for optimization according to predetermined weight thresholds, and constructs input-dependent dynamic convolutions to realize the rapid adjustment of model parameters and addresses issues related to the extraction of local features such as textures and edges in images.

(3) We construct an efficient multi-path dynamic attention branch to extract global image feature information. This branch initially utilizes a multi-scale convolution module (MSMod2) to provide fused feature information, and generate diverse QKV values through nonlinear transformations. Subsequently, it sorts global attention values and distinguishes features by selecting different numbers of key tokens. Compared to traditional methods, MDAM performs image recognition through key tokens, reduces the processing of redundant token information, thereby enhancing the efficiency of feature processing and further improving the model's real-time detection ability.

(4) We propose a star operation feature fusion method for efficient fusion of local and global information flows in images. This method enables the fusion of local and global information within a high-dimensional nonlinear feature space without increasing the computational load of the model. The proposed LCNet was evaluated on four widely-used image classification benchmarks: Tiny-ImageNet, Mini-ImageNet, CIFAR100, and CIFAR10. The experimental results demonstrate that LCNet consistently outperforms other state-of-the-art models with fewer parameters and lower computational cost, showcasing its advantages in low computational overhead and real-time identification.

2. Related work

In 2012, AlexNet's landmark victory in the ImageNet competition heralded the broad adoption of deep learning technologies within the image classification arena [25]. The performance of CNNs has been progressively enhanced through meticulous enhancements in network depth, width, and the integration of sophisticated attention mechanisms [26–28]. Nevertheless, the constrained receptive fields of CNNs render them suboptimal for modeling global contexts and capturing long-range dependencies. To mitigate these deficiencies, researchers have devised strategies such as optimizing multi-branch architectures and network widths, alongside enhancing feature representation through the strategic reuse of features and the integration of hierarchical information [19,29,30]. Additionally, the employment of attention-driven dynamic convolutions, which adaptively allocate distinct weights across channels, has further refined the models' capability to represent features [23,31].

$$Y = \sum_{i=1}^{N} \alpha_i(X) \times K_i(X) \tag{1}$$

Here, $X \in R^{C\times H\times W}$ represents the input feature map, with C denoting the number of input channels, and H and W specifying the height and width of the input feature map, respectively. $Y \in R^{C'\times H'\times W'}$ is the output feature map, and C' specifies the number of output channels. H' and W' representing the height and width of the output feature map. N denotes the number of dynamically generated convolution kernels. $\alpha_i(X)$ comprises dynamic weights produced by the attention mechanism, with indicating the significance coefficient of the ith convolution kernel relative to input X, fulfilling condition $\sum_{i=1}^{N} \alpha_i(X) = 1$. $K_i(X)$ represents the ith convolution kernel, dynamically generated based on the input feature X.

While these innovations have somewhat alleviated the pressures on CNNs in global feature modeling, their impact remains modest. Inspired

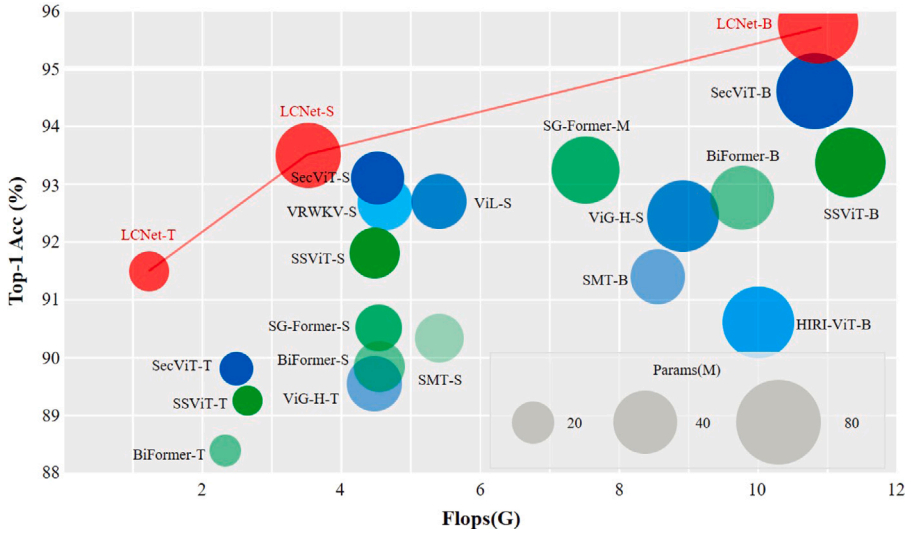

Fig. 1. LCNet achieves state-of-the-art performance on CIFAR10 public datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by the success of transformers in natural language processing, transformers have been extended to the field of computer vision and have produced state-of-the-art results across a wide range of tasks [32–35]. The Transformer architecture primarily consists of several components.

Initially, the input image is segmented into patches, embedded, and subjected to positional encoding.

$$\begin{aligned} &X = \{x_1, x_2, \ldots, x_{n-1}, x_n\} \\ &z_i = wx_i + b \quad i \in [1, n] \\ &z_i^{pos} = z_i + E_i^{pos} \end{aligned} \tag{2}$$

Here, w represents the embedding matrix, b represents the bias term, while E_i^{pos} denotes the positional encoding.

Subsequently, the model computes the QKV matrices to derive attention weights, which are then applied to generate the output feature map.

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

the function $Attention(Q, K, V)$ computes the weighted representation of the input data using the query (Q), key (K), and value (V) matrices. The multi-head attention mechanism concatenates and maps the outputs from multiple attention heads.

$$MHSA = Concat(head_1, \ldots, head_n) \times W^O \tag{4}$$

Within this mechanism, $head_i = Attention(Q_i, K_i, V_i)$, W^O are the output projection matrices, $MHSA$ is a multi head attention mechanism. $Concat$ denotes feature concatenation, where the outputs of multiple attention heads are concatenated along a specific feature dimension.

However, Transformers are challenged by their extensive parameter count, the complexity of their training processes, and their limited capacity for inductive bias. To alleviate the dependency of Transformers on extensive datasets, researchers have proposed a hybrid approach integrating both CNN and Transformer models. On one hand, CNNs and Transformers are treated as distinct modules, with their direct combination forging novel network architectures [36–38]. Conversely, a soft integration technique is employed to infuse inductive biases into the Transformer framework [39–42]. The development of CNN-Transformer hybrid models to capitalize on complementary strengths has emerged as a trend in the field of computer vision research, achieving notable successes across a spectrum of visual tasks.

The aforementioned methods have improved parameter and computational efficiency; however, they still pose significant challenges for low-performance devices and fail to achieve a balance between model performance and efficiency. To address this issue, this paper proposes the LCNet model, which features high-efficiency feature processing and is capable of completing image recognition within the millisecond range. On a low-performance device, LCNet achieved a recognition accuracy of 99.50% with an inference time of 7.2 ms. (Device specifications: Intel(R) Core (TM) i5-8300H CPU @ 2.30 GHz, 8 GB RAM, and an NVIDIA GeForce GTX 1050Ti with 4 GB.)

3. Method

In current image recognition models, there is a critical challenge: achieving a balanced trade-off among feature extraction ability, extraction efficiency, recognition accuracy, computational overhead, and real-time performance. To address this, we propose an efficient dual-path fusion model, LCNet, which has the advantages of low computational overhead and real-time detection while maintaining high recognition accuracy. LCNet consists of two components DTConv and MDAM, and integrates the dual-path information flows using a star operation feature fusion method. The structure of LCNet is illustrated in Fig. 2.

The workflow of LCNet proceeds as follows: Firstly, the input X is processed through a 3×3 depthwise separable convolution and batch normalization to output X_1.Subsequent, X_1 undergoes DWConv and is then fed into the DTConv and MDAM branches to generate local and global image feature information, respectively. Thirdly, the outputs of the DTConv and MDAM branches are fused by the star operation to output the fused image information X_2. Fourthly, X_2 is subsequently processed through a 1×1 convolution followed by a 3×3 convolution, producing X_3. Finally, X_3 is added to the original input through a residual connection to output the final result Y.

$$\begin{aligned} &X_1 = BN(Conv_{3\times3}(X)) \\ &X_2 = Star(DTConv(DWConv(X_1)), MDAM(DWConv(X_1))) \\ &X_3 = DWConv(Conv_{1\times1}(X_2)) \\ &Y = X_3 + X \end{aligned} \tag{5}$$

BN refers to batch normalization, and $Star$ represents a feature fusion method proposed in this study. This equation integrates the features from the $DTConv$ and $MDAM$ branches in an implicit high-dimensional space. $DTConv$ is a dynamic threshold convolution module for local feature extraction (detailed in Section 3.1). $Conv$ indicates a standard convolution operation, specifically with a 3×3 kernel in this context.

