Information Fusion 105 (2024) 102248

Contents lists available at ScienceDirect

Information Fusion

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

Vision transformer: To discover the "four secrets" of image patches

Tao Zhou [a,d], Yuxia Niu [a,*], Huiling Lu [b], Caiyue Peng [a], Yujie Guo [a], Huiyu Zhou [c]

[a] School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China
[b] School of medical information & Engineering, Ningxia Medical University, Yinchuan 750004, China
[c] School of Computing and Mathematical Sciences, University of Leicester, United Kingdom
[d] Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China

ARTICLE INFO

Keywords:
Transformer
Patch division mechanism
Token selection mechanism
Position encoding mechanism
Attention mechanism

ABSTRACT

Vision Transformer (ViT) is widely used in the field of computer vision, in ViT, there are four main steps, which are "four secrets", such as patch division, token selection, position encoding addition, attention calculation, the existing research on transformer in computer vision mainly focuses on the above four steps. Therefore, "how to divide patch?", "how to select token?", "how to add position encoding?", and "how to calculate attention?" are crucial to improve ViT performance. But so far, most of the review literatures are summarized from the perspective of application, and there is no corresponding literature to comprehensively summarize these four steps from the technology perspective, which restricts the further development of ViT in some degree. To address the above questions, the 4 major mechanisms and 5 applications of ViT are summarized in this paper, the main innovative works are as follows: Firstly, the basic principle and model structure of ViT are elaborated; Secondly, aiming to "how to divide patch?", the 5 key techniques of patch division mechanism are summarized: from single-size division to multi-size division, from fixed number division to adaptive number division, from non-overlapping division to overlapping division, from semantic segmentation division to semantic aggregation division, and from original image division to feature map division; Thirdly, aiming to "how to select token?", the 3 key techniques of token selection mechanism are summarized: token selection based on score, token selection based on merge, token selection based on convolution and pooling; Fourthly, aiming to "how to add position encoding?", the 5 key techniques of position encoding mechanism are summarized: absolute position encoding, relative position encoding, conditional position encoding, locally-enhanced position encoding, and zero-padding position encoding; Fifthly, aiming to "how to calculate attention?", 18 attention mechanisms are summarized based on the timeline; Sixthly, these models that Transformer is combined with U-Net, GAN, YOLO, ResNet, and DenseNet are discussed in the medical image processing field; Finally, around these four questions proposed in this paper, we look forward to the future development direction of frontier technologies such as patch division mechanism, token selection mechanism, position encoding mechanism, and attention mechanism et al, which play an important role in the further development of ViT.

1. Introduction

Transformer [1] is an encoder-decoder architecture model based on self-attention mechanism, which is first applied in the field of Natural Language Processing(NLP). It not only can model long dependency among input sequence elements, but also supports parallel computation during training and inference, and has excellent performance in language modeling and machine translation tasks. Devlin et al. [2] proposed a masked bidirectional encoding structure Bert model based on Transformer, the model learns rich language representations through large-scale unsupervised training. In addition, many language models based on Transformer, Such as GPTv1 [3], GPTv2 [4], GPTv3 [5], Ro-BERTa [6], T5 [7] et al. are widely used in multiple language tasks.

With the rapid development of Transformer in the field of NLP, more and more researchers are attracted to the field of Computer Vision (CV). CV tasks usually deal with image or video data. There are inductive biases [8] in Convolutional Neural Networks (CNN), such as translation invariance and local sensitivity, which can capture image fine-grained features and image local features. However, the CNN-based method has the problem of limited receptive field, which makes it difficult to

* Corresponding authors.
E-mail address: niuyuxia@stu.nmu.edu.cn (Y. Niu).

https://doi.org/10.1016/j.inffus.2024.102248
Received 6 December 2023; Received in revised form 30 December 2023; Accepted 8 January 2024
Available online 11 January 2024
1566-2535/© 2024 Elsevier B.V. All rights reserved.

model the long dependency. Transformer obtains global representation through attention mechanism, which can model global features and long-distance correlation of input information. Therefore, Transformer are gradually applied in the CV fields. ViT [9] relies on its modeling capabilities to achieve excellent performance on some benchmark datasets, such as ImageNet [10], COCO [11], and ADE20k [12]. Over the past few years, hundreds of Transformer-based models are proposed for various tasks in the CV fields, such as classification [13], detection [14], segmentation [15], tracking [16], generation [17], and enhancement [18].

ViT makes breakthrough progress in the field of deep learning. At present, there are many literatures to review the application progress of ViT technology, Han et al. [19] reviewed vision transformer models according to different tasks (i.e., backbone network, high-level vision, mid-level vision, low-level vision, and video processing); Liu et al. [20] reviewed vision transformer models according to three fundamental CV tasks (i.e., classification, detection, and segmentation) and data stream types (i.e., image, point clouds, multi-stream data); Khan et al. [21] reviewed vision transformer models according to popular recognition tasks (e.g., image classification, object detection, action recognition, and segmentation), generative modeling, multi-modal tasks (e.g., visual-question answering, visual reasoning, and visual grounding), video processing (e.g., activity recognition, video forecasting), low-level vision (e.g., image super-resolution, image enhancement, and colorization), and 3D analysis (e.g., point cloud classification and segmentation).

In ViT, there are four main steps, which are “four secrets”, such as patch division, token selection, position encoding addition, attention calculation, the existing research on transformer in computer vision mainly focuses on the above four steps. Therefore, “how to divide patch?”, “how to select token?”, “how to add position encoding?”, and “how to calculate attention?” are crucial to improve ViT performance. But so far, most of the review literatures are summarized from the perspective of application, and there is no corresponding literature to comprehensively summarize these four steps from the technology perspective, which restricts the further development of ViT in some degree.

Therefore, aiming at the above problems, this paper makes a comprehensive summary of the 4 mechanisms and 5 applications of ViT. Firstly, the basic principle and model structure of ViT are summarized; Secondly, aiming to “how to divide patch?”, 5 key technologies of patch division mechanism are summarized: (1) from single-size division to multi-size division, (2) from fixed number division to adaptive number division, (3) from non-overlapping division to overlapping division, (4) from semantic segmentation division to semantic aggregation division, (5) from original image division to feature map division; Thirdly, aiming to “how to select token?”, 3 key technologies of token selection mechanism are summarized: (1) token selection based on score, (2) token selection based on merge, (3) token selection based on convolution and pooling; Fourthly, aiming to “how to add position encoding?”, 5 key technologies of position encoding mechanism are summarized: (1) absolute position encoding, (2) relative position encoding, (3) conditional position encoding, (4) locally-enhanced position encoding, (5) zero-padding position encoding; Fifthly, aiming to “how to calculate attention?”, 18 attention mechanisms are summarized based on the timeline; Sixthly, the extensive applications of ViT in medical image processing are discussed by combining with U-Net, GAN, YOLO, ResNet, and DenseNet; Finally, around these four questions proposed in this paper, we look forward to the future development direction of frontier technologies such as patch division mechanism, token selection mechanism, position encoding mechanism, attention mechanism, building a unified framework of multi-tasking, reducing high-dimensional data calculation, realizing small sample learning, ViT interpretability et al, which play an important role in the further development of ViT.

2. Basic principle of ViT

The ViT model uses the classic Transformer encoder structure to realize the image classification task, it is the beginning of the visual Transformer model. The model structure is shown in Fig. 1(left). Firstly, the input image is divided into non-overlapping patches in fixed size; Secondly, the patch is flattened into a one-dimensional vector in the channel dimension, and its corresponding token is obtained by linear mapping; Thirdly, an extra class token is added to the image token set, the class token is responsible for aggregating global image features and final classification; Fourthly, position embeddings are added to the tokens to retain position information; Finally, the vector sequences are fed into multiple serial Transformer encoders to calculate attention and extract feature.

2.1. Patch embedding

To convert the image into the input sequence of Transformer encoder:

Step 1: The 2D image $X \in R^{H\times W\times C}$ is divided into a series of patches $X_p \in R^{N\times(P^2\cdot C)}$, where (H, W) is the size of the original image, C is the number of channels, (P, P) is the size of each patch, and $N = \frac{H\cdot W}{P^2}$ is the number of patches;

Step 2: These patches are flattened in the channel dimension to obtain one-dimensional vectors x_p^i, where $x_p^i \in R^{N\times(P^2\cdot C)}$, i∈(1,... N), and then its corresponding token is obtained by linear mapping through the full connect network E. The process is shown in Fig. 2, which can be expressed as $z=[x_p^1E, x_p^2E,.., x_p^NE]$, where $E\in R^{(P^2\cdot C)\times D}$;

Step 3: For the classification task, an additional learnable embedding is inserted into the 0^{th} position of the input sequence, corresponding to the output token for subsequent classification tasks, by this operation, the sequence length increases from N to N+1: $z=[x_{class},x_p^1E,x_p^2E,..,x_p^NE]$;

Step 4: Position embeddings are added to the tokens to retain positional information, $z=[x_{class},x_p^1E,x_p^2E,..,x_p^NE]+E_{pos}$, where $E_{pos}\in R^{(N+1)\times D}$, z is used as the input sequence of the Transformer encoder.

2.2. Transformer encoder

The Transformer encoder is composed of L standard Transformer modules, each module is composed of Layer Norm (LN), Multi-Head Self-Attention (MHSA), Multi-Layer Perceptron (MLP), and residual connection. The feature calculation process is as follows:

$$Z_i^{'} = \mathrm{MHSA}(\mathrm{LN}(Z_{i-1})) + Z_{i-1} \tag{1}$$

$$Z_i = \mathrm{MLP}\left(\mathrm{LN}\left(Z_i^{'}\right)\right) + Z_i^{'} \tag{2}$$

Where, Formula (1) is that the input vector is processed by LN, MHSA, and residual connection; Formula (2) is that the previous step output is processed by LN, MLP, and residual connection.

2.3. Mutil-head self-attention

Multi-Head Self-Attention is the core part of Transformer encoder, and its main operation is the Self-Attention mechanism. As shown in Fig. 3. The Self-Attention mechanism reduces the dependence on external information and can better capture internal data correlation. First, three linear transformation matrices W_q,W_k,W_v are used to convert input vector X into three different matrices: Query matrix Q, key matrix K, and value matrix V, as shown in Formula (3), then the calculation process of Self-Attention is shown in Formula (4), where d_k is the dimension of the key vector k.

$$Q = X\cdot W_q,\ K = X\cdot W_k,\ V = X\cdot W_v \tag{3}$$

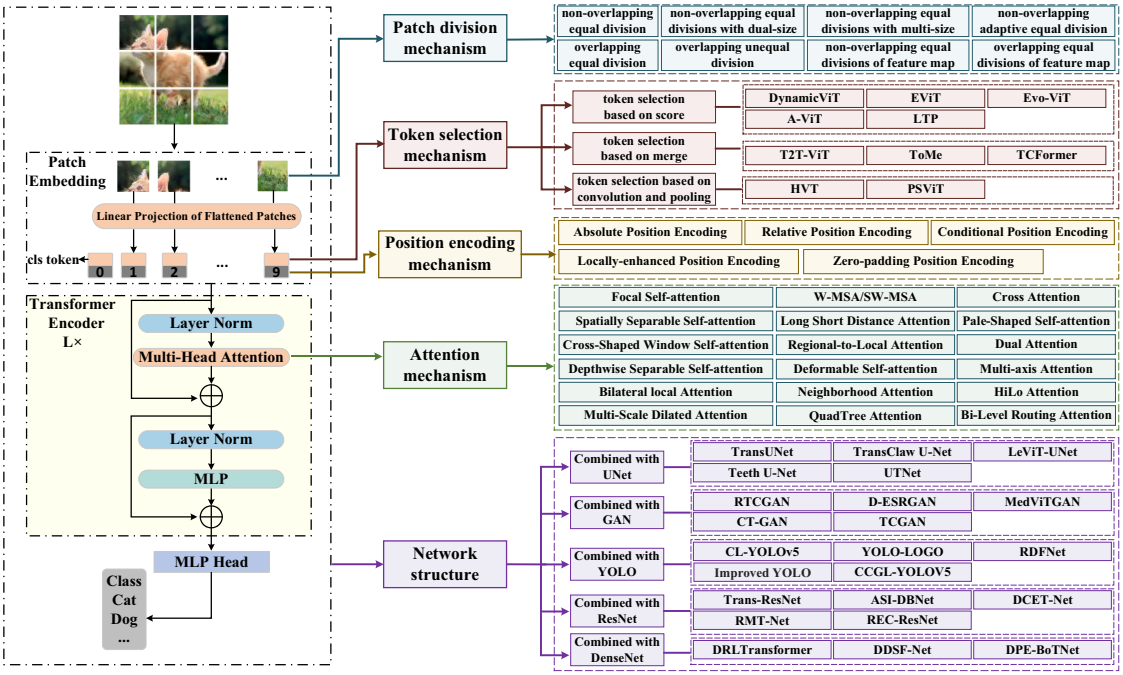


Fig. 1. The overall framework of ViT.