The advantages of adopting the aforementioned structure in LCNet are as follows: the dual-branch architecture simultaneously captures

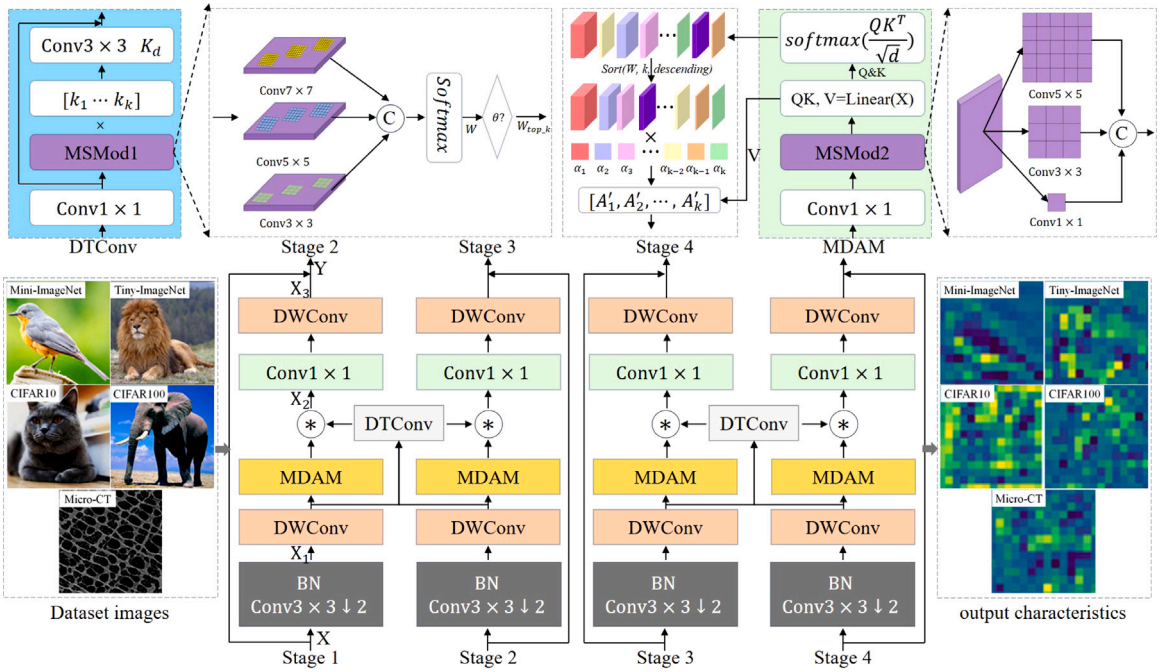

Fig. 2. The overall architecture of the proposed LCNet.

both global and local contextual information. Convolutions excel at extracting local features, while attention mechanisms effectively capture long-range dependencies and global contexts. This design enables the model to interpret data from multiple perspectives, thereby enhancing the comprehensiveness of feature extraction. By splitting the input into two parts along the channel dimension and mixing them in subsequent layers, the overall computational burden is reduced. This design allows the model to reduce its computational resource requirements while maintaining excellent performance. The dual-branch structure affords flexibility in network design, permitting adjustments to the architecture and parameters of each branch according to task-specific requirements. This flexibility also makes it easier for the network to expand and adapt to different datasets and tasks.

3.1. Dynamic threshold convolution

The proposed DTConv is one branch of LCNet. Its multi-scale structure and dynamic convolutional aggregation strategy represent an innovative approach aimed at overcoming the shortcomings of inadequate local feature extraction and low extraction efficiency in image recognition. The structure of DTConv is illustrated in Fig. 3, and its implementation can be described as follows. First, the input feature $X \in R^{C_{in}\times H\times W}$ is divided into three sub-features according to the channel dimension, namely $X_1 \in R^{\frac{C_{in}}{3}\times H\times W}$, $X_2 \in R^{\frac{C_{in}}{3}\times H\times W}$, and $X_3 \in R^{\frac{C_{in}}{3}\times H\times W}$. A linear transformation matrix W_i is then defined to project each sub-feature into a specific dimension, thereby generating the base features of multi-scale features.

$$\hat{X}_i = W_i X_i + b_i, i \in \{1,2,3\} \tag{6}$$

where $W_i \in R^{C_i'\times C_i}$ denotes the linear transformation function, and b_i represents the bias vector.

Next, the projected features $\hat{X}_i$ are processed by the multi-scale convolution module MSMod1 (with $k_1 = 3\times 3$, $k_2 = 5\times 5$, and $k_3 = 7\times 7$) and a non-linear activation function σ to generate the multi-scale feature information F_i.

$$F_i = \sigma(Conv_{k_i}(\hat{X}_i)), i \in \{1,2,3\} \tag{7}$$

$Conv$ denotes convolutions with varying kernel sizes, specifically $3\times 3, 5\times 5$, and 7×7.

Then, to enhance the numerical stability of features at different scales, each F_i undergoes batch normalization (BN), producing the normalized features $\hat{F}_i$. Subsequently, the normalized features are concatenated in the channel dimension using concat method to generate the multi-scale fusion features X_{multi}.

$$\begin{aligned} &\hat{F}_i = BN(F_i), i \in \{1,2,3\} \\ &X_{multi} = Concat(\hat{F}_1, \hat{F}_2, \hat{F}_3, axis = 1) \end{aligned} \tag{8}$$

Finally, X_{multi} is flattened into a one-dimensional vector X_n using the flatten operation, and X_n is then processed by the Softmax function to generate the one-dimensional dynamic weight W. Sort W in descending order, then select the top k weights based on the predefined weight threshold θ to obtain W_{top_k}. The selected k weights W_{top_k} are assigned to k parallel convolution kernels to implement the dynamic threshold convolution. The dynamic threshold convolution is applied to the input features, producing the output features G.

$$\begin{aligned} &x_n = flatten(X_{multi}) \\ &w_i = Softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad i \in [1,n] \\ &W = [w_1, \ldots, w_n] \\ &W_{top\_k} = sort(W, k, descending) \\ &G = X * \sum_{i=1}^{k} (W_{top\_k}^i \times H_i) \quad i \in [1,k] \end{aligned} \tag{9}$$

Here, $flatten$ refers to the operation of converting a multi-dimensional tensor into a one-dimensional vector. Softmax is an activation function that maps a real-valued vector into a probability distribution between 0 and 1. The symbol σ represents a nonlinear activation function. $Sort$ arranges the input data in descending order and selects the top k elements. For better clarity, we have renamed $Conv_i$ as H, where H denotes a filter. Thus, $W_{top\_k}^i \times H$ indicates that the $W_{top\_k}^i$ weights are multiplied by filter H.

DTConv offers several advantages. MSMod1 can reveal the details and feature information of images at different scales, thereby enhancing the image expression ability of DTConv. By setting a weight threshold to aggregate different number of convolution kernels, DTConv enables the rapid update of convolution parameters and improves the efficiency of feature processing. The dynamic threshold convolution based on input aggregation aligns more closely with the input image, thereby enhancing the model's ability to extract the local image features.

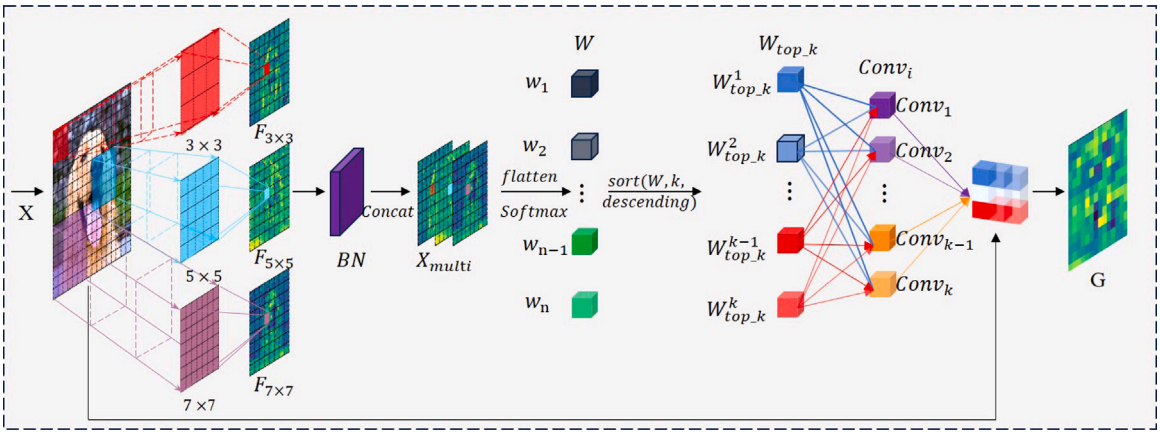

Fig. 3. DTConv structure diagram.