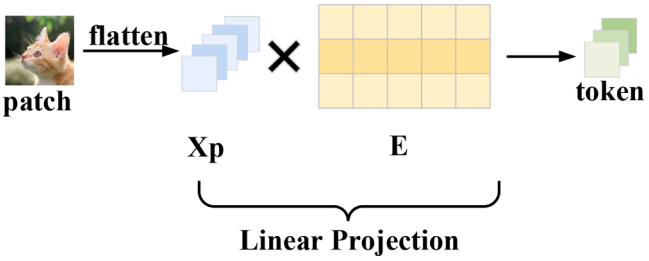


Fig. 2. Linear Projection of Flattened Patches.

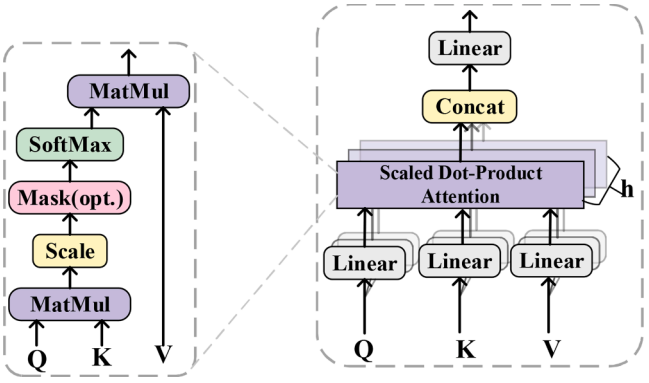


Fig. 3. Self-Attention and Multi-Head Self-Attention.

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

Multi-Head Self-Attention can be seen as an extended form of Self-Attention. By introducing multiple independent attention heads, it can better capture the associated information in the input sequence and provide a richer context representation. First, the input vectors are multiplied with multiple sets of parameter matricesW_i^Q, W_i^K, W_i^V to get multiple independent attention heads, the calculation process is shown in Formula (5), where h is the number of heads, then the outputs of multiple attention heads are concated and multiply with the trainable parameter matrix W^o, as shown in Formula (6).

$$\mathrm{head}_i=\mathrm{Attention}\left(Q\cdot W_i^Q, K\cdot W_i^K, V\cdot W_i^V\right),\ i=1,\ldots,h \tag{5}$$

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)\cdot W^o \tag{6}$$

3. Vision transfortmr improvement mechanism

The ViT model is a deep neural network that divides the input image into a series of non-overlapping patches in fixed size, where each patch is regarded as an independent feature vector, and then the context relationships are modeled between different patches by Multi-Head Self-Attention. 5 key technologies of patch division mechanism are summarized as following: from single-size division to multi-size division, from fixed number division to adaptive number division, from non-overlapping division to overlapping division, from semantic segmentation division to semantic aggregation division, and from original image division to feature map division; 3 key technologies of token selection mechanism are summarized as following: token selection based on score, token selection based on merge, token selection based on convolution and pooling; 5 key technologies of position encoding mechanism are summarized as following: absolute position encoding, relative position encoding, conditional position encoding, locally-enhanced position encoding, and zero-padding position encoding; 18 attention mechanisms are summarized based on the timeline; the extensive applications of ViT in medical image processing are discussed by combining with U-Net, GAN, YOLO, ResNet, and DenseNet.

3.1. Patch division mechanism

The patch division mechanism is that the image is divided into multiple patches for processing. By dividing the image into multiple patches, local features in the image can be better extracted, thus the efficiency and accuracy of image processing are improved. The traditional division method is that the original input images are divided into a series of non-overlapping patches in fixed size, as shown in Fig. 4(A).

This method is simple and direct, but it destructs the image local-

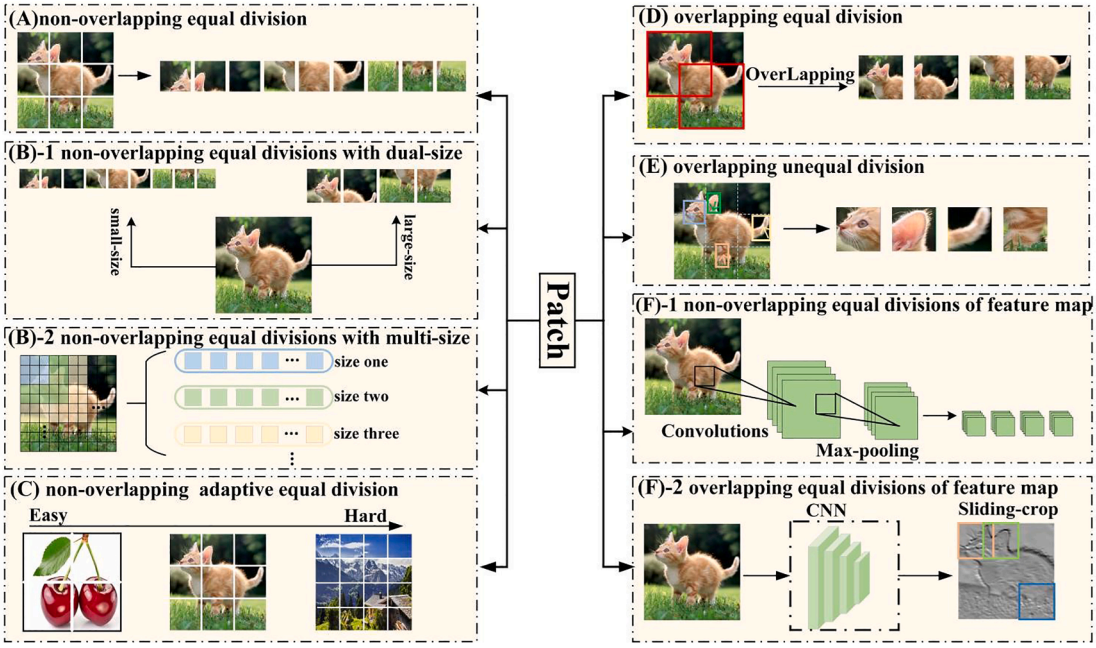


Fig. 4. Patch division mechanism.

continuity in some degree, and also limits the ViT performance in visual tasks. In order to resolve the problem, there are many researches on the patch division mechanism, which include 5 aspects: from single-size division to multi-size division, from fixed number division to adaptive number division, from non-overlapping division to overlapping division, from semantic segmentation division to semantic aggregation division, and from original image division to feature map division.

3.1.1. From single-size division to multi-size division

From single-size division to multi-size division is that the images are divided according to multiple sizes rather than a single size. Computer Vision tasks, such as detecting and classifying, require that effective multi-size feature representation. Therefore, the multi-size feature representation is obtained by dividing the image into multiple sizes. In the case of non-overlapping equal division, there are non-overlapping equal divisions with dual-size and non-overlapping equal divisions with multi-size in patch division. The non-overlapping equal divisions with dual-size is that coarse-grained large-size patches and fine-grained small-size patches are extracted by two independent branches. This idea is proposed by Chen et al. [22], in the idea, stronger image features are produced by combining patches of different sizes, as shown in Fig. 4 (B)-1. The non-overlapping equal divisions with multi-size is that patches of different sizes are obtained by multiple independent branches. This idea is proposed by Lee et al. [23], in the idea, obtained features are aggregated by independently inputting patches of different sizes into Transformer encoder, and multi-size feature representation is realized at the same feature level, as shown in Fig. 4(B)-2.

3.1.2. From fixed number division to adaptive number division

From fixed number division to adaptive number division is that the images are divided into patches in a self-adaptive number rather than a fixed number. In general, these problems are generated by more patches, such as higher prediction accuracy and more computational complexity. Therefore, in order to realize a trade-off between prediction accuracy and computational complexity, self-adaptive number patches are very important for patch division mechanism. This idea is proposed by Wang et al. [24], in the idea, multiple Transformers with increasing numbers of patches are cascaded, during the test, the model is activated sequentially starting from fewer patches, once a sufficiently confident prediction is produced, the inference is terminated immediately. as shown in Fig. 4(C). For "easy" images, only 2×2 patches are satisfied for accurate prediction, while for "hard" images, fine-grained representations are required to reduce information loss and improve computational efficiency.

3.1.3. From non-overlapping division to overlapping division

From non-overlapping division to overlapping division is that patch division mechanism is changed from non-overlapping division to overlapping division. Non-overlapping division destructs the image local-continuity in some degree. Therefore, overlapping division of images can enhance the semantic correlation between neighboring patches and effectively solve the problem of non-overlapping division destructs the feature local-continuity. This idea is proposed by Wang et al. [25], in the idea, the image is serialized by overlapping and equal division, so that the neighboring patch overlaps half of the area. In this way, more local continuity of image features is contained. As shown in Fig. 4(D).

3.1.4. From semantic segmentation division to semantic aggregation division

From semantic segmentation division to semantic aggregation division is that the semantically related target local structures are aggregated in a patch. Since the complete local structure of the target object is captured by regular patch is always difficult, so the images are adaptively divided into patches of different positions and sizes can effectively capture the complete local structure of the target object. This idea is proposed by Chen et al. [26], in the idea, the offset and size of each patch are learned according to the input visual features, and the images are divided into patches with different positions and sizes in a deformable way, which preserves the semantic information in each patch well and reduces the semantic destruction caused by image segmentation, as shown in Fig. 4(E).

3.1.5. From original image division to feature map division

From original image division to feature map division is that the generated feature maps are divided into patches instead of the original image, so that more semantic information is contained in each obtained patch. The division of feature maps includes the non-overlapping equal division of feature maps and the overlapping equal division of feature

maps. The non-overlapping equal division of feature maps is that the convolution layer and the max pooling layer are adopted to obtain the feature maps of original image, and then the feature maps are divided into non-overlapping patches. This idea is proposed by Yuan et al. [27], in the idea, the advantage of CNN in extracting low-level features is utilized, and the training difficulty of embedding is decreased by reducing the patch size, as shown in Fig. 4(F)-1. The overlapping equal division of feature maps is that pre-trained CNN is adopted to extract the intermediate convolutional feature map from the original image, and then the intermediate feature maps are divided into overlapping patches by sliding window. This idea is proposed by Liu et al. [28], in the idea, a $p \times p$ window slides at stride s, where $0 < s \leq p$, which helps to retain more edge information, as shown in Fig. 4(F)-2.

3.2. Token selection mechanism

The one-dimensional vector obtained by linear mapping after the patch flattening operation in the channel dimension is called the token. The token selection mechanism is that redundant tokens are dynamically identified and tokens of higher importance are selected in the forward propagation process of the model. Considering that when using Transformer to solve visual tasks, the calculated amount of ViT increases exponentially as the number of tokens increases, and the final prediction result of ViT is often determined by some tokens with large amounts of information, most of tokens are redundant. Therefore, if the redundant tokens can be dynamically identified according to the input image and tokens containing important information are selected in the process of forward propagation, the reasoning speed of the ViT model can be greatly accelerated. There are many researches on token selection mechanism, which include 3 aspects: The first type is the token selection mechanism based on score, as shown in Fig. 5(A), with typical models DynamicViT, EViT, Evo-ViT, A-VIT, and LTP; The second type is the token selection mechanism based on merge, as shown in Fig. 5(B), with typical models T2T-ViT, ToMe, and TCFormer; The third type is the