The proposed DTConv addresses the issues of inefficient local feature extraction and incomplete feature acquisition. The subsequent experimental evaluations and comparison with other dynamic convolution methods further validate DTConv′s superior performance.

3.2. Multi-path dynamic attention mechanism

Multi-path Dynamic Attention Mechanism (MDAM) proposed in this study is the other branch of LCNet. The token-based dynamic selection method in MDAM introduces a degree of innovation, aiming to solve the problem of low feature extraction efficiency of traditional attention mechanisms. MDAM comprises three components: QKV retrieval based on multi-scale features, dynamic modeling, and attention modification, as illustrated in Fig. 4. The detailed implementation is as follows.

QKV retrieval is based on multi-scale features. Firstly, convolution kernels of different sizes are applied to the input feature $X \in R^{C\times H\times W}$ to obtain multi-scale feature representations. Specifically, kernels of sizes 1×1, 3×3, and 5×5 are employed, with the feature extraction formula as follows.

$$X'_k = \sigma(W_{k\times k}X + b), \quad k \in \{1, 3, 5\} \tag{10}$$

Here, X'_k denotes the feature map obtained by applying a $k \times k$ convolution operation to the input X, σ is the activation function.

Secondly, the multi-scale feature representations are concatenated in the channel dimension to form the extended feature map $X_{multi} \in R^{C'\times H\times W}$, where C' denotes the number of channels after merging.

$$X_{multi} = Concat(X'_1, X'_3, X'_5) \tag{11}$$

Here, $Concat(\cdot)$ denotes the concatenation operation along the channel dimension.

Finally, X_{multi} is projected into the spaces of the query vector Q, the key vector K, and the value vector V by linear transformations, as given by the following equations.

$$\begin{aligned} Q &= W_Q X_{multi} + b_Q \\ K &= W_K X_{multi} + b_K \\ V &= W_V X_{multi} + b_V \end{aligned} \tag{12}$$

where W_Q, W_K, and W_V are linear transformation matrices responsible for generating the query, key, and value vectors, respectively, while b_Q, b_K, and b_V are the corresponding bias terms.

Dynamic Modeling. In our designed multi-head attention mechanism, the attention matrix of each head is calculated first. Q_t and K_t denote the query and key matrices of the tth head, respectively, where d represents the feature dimension. For each head t, the attention is calculated by the dot product between the query matrix Q_t and the key matrix K_t.

$$a_t = softmax(\frac{Q_t K_t^T}{\sqrt{d}}) \tag{13}$$

Next, the attention matrices of all heads are aggregated to form a global attention matrix A. Each element in A, denoted as $A[i, j]$, represents the total attention of a query token i on a key token j. The total attention for each key token, denoted as A_s, is calculated by summing the attention values of all query tokens.

$$\begin{aligned} A &= \sum_{t=1}^{num} a_t \\ A_s &= \sum_i A[i, j] \end{aligned} \tag{14}$$

where A is a matrix of dimension $R^{N\times N}$, and num denotes the number of heads, s represents the number of key tokens, and A_s denotes the total attention value of each K token, which is obtained by summing over the Q tokens.

Finally, the attention values of all key tokens are sorted in descending order. The first k key tokens (i.e., those with the highest attention values) are determined experimentally as the crucial tokens, while the others are discarded. As illustrated in the middle part of Fig. 4, the tokens related to the dog are retained, while other background tokens are ignored. In this manner, the model identifies high-attention key tokens and disregards irrelevant information.

$$W_{top\_k} = sort(A_s, k, descending) \tag{15}$$

Attention Correction. The selected k key tokens form the set τ, and the importance of each element in τ is adjusted by a scaling factor α to generate the corrected attention map A'. The value of α ranges from [0,1], where $\alpha = 0$ indicates that the attention values are entirely overridden and not activated, while $\alpha = 1$ signifies that the attention remains unmodified and plays a crucial role. As illustrated in Fig. 4, the attention weights of the dog′s mouth and eyes remain unchanged, while the attention values of other features, such as fur, undergo different degrees of attenuation

$$A'_t[i, j] = \begin{cases} 0 & j \notin \tau \\ \alpha_t \cdot A_t[i, j] & j \in \tau \end{cases} \tag{16}$$

The corrected attention matrix A'_t is multiplied by each head′s value matrix V_t, producing the improved output for each head. Subsequently, the outputs of all heads are concatenated ($H = Concat(A'_t \cdot V_t) \cdot W$) and passed to the next module. W is a learnable weight matrix used to project the concatenated vector into an appropriate feature space, forming the final hidden representation H.

The MDAM structure offers several advantages. MSMod2 can enhance its feature representation ability by providing more comprehensive input feature information for the generation of QKV, thus making the model more stable when focusing on the salient regions of the image. Image categories are distinguished by focusing on key feature tokens to reduce the computational overhead of the model. Image identification through key tokens can minimize the interference of external factors (e.g., image background, noise, etc.), so as to improve the stability and generalization ability of the model.

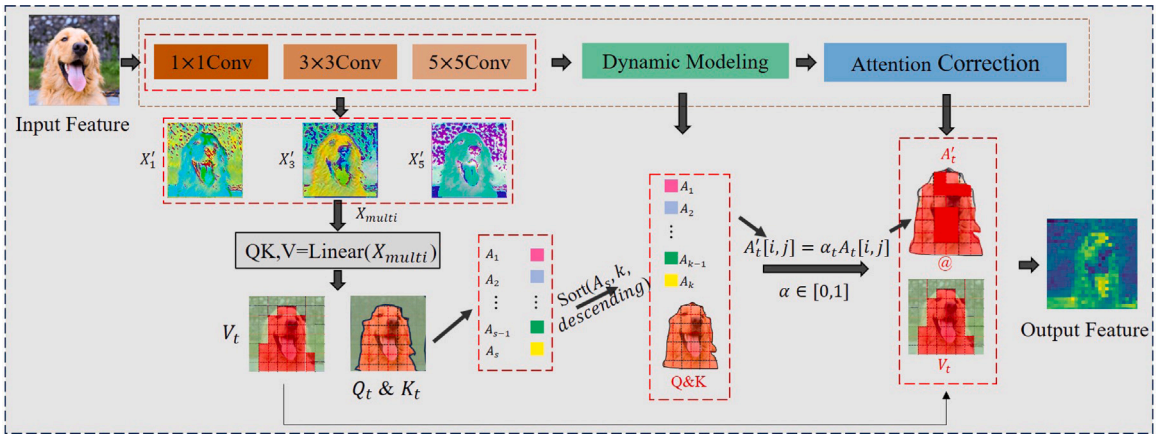

Fig. 4. MDAM structure diagram.

The traditional attention mechanism can fully extract the global features of the image, but the model training efficiency is often reduced due to the large number of parameters. In contrast, the proposed MDAM reduces the amount of attention parameters by optimizing the number of tokens, thereby decreasing computational overhead. Subsequent experimental evaluations and comparisons with other attention mechanisms further verify MDAM′s efficiency.

3.3. Feature fusion method

DTConv addresses the issues of low efficiency and insufficient extraction of local image features. MDAM solves the problem of high computational overhead when extracting global image features. In order to efficiently fuse global and local image information, this study presents a highly efficient and concise feature fusion method, called the star operation method, which can map inputs into a high-dimensional nonlinear feature space without increasing network parameters to improve the information fusion efficiency. The implementation of the star operation feature fusion method is as follows.

The star operation is realized as Eq. (17).

$$\begin{aligned} & w_1^T(x) * w_2^T(x) \\ &= \left(\sum_{i=1}^{d+1} w_1^i x^i\right) * \left(\sum_{j=1}^{d+1} w_2^j x^j\right) \\ &= \sum_{i=1}^{d+1}\sum_{j=1}^{d+1} w_1^i w_2^j x^i x^j \\ &= \underbrace{\alpha_{(1,1)} x^1 x^1 + \cdots + \alpha_{(d+1,d+1)} x^{d+1} x^{d+1}}_{(d+2)(d+1)/2 \text{ items}} \end{aligned} \tag{17}$$

By employing the star operation with high computational efficiency in a high-dimensional space, an implicit dimensional feature space of $(d+2)(d+1)/2$ nonlinear terms is obtained. This approach enables information fusion in high-dimensional feature spaces without incurring any additional computational overhead in a single layer.

$$\alpha_{(i,j)} = \begin{cases} w_1^i w_2^j & if\ i = j \\ w_1^i w_2^j + w_1^j w_2^i & if\ i \neq j \end{cases} \tag{18}$$

where i and j denote the channel subscripts, and α represents the coefficient of each subterm

By employing the star operation method to integrate DTConv and MDAM feature information in a high-dimensional feature space, no additional computational overhead is introduced. This approach not only ensures the full fusion of DTConv and MDAM but also compresses the model and accelerates the training of the model. By fusing the local and global feature information of the image by the star operation method, LCNet attains the advantages of low computational overhead, high recognition accuracy, and real-time detection.