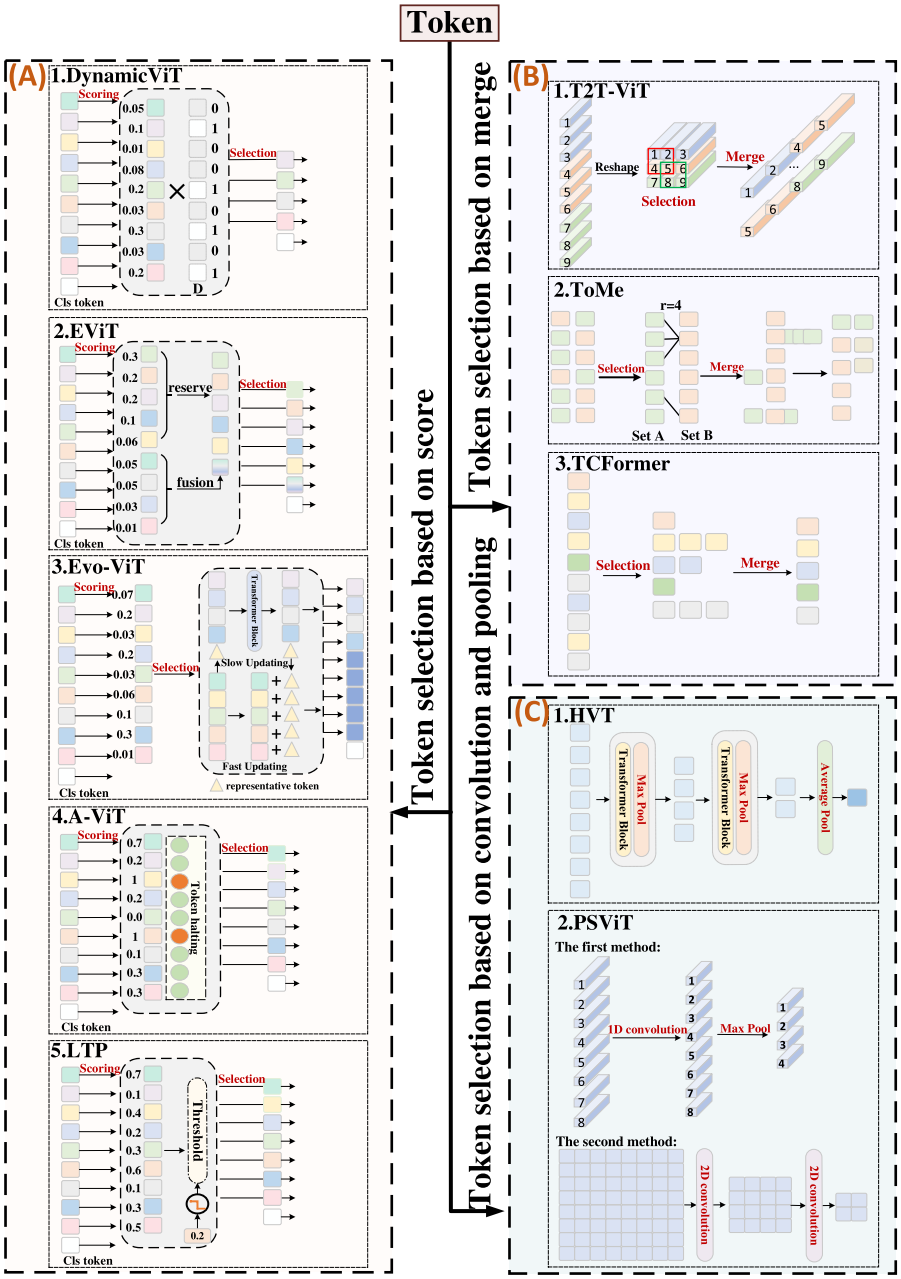


Fig. 5. Token selection mechanism.

token selection based on convolution and pooling, as shown in Fig. 5(C), with typical models HVT and PSViT. Meanwhile, the performances of different token selection mechanisms are compared from 5 aspects: Params, FLOPs, Throughput, Dataset, and Top-1 acc, as listed in Table 1.

3.2.1. Token selection based on score

The token selection mechanism based on score is that the importance of tokens is scored by the scoring function, high-score tokens are retained through pruning operation, and low-score tokens are deleted, so as to reduce the calculation amount and improve the calculation efficiency of the model. The key operations in this method are scoring strategy and selection strategy. At present, there are 5 major scoring strategies and selection strategies: DynamicViT strategy, EViT strategy, Evo-ViT strategy, A-ViT strategy, and LTP strategy.

(1) DynamicViT strategy. This strategy is proposed by Rao et al. [29], as shown in Fig. 5(A)-1, the performance is improved by layer pruning on 66% of the input tokens, such as the flops is reduced by 31–37%, and running speed is increased by more than 40% in the model. In the scoring strategy section: Firstly, a binary decision mask D is initialized, and all its element values are set to 1; Secondly, the local feature is obtained by using an MLP to project the tokens, as shown in Formula (7); Thirdly, the global feature is computed by using an Agg function to aggregate the information of all the existing tokens, as shown in Formula (8); Finally, the local features and global features are concated in the channel dimension and fed into another MLP to predict the probabilities of drop/keep the tokens, as shown in Formula (9), the result P is the importance score for each token; In the selection strategy section: according to the importance score of each token, the binary mask D is updated to select tokens, In mask D, if the value of the element in D is 0, its corresponding token is deleted, if the value of the element in D is 1, its corresponding token is retained.

$$Z^{local} = \mathrm{MLP}(x) \tag{7}$$

$$Z^{global} = \mathrm{Agg}(\mathrm{MLP}(x), D),$$

$$\text{where } \mathrm{Agg}(u, D) = \frac{\sum_{i=1}^{N} D_i u_i}{\sum_{i=1}^{N} D_i} \tag{8}$$

$$P = \mathrm{softmax}\left(\mathrm{MLP}\left(\left[Z_i^{local}, Z_i^{global}\right]\right)\right) \tag{9}$$

(2) EViT strategy. This strategy is proposed by Liang et al. [30], as shown in Fig. 5(A)-2, this strategy is applied to DeiT-S, and the reasoning speed of DeiT-S is improved by 50% in ImageNet, while the recognition accuracy is decreased by only 0.3%. In the scoring strategy section: the attention value a_i between the class token and other tokens is obtained by performing the attention mechanism, and the attention value represents the importance score for each token, as shown in Formula (10), where q_{class}, K and d represent the query vector of the class token, the key matrix, and the dimension of k vector respectively, N is the number of image tokens; In the selection strategy section: top-k high-score tokens are selected, other tokens are merged into a new token through weighted average operation, and the new token is appended to the top-k tokens and sent to the subsequent layer.

$$a_i = \mathrm{softmax}\left(\frac{q_{class} \cdot K^T}{\sqrt{d}}\right), \text{ where } i \in (1, 2, ..., N) \tag{10}$$

(3) Evo-ViT strategy. This strategy is proposed by Xu et al. [31], as shown in Fig. 5(A)-3, this strategy is applied to DeiT-S, and the throughput of DeiT-S is improved by 60% in ImageNet, while the recognition accuracy is decreased by only 0.4%. In the scoring strategy section: the similarity between the class token and image tokens is represented as class attention A_{cls}, as shown in Formula (11), where q_{cls}, K and d represent the query vector of the class token, the key matrix, and the dimension of k vector respectively, and A_{cls} is the importance score for each token; In the selection strategy section: low-score tokens are aggregated into a representative token, top-k high-score tokens and the representative token are updated through Transformer block, meanwhile, low-score tokens are weighted separately with updated representative tokens, the weighted low-score tokens are appended to the top-k tokens and sent to the subsequent layer.

$$A_{cls} = \mathrm{softmax}\left(\frac{q_{cls} \cdot K^T}{\sqrt{d}}\right) \tag{11}$$

(4) A-ViT strategy. This strategy is proposed by Yin et al. [32], as shown in Fig. 5(A)-4, this strategy is applied to Deit-Tiny and Deit-Small, the speed of Deit-Tiny and Deit-Small are improved by 62% and 38% respectively, while the accuracy is decreased by only 0.3%. In the scoring strategy section: the h_k^l is introduced for each token as the halting probability of token k at layer l, as shown in Formula (12), where H(.) is a halting module, h_k^l range is [0,1]; In the selection strategy section: as enter deeper layers, the cumulative halting probability is used according to the output of the halting module, and when the cumulative halting score exceeds 1, the calculation of tokens is halted.

$$h_k^l = H\left(t_k^l\right) \tag{12}$$

(5) LTP strategy. This strategy is proposed by Kim et al. [33], as shown in Fig. 5(A)-5. In the scoring strategy section: firstly, the attention probability of head h between token x_i and token x_j is

Table 1
Performance comparison of different token selection mechanisms.

No.	Strategy	Model	Params(M)	FLOPs(G)	Throughput (images/s)	Dataset	Top-1 acc
1	DynamicViT[29]	DynamicViT-LV-S/0.5	26.9	3.7	-	ImageNet-1K	82.0
2	EViT[30]	EViT-LV-S/0.5	-	3.9	3603	ImageNet-1K	82.5
3	Evo-ViT[31]	Evo-LeViT-256/0.5	19.0	-	4277	ImageNet-1K	78.8
4	A-ViT[32]	A-ViT-S	22	3.6	1.1k	ImageNet-1K	78.6
5	LTP[33]	LTP	*				
6	T2T-ViT[34]	T2T-ViT-14	21.5	4.8	-	ImageNet-1K	81.5
7	ToMe[35]	ToMe-DeiT/r_{13}	-	2.7	1552	ImageNet-1K	79.4
8	TCFormer[36]	TCFormer	25.6	5.9	-	ImageNet-1K	82.4
9	HVT[37]	HVT-S	22.09	2.4	-	ImageNet-1K	78.0
10	PSViT[38]	PSViT-1D-Base	-	18.9	-	ImageNet-1K	82.6
		PSViT-2D-Base	-	15.5	-	ImageNet-1K	82.9

Notes: "*" indicates that the model was not experimented on ImageNet-1K, so it is not given here,
"-" Indicates that the data is not given in the original paper.

obtained, as shown in Formula (13), then the importance score of token x_i in layer l is computed, as shown in Formula (14), where N_h is the number of heads, n is the number of tokens; In the selection strategy section: tokens with scores are lower than the learnable threshold are pruned at each layer.

$$A^{(h,l)}(x_i, x_j) = \text{softmax}\left(\frac{x^T W_q^T W_k x}{\sqrt{d}}\right)_{(i,j)} \tag{13}$$

$$S^{(l)}(x_i) = \frac{1}{N_h}\frac{1}{n}\sum_{h=1}^{N_h}\sum_{j=1}^{n} A^{(h,l)}(x_i, x_j) \tag{14}$$

3.2.2. Token selection based on merge

The token selection mechanism based on merge is that similar tokens are selected and combined by matching algorithm, so as to reduce the information loss and improve the training speed. The key operations in this method are selection strategy and merge strategy. At present, there are 3 major selection strategies and merge strategies: T2T-ViT strategy, ToMe strategy, and TCFormer strategy.

(1) T2T-ViT strategy. This strategy is proposed by Yuan et al. [34], as shown in Fig. 5(B)-1. In this strategy, the neighboring tokens are recursively aggregated into one token, such that local structure can be modeled of neighboring tokens and token length can be reduced. In the selection strategy section: all tokens are reshaped as an image in the spatial dimension, and then split into patches with overlapping; In the merge strategy section: the tokens in each split patch are concatenated as one new token.

(2) ToMe strategy. This strategy is proposed by Bolya et al. [35], as shown in Fig. 5(B)-2. In this strategy, a large number of redundant tokens are merged, and the model training and reasoning speed are greatly improved. In the selection strategy section: tokens are divided into two sets A and B, which these two sets are roughly equal in size, and then the dot product similarity is used to select the r tokens from the B set that are most similar to the A set; In the merge strategy section: the binary soft matching algorithm is used to merge the most similar r tokens.

(3) TCFormer strategy. This strategy is proposed by Zeng et al. [36], as shown in Fig. 5(B)-3. In this strategy, the important areas are focused on to capture details. In the select strategy section: the K-nearest-neighbor based density peaks clustering algorithm is used to group tokens into a certain number of clusters; In the merge strategy section: tokens in the same cluster are merged by weighted average.

3.2.3. Token selection based on convolution and pooling

The token selection mechanism based on convolution and pooling is that the tokens' sequence length is shrunk by convolution and pooling operations, similar to feature maps downsampling in CNN, so as to reduce the redundant information and computing costs. The key operations in this method are convolution strategy and pooling strategy. At present, there are 2 major convolution strategies and pooling strategies: HVT strategy, and PSViT strategy.

(1) HVT strategy. This strategy is proposed by Pan et al. [37], as shown in Fig. 5(C)-1. In this strategy, the tokens are hierarchical pooling to shorten sequence length. In the pooling strategy section: Firstly, the ViT blocks are divided into several stages; secondly, At each stage, a max pooling layer is inserted after the Transformer block to perform down-sampling; finally, the average pooling is performed on tokens in the last stage for the final result prediction.

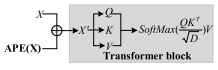
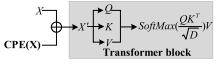
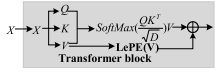
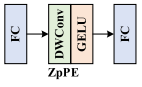
(2) PSViT strategy. This strategy is proposed by Chen et al. [38], as shown in Fig. 5(C)-2. In this strategy, there are 2 methods to reduce the number of tokens to eliminate spatial redundancy. In the first method, a 1D convolution with a small kernel size is used to change the dimension of each token and then a 1D maxpooling is used to decrease the number of tokens; In the second method, a 2D convolutional layer with stride 2 is adopted for token down-sampling, which is widely applied in many convolutional networks.

3.3. Position encoding mechanism

Since position information is not taken into account in the process of self-attention calculation, and the relationship among data is affected by position information, in order to enable the model to accept the position information of the input image, the position encoding mechanism is introduced. The position encoding mechanism is that the position information is integrated into the input sequence. By capturing the position information of the input sequence and maintaining the spatial position relationship among these sequences, the position information in the image can be effectively expressed, thus the performance of the model is improved. The sequence encoded by the position information can be input in parallel, and the computational efficiency is greatly improved. There are 5 typical Position Encoding mechanisms: Absolute Position Encoding, Relative Position Encoding, Conditional Position Encoding, Locally-enhanced Position Encoding, and Zero-Padding Position Encoding. Meanwhile, different position encoding mechanisms are compared, as listed in Table 2.