3.4. Architecture variants

In this paper, DTConv local branch and MDAM global branch are proposed, whose information flow is fused by the star operation method. Finally, an efficient dual-path image recognition model LCNet consisting of four stages is constructed, which aims to reduce the computational overhead and improve the ability of real-time detection of the model while ensuring the high accuracy. To facilitate comparisons with other SOTA backbone networks, we construct different versions of LCNet by adjusting the number of convolution kernels, the number of convolution layers, the number of attention layers, and the depth of the model, as shown in Table 1. Specifically, we design three variants of LCNet — LCNet-T, LCNet-S, and LCNet-B — differentiated by parameter scale and computational cost. As model depth increases, more convolution kernels are required to extract richer features, resulting in doubling their number at each stage. In addition, to ensure that the attention layers can adequately capture the convolutional features and their relative positional information, the embedding size is set to 64.

4. Result and discussion

4.1. Classification result

To ensure the general applicability of LCNet, experiments were conducted on five diverse datasets: Tiny-ImageNet [43], Mini-ImageNet [44], CIFAR100 [45], CIFAR10 [45], and Micro-CT. The proposed models are implemented on the PyTorch framework. AdamW [46] is adopted as the optimizer, with the learning rate set to 0.001 and the weight decay set to 0.001 by default. The learning rate is adjusted through the cosine annealing method. The experimental results are detailed in Table 2.

The experimental results indicate that LCNet outperforms all other models with fewer parameters and lower computational costs. Specifically, on the Tiny-ImageNet dataset, LCNet-T achieved a Top-1 accuracy of 58.62% with 1.28 GFLOPs and 19.48 M parameters. This represents an improvement of 1.74% over Vi-H-T (4.46 GFLOPs vs. 1.28 GFLOPs, a reduction of 71.3%; 29.15 M vs. 19.48 M parameters, a reduction of 31.9%), an improvement of 1.41% over SMT-S (5.37 GFLOPs vs. 1.28 GFLOPs, a reduction of 76.2%; 22.55 M vs. 19.48 M parameters, a reduction of 13.6%), and an improvement of 1.06% over SG-Former-S (4.57 GFLOPs vs. 1.28 GFLOPs, a reduction of 72.0%; 22.31 M vs. 19.48 M parameters, a reduction of 12.7%). Furthermore, in comparison with other SOTA models on the other four datasets, LCNet-T achieved Top-1 accuracies of 62.25%, 66.71%, 91.50%, and 98.89%, respectively, maintaining its leading position. Notably, LCNet-B achieved Top-1 accuracies of 62.75%, 66.62%, 70.24%, 95.82%, and 99.75% on the five datasets, with a computational cost of only

Table 1
Only vary the embed width and the depth to build different sizes of LCNet.

Stage	Input size	LCNet-T	LCNet-S	LCNet-B
1	224 × 224	$\begin{bmatrix} 3\times3, 24 \\ DTConv \\ MDAM \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 32 \\ DTConv \\ MDAM \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 32 \\ DTConv \\ MDAM \end{bmatrix} \times 4$
2	56 × 56	$\begin{bmatrix} 3\times3, 48 \\ DTConv \\ MDAM \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64 \\ DTConv \\ MDAM \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 32 \\ DTConv \\ MDAM \end{bmatrix} \times 6$
3	28 × 28	$\begin{bmatrix} 3\times3, 96 \\ DTConv \\ MDAM \end{bmatrix} \times 8$	$\begin{bmatrix} 3\times3, 128 \\ DTConv \\ MDAM \end{bmatrix} \times 12$	$\begin{bmatrix} 3\times3, 128 \\ DTConv \\ MDAM \end{bmatrix} \times 16$
4	14 × 14	$\begin{bmatrix} 3\times3, 192 \\ DTConv \\ MDAM \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 256 \\ DTConv \\ MDAM \end{bmatrix} \times 4$	$\begin{bmatrix} 3\times3, 256 \\ DTConv \\ MDAM \end{bmatrix} \times 4$
	7 × 7		Global average pooling	
	1 × 1		Fully connected layer, 1000	
#Flops		1.28 G	3.39 G	4.92 G
#Params		19.48 M	40.21 M	75.42 M

Table 2
Comparison of top-1 accuracy (%) with SOTA schemes.

Model	Top-1_Acc/%					#F (G)	#P (M)
	Tiny-ImageNet	Mini-ImageNet	CIFAR100	CIFAR10	Micro-CT		
BiFormer-T [9]	55.29	58.92	63.19	88.31	97.13	2.21	15.42
CAT-T [10]	55.73	59.21	63.34	88.45	97.27	2.67	16.65
SMT-T [14]	56.27	59.64	63.89	88.76	97.36	2.43	16.54
SSViT-T [11]	56.64	59.96	64.31	89.12	97.73	2.62	16.13
ViG-H-T [32]	56.88	60.31	64.67	89.64	97.95	4.46	29.15
SecViT-T [12]	56.89	60.68	65.39	89.86	98.11	2.56	15.84
SMT-S [14]	57.21	61.13	65.84	90.23	98.32	5.37	22.55
SG-Former-S [47]	57.56	61.69	66.21	90.68	98.64	4.57	22.31
LCNet-T	**58.62**	**62.25**	**66.71**	**91.50**	**98.89**	**1.28**	**19.48**
BiFormer-S [9]	56.45	59.93	64.55	89.64	97.84	4.57	28.32
CAT-S [10]	56.96	60.73	64.83	90.21	97.98	5.91	36.97
HIRI-ViT-B [17]	57.49	61.22	65.26	90.76	98.13	9.95	49.46
SMT-B [14]	58.12	61.69	65.79	91.35	98.22	8.65	32.04
SSViT-S [11]	58.46	62.25	66.25	91.89	98.39	4.44	28.35
ViG-H-S [32]	58.73	62.83	66.77	92.36	98.56	8.91	49.36
ViL-S [33]	59.12	63.17	66.92	92.71	98.77	5.13	27.32
VRWKV-S [34]	59.31	64.06	67.23	92.88	98.91	4.66	26.92
SecViT-S [12]	59.66	65.51	67.56	93.09	99.09	4.52	28.64
SG-Former-M [47]	59.95	64.06	67.94	93.21	99.21	7.52	43.72
LCNet-S	**60.16**	**64.34**	**68.42**	**93.50**	**99.34**	**3.39**	**40.21**
BiFormer-B [9]	59.33	63.29	67.62	92.77	98.46	9.83	64.98
HIRI-ViT-L [17]	59.88	63.62	67.91	93.05	98.55	19.92	94.42
SSViT-B [11]	60.31	64.12	68.15	93.29	98.89	11.25	66.62
ViG-H-B [32]	60.78	64.68	68.39	93.66	99.02	15.69	88.91
ViL-B [33]	61.26	64.91	68.73	94.17	99.28	18.65	88.56
VRWKV-B [34]	61.47	65.23	69.16	94.39	99.36	I18.21	93.64
SecViT-B [12]	61.89	65.79	69.38	94.78	99.46	10.86	62.35
SG-Former-B [47]	62.39	66.19	69.86	95.26	99.52	14.99	77.61
LCNet-B	**62.75**	**66.62**	**70.24**	**95.82**	**99.75**	**10.92**	**75.42**

10.92 GFLOPs and 75.42 M parameters. This surpasses many larger models, such as ViL-B, VRWKV-B, and ViG-H-B, which have more than 85 M parameters and more than 15 GFLOPs computational costs. In summary, by combining dynamic threshold convolution and multi-path dynamic attention, LCNet can efficiently and accurately extract both macroscopic features (such as image contours and spatial layouts) and microscopic features (such as image textures and edges). Furthermore, LCNet can achieve maximum recognition accuracy with only a minimal computational overhead. Additionally, the three variants of LCNet developed in this study are designed to meet diverse requirements. LCNet-T features low computational overhead and strong real-time detection capabilities, making it suitable for applications such as real-time detection of timber and rapid medical diagnosis. LCNet-B offers the highest recognition accuracy with slightly longer detection time, which is suitable for high-precision, non-real-time detection scenarios. LCNet-S combines the features of both variants, providing stronger universality.