Firstly, Absolute Position Encoding, Absolute Position Encoding is generated by predefined functions or learned by training. Its dimension is the same as the input sequence, and the position information is added to the input sequence by an add operation. In ViT, the encoding method is generated by sine and cosine functions of different frequencies, as shown in Formulas (15) and (16), where *pos* represents the position of each element in the sequence, and *d* represents the dimension of the position encoding, which is consistent with the input dimension. *2i* represents the even dimension of the position encoding, *2i+1* represents the odd dimension of the position encoding, and the value range of i is [0,1,...,d/2). The disadvantage of this method is that the serialization length is fixed, which cannot be handled when facing high-resolution

Table 2
Comparison among different positional encoding mechanisms.

Position Encoding	Model	Position	Diagram
Absolute Position Encoding (APE)	ViT[9]	Introducing the positional information before feeding into the Transformer blocks	
Relative Position Encoding (RPE)	Swin [39]	Introducing the positional information in each Transformer block	
Conditional Position Encoding (CPE)	CPVT [40]	Introducing the positional information before feeding into the Transformer blocks	
Locally-enhanced Position Encoding (LePE)	CSWin [41]	Introducing the positional information in each Transformer block	
Zero-padding Position Encoding (ZpPE)	PVT v2 [25]	Introducing the positional information in feed-forward networks	

images during testing.

$$PE_{(pos,2i)} = \sin\left(pos \,/\, 10000^{2i/d}\right) \tag{15}$$

$$PE_{(pos,2i+1)} = \cos\left(pos \,/\, 10000^{2i/d}\right) \tag{16}$$

Secondly, Relative Position Encoding, unlike absolute position encoding, which directly adds position information to the input tokens sequence, the relative distance between the current position and the attended position is taken into account by relative position encoding. The relative positions between element pairs are calculated and encoded as part of the model input, and spatial relationships among different elements can be learned. In Swin [39], a relative position bias $B \in R^{M^2 \times M^2}$ is added to the attention calculation process, as shown in Formula (17), where Q, K, $V \in R^{M^2 \times d}$, d is the dimension of k, M^2 is the number of patches in a window, since the relative positions along each axis are in the range [-M+1, M-1], parameterizing a bias matrix $\overline{B} \in R^{(2M-1)\times(2M-1)}$, where the values in B are taken from $\overline{B}$. Relative position encoding is translation invariant and can naturally handle longer token sequences during training.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}} + B\right) \cdot V \tag{17}$$

Thirdly, Conditional Position Encoding [40], unlike fixed or learnable position encodings, conditional position encoding is predefined and dynamically generated by a simple Positional Encoding Generator (PEG) based on the input's local neighborhood. In order to take the local neighborhood as the condition, the flattened input token sequence $X \in R^{N \times C}$ is reshaped into a three-dimensional image space $X' \in R^{H \times W \times C}$, and then the function F is repeatedly applied in the local patch of X', and then the conditional position encoding $E^{B \times H \times W \times C}$ is generated. PEG can be effectively implemented by a 2D convolution operation with a kernel size of k(k≥3) and zero-padding of $\frac{k-1}{2}$, where F can be multiple forms, such as depth convolution, separable convolution, or others. Conditional position encoding not only is easily applied to a longer input token sequence than the model in the training process, but also maintains the translation invariance, and the classification accuracy is improved in image classification tasks.

Fourthly, Locally-enhanced Position Encoding [41], Locally-enhanced Position Encoding proposed that the position encoding is added to self-attention operations as a parallel module. This design decouples positional encoding from the self-attention calculation process, and local inductive bias is enforced stronger. The value V of position information is learned directly through deep convolution operation and acts as a parallel module as shown in Formula (18). Locally-enhanced Position Encoding can better process local position information and support input images with arbitrary resolution, which can be effectively used in some application fields.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V + \text{DWConv}(V) \tag{18}$$

Fifthly, Zero-padding Position Encoding, studies [42] have shown that position information can be learned implicitly from Zero-padding in CNN, Zero-padding position encoding is that a deep convolution with kernel size 3×3 is added to learn position information between the first fully connected layer in Feedforward Neural Network and the GELU activation function. Zero-padding position encoding is introduced to expand different-size images into uniform-size images, so that the model can flexibly deal with variable-resolution input images.

3.4. Attention mechanism

The idea of attention was first proposed in the field of image recognition by Mnih et al. [43] in 2014, which is a model to simulate the attention mechanism of the human brain. As the core component of ViT, self-attention is a powerful tool to capture long-distance dependency relationships. However, the global information interactions of all pairs of patches in the spatial position are calculated by the global attention in the original ViT model, as shown in Fig. 6-1, hence, high time complexity and space complexity are resulted, especially in high-resolution vision tasks. There are a series of works to improve attention performance and reduce computing and storage costs. Based on the timeline, 18 attention mechanisms are summarized in this paper, as shown in Fig. 6. The main attention mechanisms include: Focal Self-attention, W-MSA and SW-MSA, Cross-Attention, Spatially Separable Self-Attention, Long Short Distance Attention, Pale-Shaped Self--attention, Cross-Shaped Window Self-Attention, Regional-to-Local Attention, Dual Attention, Depthwise Separable Self-attention, Deformable Attention, Multi-axis Attention, Bilateral Local Attention, Neighborhood Attention, HiLo Attention, Multi-Scale Dilated Attention, QuadTree Attention, and Bi-Level Routing. Meanwhile, the performances of 18 attention mechanisms are compared from 5 aspects: Params, FLOPs, Throughput, Dataset, and Top-1 acc, as listed in Table 3.

Firstly, Focal Self-attention(FSA), FSA is an attention mechanism that incorporates both fine-grained local features and coarse-grained global features, where each token focuses on the nearest tokens around it at fine granularity and focuses on the long-distance tokens at coarse granularity, as shown in Fig. 6-2. Yang et al. [44] considered that the visual dependence among neighbor regions is often stronger than that among non-neighbor regions, and proposed FSA to model local and global interaction for high-resolution prediction tasks, which can effectively capture the visual dependence between short and long distances.