4.2. Model validation

To validate the practical applicability of LCNet, this study selects five SOTA models with similar computational complexity and parameter sizes as LCNet, and evaluates their recognition accuracy and real-time detection performance on the Micro-CT dataset. The results are detailed in Table 3.

Among the six models evaluated, the proposed LCNet-T outperforms all other SOTA models with the highest recognition accuracy of 99.50%. Additionally, LCNet-T stands out as the most efficient model with only 0.0072 s of recognition time, highlighting its real-time detection capabilities in practical applications. BiFormer-T records the lowest recognition accuracy and efficiency. Although the SMT-S model′s recognition accuracy is comparable to that of LCNet-T, its parameter count is significantly higher than that of LCNet-T by 4.2 times. Furthermore, SMT-S′s detection time is twice that of LCNet-T, resulting in higher time costs for large-scale image recognition

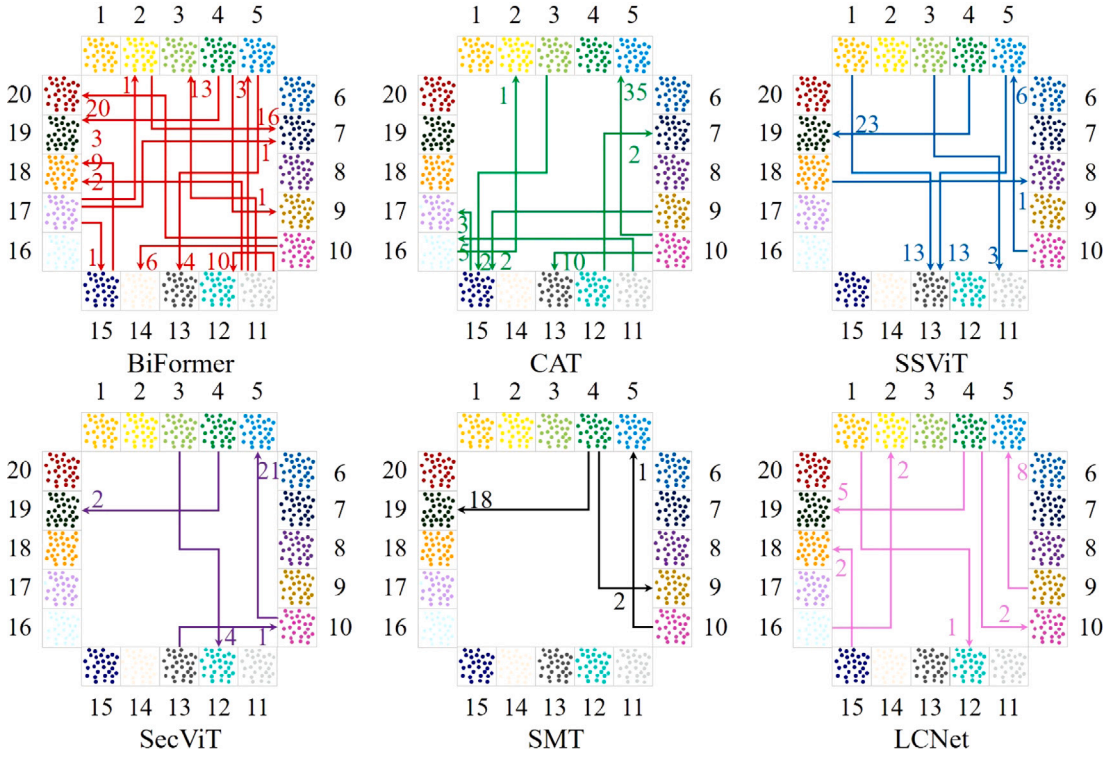

Fig. 5. The verification results of LCNet and other SOTAs on Micro-CT.

tasks. LCNet's ability to achieve high recognition accuracy and efficient real-time detection capabilities can be attributed to two key factors: First, the dynamic threshold convolution employed by LCNet enriches image features through MSMod1, thereby enhancing the model's representation ability. Additionally, by using dynamic weights to rapidly aggregate multiple convolutions, LCNet generates input-dependent dynamic convolutions, enabling accurate and comprehensive extraction of input image features. Secondly, the multi-path dynamic attention mechanism in LCNet optimizes the number of selected tokens, which not only reduces the model's parameter count, but also avoids the computational overhead on redundant information. The combination of the two mechanisms not only ensures LCNet's high accuracy but also improves the processing speed of feature information and real-time detection efficiency. This advantage of balancing model accuracy and detection efficiency enables LCNet to stand out among numerous models, making it an ideal choice for real-time recognition tasks or large-scale data processing.

To further demonstrate the specific recognition performance of the six models on the Micro-CT dataset, we present their confusion matrices as shown in Fig. 5. In the image, the numbers 1–20 represent the labels of 20 different tree species, with polyline arrows indicating the direction from actual labels to predicted labels. The numbers denote the quantity of misclassified images. In the validation experiments, BiFormer, CAT, SSViT, SecViT, SMT, and LCNet all misclassified wood images with label 10 as wood images with label 5. This is due to the fact that the woods corresponding to these two labels both have separate duct pores with obturators, and the cells are arranged in separate radial directions, making their microscopic images extremely similar in morphology and structural features. This similarity hinders the models to accurately capture the distinguishing features between species 10 and species 5, resulting in misclassification. Among the six models, BiFormer consistently exhibited errors in the recognition of multiple wood species, which indicates that the image features extracted by BiFormer are insufficient to analyze distinguishable salient features, resulting in the lowest recognition accuracy. This is attributed to the

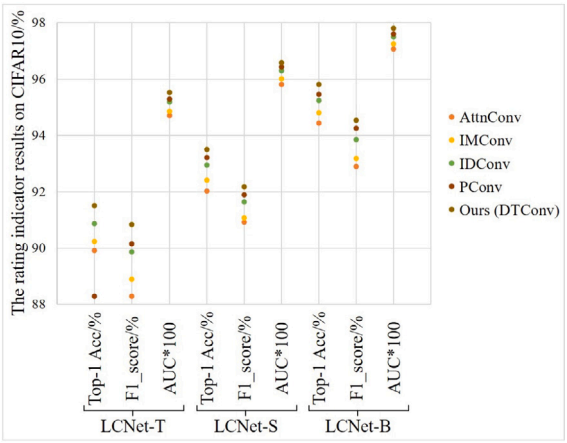

Fig. 6. The results of DTConv and other dynamic convolutions on CIFAR10.

design of BiFormer, which directly reduces the memory usage and computational cost by reducing key–value pairs in rough regions without sufficient image features, making the model parameters unable to train adequately. In contrast, LCNet can accurately and efficiently extract both local and global information of the images through DTConv and MDAM. By employing different convolution types and varying the number of tokens for different input images, the model is capable of capturing unique features inherent to each image, thereby enhancing its recognition ability.

4.3. Ablation experiment

Comparison between DTConv and other dynamic convolutions. To evaluate the effectiveness of DTConv, we replaced DTConv in LCNet-T with a series of alternative methods, including AttnConv [35], IMConv [23], IDConv [18], and PConv [48]. To ensure a fair comparison,

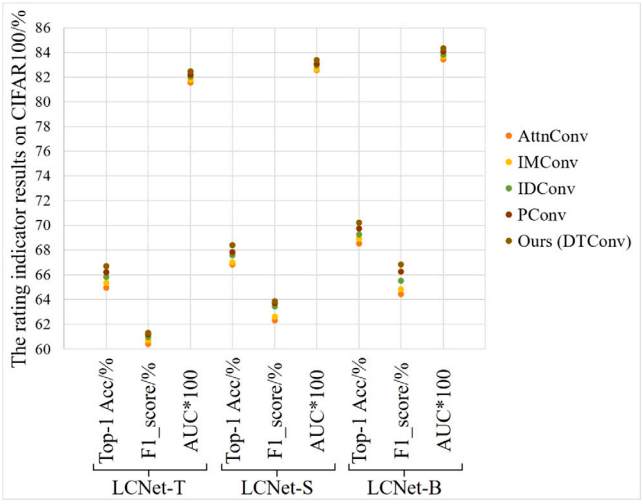

Fig. 7. The results of DTConv and other dynamic convolutions on CIFAR100.

Table 3
Comparison of validation results on Micro-CT with SOTA model.

Model	#F (G)	#P (M)	Top-1_Acc/%	Val_Time/s
BiFormer-T [9]	2.21	15.42	97.45	0.0081
CAT-T [10]	2.67	16.65	98.50	0.0083
SSViT-T [11]	2.62	16.13	98.78	0.0085
SecViT-T [12]	2.56	15.84	99.20	0.0082
SMT-S [14]	5.37	22.55	99.48	0.0141
LCNet-T	**1.28**	**19.48**	**99.50**	**0.0072**

the convolution kernel sizes of all methods were standardized, and the performance was evaluated using Top-1 accuracy, F1_score, and AUC, with results presented in Table 4, Figs. 6, and 7. DTConv achieved the highest accuracy on both CIFAR10 and CIFAR100 datasets, with Top-1 accuracies of 95.82% and 70.24%, respectively. This advantage is attributed to DTConv′s adoption of multi-scale dynamic convolutions, which capture salient features such as multi-scale local textures and edges, while enhancing the adaptability of convolution operations to different targets, thus improving overall classification performance. In addition, the model achieved F1 scores of 94.54% and 65.84% on the CIFAR10 and CIFAR100 datasets, respectively, indicating that the introduction of DTConv enables LCNet to quickly adapt to imbalanced datasets, thereby achieving optimal training and strong recognition ability. This is due to DTConv′s construction through the dynamic aggregation of k parallel convolutions, which enables it to efficiently adjust the model parameters according to input data and quickly adapt to multiple categories and quantities of image features. Furthermore, LCNet achieved AUC values of 97.80% and 84.34% on the CIFAR-10 and CIFAR100 datasets, respectively, indicating that LCNet effectively reduces the influence of class sample imbalance and can accurately distinguish between positive and negative classes, thereby demonstrating strong robustness.

Comparison between MDAM and other attention mechanisms. This study investigates the impact of various attention mechanisms on model classification performance using a LCNet-T. The comparison results between MDAM and SSSA [11], SECA [12], MAttn [49], and BFSA [50] are presented in Table 5, Figs. 8, and 9. MDAM achieved the highest Top-1 accuracy, F1 score, and AUC values on both CIFAR10 and CIFAR100 datasets. In terms of recognition accuracy, MDAM improved by 0.85%, 1.28%, 1.58%, and 1.77% compared with MAttn, BFSA, SSSA, and SECA, respectively. The superior performance of MDAM can be attributed to two main factors: First, MDAM uses multi-scale convolutions to capture image features at different scales, thereby enriching the model′s understanding of image patterns. Second, MDAM

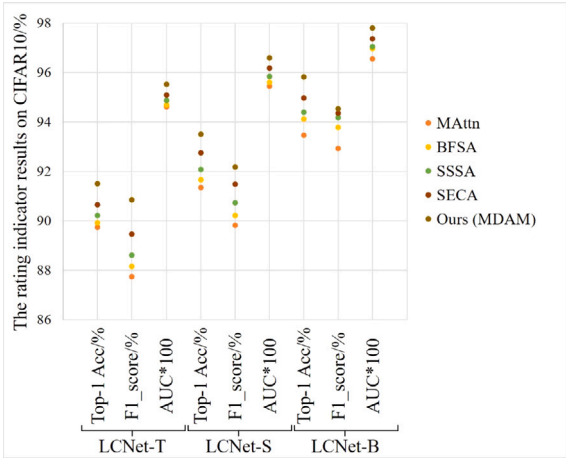

Fig. 8. The results of DTConv and other dynamic convolutions on CIFAR10.

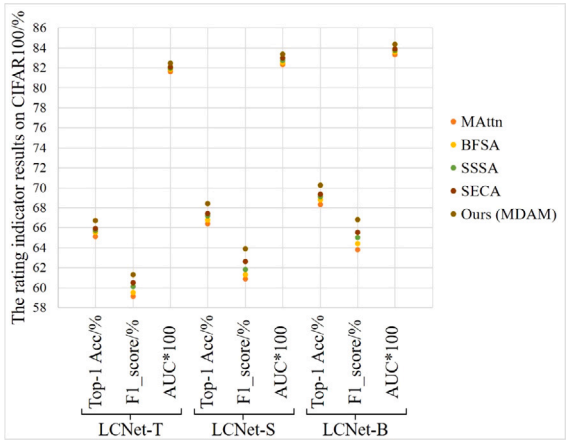

Fig. 9. The results of DTConv and other dynamic convolutions on CIFAR100.

dynamically selects feature tokens to eliminate redundant tokens, while enhancing the weight proportions of key tokens in salient and distinguishable regions and suppressing the weights of other regions. The combination of these methods not only improves the feature reuse rate of MDAM but also significantly reduces the parameter count and computational load of LCNet, thereby improving its image recognition accuracy and efficiency.

4.4. Visualization analysis

Features Visualization [29]. To visualize more intuitively the hierarchical features learned by LCNet, we conducted feature visualizations on images from five distinct datasets, as illustrated in Fig. 10. During the experiment, we randomly visualized the feature maps of 6 channels at each of the four stages (stage 1 through stage 4) of LCNet. From stage 1 to stage 4, LCNet incrementally captures features from shallow to deep layers, progressively extracting information across multiple levels. In the shallow stages (stage 1 and stage 2), the model typically captures low-level features, such as edges, textures, shapes, and contours, whereas in the deeper stages, the feature maps reveal more abstract and semantic-level features. Through this gradual process of hierarchical feature extraction, the model incrementally constructs high-level semantic features from basic edge information in the images. Fig. 10 further illustrates the model′s adaptability across diverse datasets, with samples from each dataset exhibiting unique feature distributions at the same hierarchical level, emphasizing specific feature types (e.g., animal eyes, feathers, noses) at different stages. This

Table 4
Comparison between DTConv and other dynamic convolutions.

Model	Convolution type	Top-1_Acc/%		F1_score/%		AUC*100	
		CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
LCNet-T	AttnConv [35]	89.92	64.95	88.29	60.42	94.70	81.55
	IMConv [23]	90.23	65.38	88.79	60.70	94.86	81.78
	IDConv [18]	90.87	65.84	89.86	60.99	95.20	82.03
	PConv [48]	90.23	66.21	90.15	61.19	95.29	82.21
	Ours (DTConv)	**91.50**	**66.71**	**90.84**	**61.32**	**95.52**	**82.47**
LCNet-S	AttnConv [35]	63.17	66.92	92.71	98.77	5.13	27.32
	IMConv [23]	64.06	67.23	92.88	98.91	4.66	26.92
	IDConv [18]	65.51	67.56	93.09	99.09	4.52	28.64
	PConv [48]	64.06	67.94	93.21	99.21	7.52	43.72
	Ours (DTConv)	**93.50**	**68.42**	**92.17**	**63.90**	**96.59**	**83.38**
LCNet-B	AttnConv [35]	94.43	68.53	92.90	64.46	97.07	83.43
	IMConv [23]	94.81	68.94	93.19	64.86	97.26	83.65
	IDConv [18]	95.24	69.29	93.86	65.53	97.51	83.84
	PConv [48]	95.46	69.73	94.26	66.27	97.61	84.08
	Ours (DTConv)	**95.82**	**70.24**	**94.54**	**66.84**	**97.80**	**84.34**

Table 5
Comparison between MDAM and other attention mechanism.

Model	Attention type	Top-1_Acc/%		F1_score/%		AUC*100	
		CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
LCNet-T	MAttn [49]	89.73	65.12	87.73	59.12	94.60	81.64
	BFSA [50]	89.92	65.54	88.16	59.51	94.69	81.87
	SSSA [11]	90.22	65.73	88.61	60.13	94.86	81.96
	SECA [12]	90.65	65.93	89.46	60.48	95.05	82.08
	Ours (MDAM)	**91.50**	**66.71**	**90.84**	**61.32**	**95.52**	**82.47**
LCNet-S	MAttn [49]	91.34	66.42	89.81	60.86	95.45	82.32
	BFSA [50]	91.65	66.78	90.21	61.29	95.61	82.51
	SSSA [11]	92.67	67.21	90.72	61.82	95.83	82.75
	SECA [12]	92.74	67.64	91.47	62.64	96.18	82.97
	Ours (MDAM)	**93.50**	**68.42**	**92.17**	**63.90**	**96.59**	**83.38**
LCNet-B	MAttn [49]	93.46	68.34	92.29	63.79	96.55	83.34
	BFSA [50]	94.11	68.82	93.78	64.43	96.91	83.59
	SSSA [11]	94.39	69.12	94.18	65.03	97.05	83.75
	SECA [12]	94.97	69.39	94.35	65.54	97.37	83.89
	Ours (MDAM)	**95.82**	**70.24**	**94.54**	**66.84**	**97.80**	**84.34**

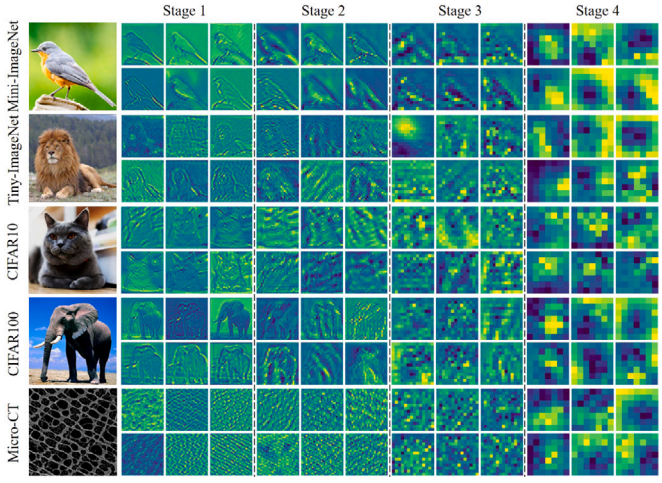

Fig. 10. Visualization of features from 6 random channels in the last convolutional layer of LCNet.

adaptability highlights LCNet′s ability to capture features tailored to varied image characteristics.

Grad-CAM Visualization [51]. In this study, one image was randomly selected from each dataset for visualization analysis using the Grad-CAM technique on BiFormer, SSViT, and LCNet models. These activation maps provide insight into the significance of individual pixels in depicting class discrimination for each input image. As demonstrated in Fig. 11, our method more effectively reveals critical feature points within images that play key roles in identification, resulting in a more comprehensive extraction of image features. Moreover, the features captured by LCNet are focused on the object itself, enabling the filtering of external disturbances and exhibiting robust anti-interference capabilities. This compellingly demonstrates the robust visual representation capabilities of our method compared to other competitors.

T-SNE Visualization [52]. We use some of the categories of the Micro-CT dataset as examples for 2D visualization experiments as shown in Fig. 12, in order to validate the effectiveness of our proposed method more intuitively. We selected BiFormer-T and SSViT-T to compare with LCNet-T. The projection of the low-dimensional manifold features extracted by the three methods onto the Euclidean space is mapped onto the 2-D space by the t-SNE technique. Fig Q reveals that the visualized features of BiFormer-T are the most dispersed, with significant overlap among different features. SSViT-T′s visualized features exhibit better distinction among different classes, although there is still some overlap and lower intra-class compactness. The visualized results of the LCNet-T, proposed in this paper, demonstrate greater intra-class compactness and clear distinction among different categories, thereby facilitating improved classification outcomes via the classifier. The efficacy of our proposed method can be directly visualized more effectively.

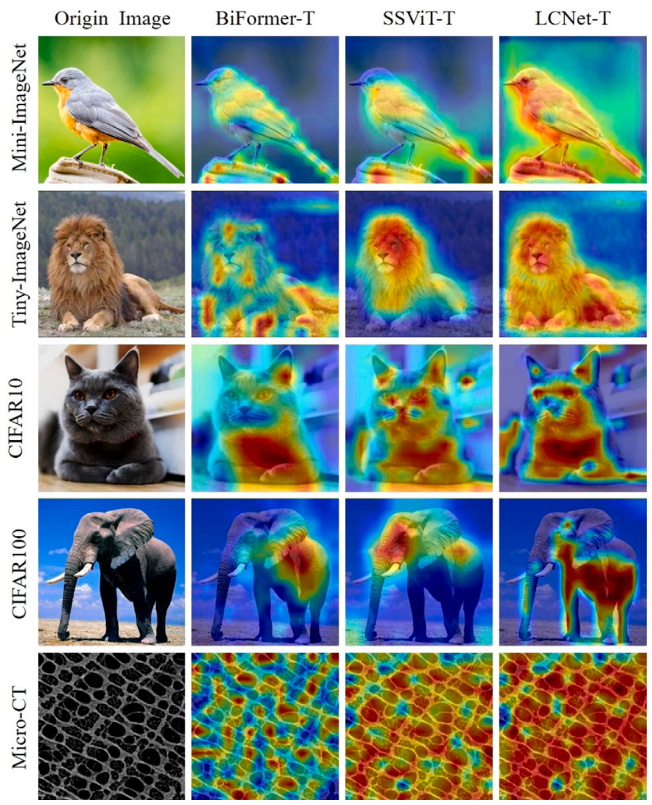

Fig. 11. Grad-CAM visualization of BiFormer, SSViT, and LCNet on the five datasets.

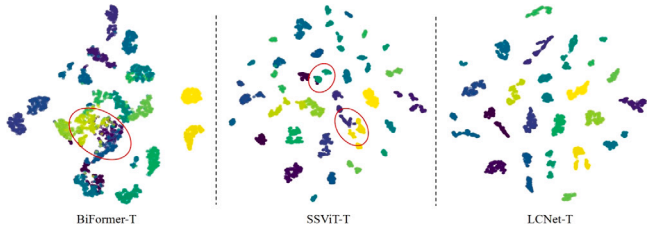

Fig. 12. T-SNE visualization of BiFormer, SSViT, and LCNet on the Micro-CT dataset.

5. Conclusion

In this study, a lightweight and efficient dual-path real-time identification network is proposed to solve the problem of imbalance between the efficiency and accuracy of real-time identification and the inability to be applied to low-performance devices. The model comprises DTConv and MDAM, and the information exchange between them is realized by star operation method. LCNet provides image context information at different scales through MSMod1 and MSMod2, thereby enriching feature diversity and improving the model′s recognition accuracy. Meanwhile, LCNet selectively aggregates different numbers of convolutions based on input features while optimizing the number of tokens in the attention mechanism, thereby reducing the model′s parameter count and enhancing its real-time recognition capabilities. The experimental results demonstrate that LCNet outperforms the latest state-of-the-art (SOTA) models, including SecViT and SG-Former, on public datasets such as Tiny-ImageNet, Mini-ImageNet, CIFAR100, and CIFAR-10. In applications on the self-constructed wood dataset, LCNet achieved a recognition accuracy of 99.50% with the processing time of only 0.0072 s per image, showcasing its superior real-time detection capabilities. The Ablation experiments further confirm that the proposed DTConv and MDAM outperform other dynamic convolution and attention mechanisms in terms of lower computational overhead and higher recognition accuracy. LCNet constructed by combining DTConv, MDAM, and the star operation fusion method is capable of recognizing nearly 140 images within 1 s, demonstrating significant practical application value in areas such as customs real-time inspection, video surveillance, and road traffic accident judgment.

CRediT authorship contribution statement

Xiaoxia Yang: Writing – original draft, Methodology, Conceptualization. **Zhishuai Zheng:** Writing – review & editing, Data curation. **Huanqi Zheng:** Visualization, Validation. **Zhedong Ge:** Project administration, Funding acquisition, Formal analysis. **Xiaotong Liu:** Supervision. **Bei Zhang:** Data curation. **Jinyang Lv:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is partly supported by Funding Statement: The Natural Science Foundation of Shandong Province, China (ZR2024MC185); The Philosophy and Social Science Research Project of Guangxi Province (23FMZ025); The Taishan Scholar Advantage Characteristic Discipline Talent Team Project of Shandong Province of China (2015162).

Data availability

Data will be made available on request.

References

[1] Kun Hu, Yuanbin Mo, An efficient multi-threshold image segmentation method for COVID-19 images using reinforcement learning-based enhanced sand cat algorithm, J. Supercomput. 81 (1) (2025) 113.
[2] Shihe Zhang, Ruidong Chen, Jingxue Chen, et al., L-GraphSAGE: A graph neural network-based approach for IoV application encrypted traffic identification, Electronics 13 (21) (2024) 4222.
[3] Kai Zhao, Shuoshuo Xu, James Loney, et al., Road pavement health monitoring system using smartphone sensing with a two-stage machine learning model, Autom. Constr. 167 (2024) 105664.
[4] Nan Deng, Zhengguang Xu, Xiuyun Li, et al., Deep learning and face recognition: Face recognition approach based on the DS-CDCN algorithm, Appl. Sci.- Basel 14 (13) (2024) 5739.
[5] GuiRong You, YeouRen Shiue, ChaoTon Su, et al., Enhancing ensemble diversity based on multiscale dilated convolution in image classification, Inform. Sci. 606 (2022) 292–312.
[6] Haizhu Pan, Hui Yan, Haimiao Ge, et al., Pyramid cascaded convolutional neural network with graph convolution for hyperspectral image classification, Remote. Sens. 16 (16) (2024) 2942.
[7] Luoxuan Chen, Chengchuang Lin, Zhaoliang Zheng, et al., Review of transformer in computer vision, Comput. Sci. 50 (12) (2023) 130–147.
[8] Lijuan Zhou, Jianning Mao, Vision transformer-based recognition tasks: A critical review, J. Image Graph. 28 (10) (2023) 2969–3003.
[9] Lei Zhu, XinjiangWang, Zhanghan Ke, et al., BiFormer: Vision transformer with bi-level routing attention, Comput. Vis. Pattern Recognit. (2023) 10323–10333.
[10] Hezheng Lin, Xing Cheng, Xiangyu Wu, et al., CAT: Cross attention in vision transformer, Comput. Vis. Pattern Recognit. (2022) 1–6.
[11] Qihang Fan, Huaibo Huang, Mingrui Chen, et al., Vision transformer with sparse scan prior, Comput. Vis. Pattern Recognit. (2024) 1–16.
[12] Qihang Fan, Huaibo Huang, Mingrui Chen, et al., Semantic equitable vlustering: A simple, fast and effective strategy for vision transformer, Comput. Vis. Pattern Recognit. (2024) 1–15.
[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, Comput. Vis. Pattern Recognit. (2021) 1–22.
[14] Weifeng Lin, Ziheng Wu, Jiayu Chen, et al., Scale-aware modulation meet transformer, Int. Conf. Comput. Vis. (2023) 6015–6026.

[15] Yanyu Li, Geng Yuan, Yang Wen, et al., Efficientformer: Vision transformers at mobilenet speed, Adv. Neural Inf. Process. Syst. 35 (2022) 12934–12949.
[16] Wang Zeng, Sheng Jin, Lumin Xu, et al., TCFormer: Visual recognition via token clustering transformer, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–16.
[17] Ting Yao, Yehao Li, Yingwei Pan, et al., HIRI-ViT: Scaling vision transformer with high resolution inputs, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–12.
[18] Meng Lou, Hong-Yu Zhou, Sibei Yang, et al., TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition, Comput. Vis. Pattern Recognit. (2023) 1–12.
[19] Jianyuan Guo, Kai Han, Han Wu, et al., CMT: convolutional neural networks meet vision transformers, Comput. Vis. Pattern Recognit. 12 (2022) 175–185.
[20] Mingzhe Li, Ningfeng Que, Juanhua Zhang, et al., VTCNet: A feature fusion DL model based on CNN and ViT for the classification of cervical cells, Int. J. Imaging Syst. Technol. 34 (5) (2024) e23161.
[21] Liwen Zhang, Rongwei Xia, Baiyang Yang, et al., WangMSFNet-2SE: A multi-scale fusion convolutional network for alzheimer′s disease classification on magnetic resonance images, Int. J. Imaging Syst. Technol. 34 (4) (2024) e23112.
[22] Guangzhe Zhao, Chen Zhang, Xueping Wang, et al., PMANet: Progressive multi-stage attention networks for skin disease classification, Image Vis. Comput. 149 (2024) 105166.
[23] Zhishuai Zheng, Zhedong Ge, Zhikang Tian, et al., WoodGLNet: A multi-scale network integrating global and local information for real-time classification of wood images, J Real- Time Image Proc 21 (4) (2024) 147.
[24] Zhikang Tiao, Zhedong Ge, Huanqi Zheng, et al., Microscopic identification methods for 75 types of hardwood based on deep neural network, Sci. Silvae Sin. 60 (10) (2024) 94–103.
[25] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 1–9.
[26] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, et al., RepVGG: Making vgg-style convnets great again, Comput. Vis. Pattern Recognit. (2021) 13733–13742.
[27] Christian Szegedy, Wei Liu, Yangqing Jia, et al., Going deeper with convolutions, Comput. Vis. Pattern Recognit. (2015) 1–9.
[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., Deep residual learning for image recognition, Comput. Vis. Pattern Recognit. (2016) 770–778.
[29] Chenhao Xu, Chang-Tsun Li, Chee Peng Lim, et al., HSViT: Horizontally scalable vision transformer, Comput. Vis. Pattern Recognit. (2024) 1–9.
[30] Qihang Fan, Huaibo Huang, Xiaoqiang Zhou, et al., Lightweight vision transformer with bidirectional interaction, Adv. Neural Inf. Process. Syst. 36 (2024) 1–17.
[31] Meng Lou, Hong-Yu Zhou, Sibei Yang, et al., TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition, Comput. Vis. Pattern Recognit. (2023) 1–12.
[32] Bencheng Liao, Wang Xinggang, Lianghui Zhu, et al., ViG: Linear-complexity visual sequence learning with gated linear attention, Comput. Vis. Pattern Recognit. (2024) 1–18.
[33] Benedikt Alkin, Maximilian Beck, Korbinian Poppel, et al., Vision-LSTM: xLSTM as generic vision backbone, Comput. Vis. Pattern Recognit. (2024) 1–18.
[34] Yuchen Duan, Weiyun Wang, Zhe Chen, et al., Vision-RWKV: Efficient and scalable visual perception with RWKV-like architectures, Comput. Vis. Pattern Recognit. (2024) 1–18.
[35] Qihang Fan, Huaibo Huang, Jiyang Guan, et al., Rethinking local perception in lightweight vision transformer, Comput. Vis. Pattern Recognit. (2024) 1–14.
[36] Weijie Zhang, Chang Li, Hu Peng, et al., CTCNet: A CNN transformer capsule network for sleep stage classification, Measurement 226 (2024) 114157.
[37] Guoan Xu, Juncheng Li, Guangwei Gao, et al., Lightweight real-time semantic segmentation network with ffficient transformer and CNN, IEEE Trans. Intell. Transp. Syst. 24 (12) (2023) 15897–15906.
[38] Jiahui Chen, Peng Lu, Xiaoling Luo, A hybrid algorithm for remote sensing image land cover classification combining CNN and ViT, Remote. Sens. Inf. 39 (03) (2024) 121–127.
[39] Shuoxi Zhang, Hanpeng Liu, Stephen Lin, et al., You only need lless attention at each stage in vision transformers, Comput. Vis. Pattern Recognit. (2024) 1–10.
[40] Wei Wang, Yujie Sun, Xin Wang, Lightweight frequency and spatial feature fused multi-scale remote sensing scene classification network, J. Jilin Univ. (Eng. Technol. Edition) (2024) 1–11.
[41] Ao Wang, Hui Chen, Zijia Lin, et al., RepViT: Revisiting mobile CNN from ViT perspective, Comput. Vis. Pattern Recognit. (2024) 15909–15920.
[42] Xu Ma1, Xiyang Dai, Yue Bai, et al., Rewrite the stars, Comput. Vis. Pattern Recognit. (2024) 5694–5703.
[43] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., Matching networks for one shot learning, Mach. Learn. 29 (2016) (2016) 1–12.
[44] Alex Krizhevsky, Geoffrey Hinton, et al., Learning multiple layers of features from tiny images, 2009, pp. 1–60.
[45] Alex Krizhevsky, Learning multiple layers of features from tiny images, Comput. Sci. (2020) 1–60.
[46] Ilya Loshchilov, Frank Hutter, DecoupledWeight decay regularization, in: International Conference on Learning Representations, 2019, pp. 1–19.
[47] Sucheng Ren, Xingyi Yang, Songhua, Liu, et al., SG-former: Self-guided transformer with evolving token reallocation, IEEE/CVF Int. Conf. Comput. Vis. (2023) 5980–5991.
[48] Jierun Chen, Shiu-hong Kao, Hao He, et al., Run, don′t walk: Chasing higher flops for faster neural networks, Comput. Vis. Pattern Recognit. (2023) 1–15.
[49] Xiangyong Lu, Masanori Suganuma1, Takayuki Okatani, et al., SBCFormer: Lightweight network capable of full-size ImageNet classification at 1 FPS on single board computers, Comput. Vis. Pattern Recognit. (2023) 1–11.
[50] Yulong Shi, Mingwei Sun, Yongshuai Wang, et al., EViT: An eagle vision transformer with bi-fovea self-attention, Comput. Vis. Pattern Recognit. (2023) 1–11.
[51] Kun Yuan, Shaopeng Guo, Ziwei Liu, et al., Incorporating convolution designs into visual transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 579–588.
[52] Laurens van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.