Secondly, W-MSA and SW-MSA, W-MSA represents Windows Multi-head Self-Attention, in layer l, the feature maps are divided into multiple non-overlapping windows, and self-attention is performed in each window, as shown in Fig. 6-3(left); SW-MSA represents Shifted Windows Multi-Head Self-Attention, in the next layer $l+1$, the windows, that is divided in layer l, are offset by $\frac{M}{2}$ pixels from the upper left corner to the right side and the lower side respectively, where M is the window size, so that information is transmitted in neighboring Windows. as shown in Fig. 6-3(right); Liu et al. [39] proposed W-MSA and SW-MSA, connections among neighboring non-overlapping Windows in the previous layer are introduced, and modeling capabilities are significantly enhanced.

Thirdly, Cross-Attention(CA), since the semantic information among all image tokens is already learned by the class token in its own branch, the interaction can fuse information in different scales between the class token at one branch and the image tokens at the other branch. as shown in Fig. 6-4. Chen et al. [22] proposed CA to realize the information exchange between two branches, and fuse multi-scale features more effectively.

Fourthly, Spatially Separable Self-Attention(SSSA), SSSA is composed of locally-grouped self-attention (LSA) and global sub-sampled attention (GSA). In LSA, the feature maps are equally divided into sub-windows, and self-attention calculations are performed within each sub-window; In GSA, a single representative is used to summarize the important information for each of m×n sub-windows, and the representative is used to communicate with other sub-windows, as shown in Fig. 6-5. Chu et al. [45] proposed SSSA to capture fine-grained short-range information and to process global information over long distances.

Fifthly, Long Short Distance Attention(LSDA), LSDA is composed of Short Distance Attention(SDA) and Long Distance Attention(LDA). SDA groups neighboring embeddings to establish dependencies, while LDA groups remote embeddings to establish dependencies by interval sampling, as shown in Fig. 6-6. Wang et al. [46] proposed LSDA to establish cross-scale interaction ability on the basis of not undermining either small-scale or large-scale features.

Sixthly, Pale-Shaped Self-attention(PSA), In PSA, the input feature

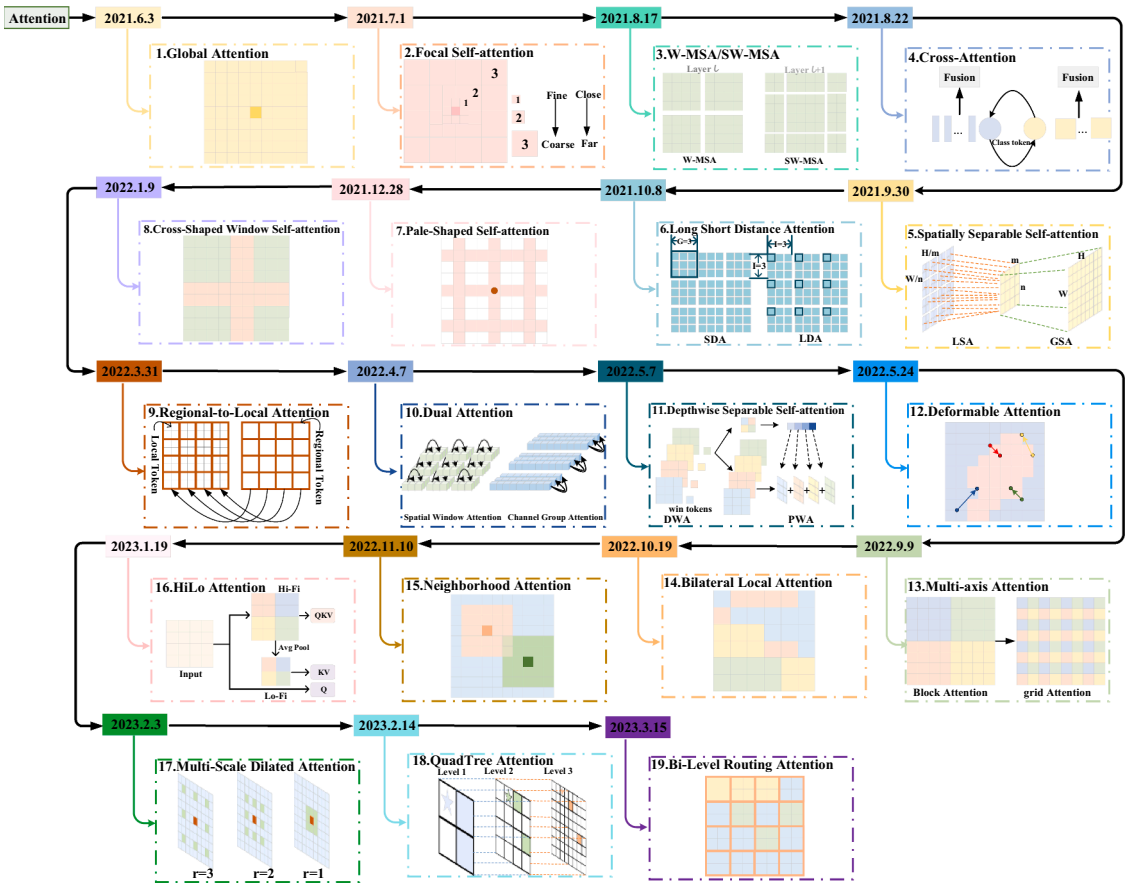


Fig. 6. Attention mechanism.

Table 3
Performance comparison of 18 attention mechanisms.

No.	Model	Attention	Params(M)	FLOPs(G)	Throughput (images/s)	Dataset	Top-1 acc
1	ViT(Base)[9]	Global Attention	86	55.4	85.9	ImageNet-1K	77.9
2	Focal Transformer (Base)[44]	Focal Self-attention	89.8	16.0	-	ImageNet-1K	83.8
3	Swin Transformer (Base)[39]	W-MSA,SW-MSA	88	15.4	278.1	ImageNet-1K	83.5
4	CrossViT (Base)[22]	Cross-Attention	104.7	21.2	239	ImageNet-1K	82.2
5	Twins (Base)[45]	Spatially Separable Self-attention	56	8.6	469	ImageNet-1K	83.2
6	CrossFormer (Base)[46]	Long Short Distance Attention	52.0	9.2	-	ImageNet-1K	83.4
7	Pale Transformer (Base)[47]	Pale-Shaped self-Attention	85	15.6	-	ImageNet-1K	84.9
8	CSWin Transformer (Base)[41]	Cross-Shaped Window Self-attention	78	15.0	250	ImageNet-1K	84.2
9	RegionVIT (Base)[48]	Regional-to-Local Attention	72.7	13.0	-	ImageNet-1K	83.2
10	DaViT (Base)[49]	Dual Attention	87.9	15.5	-	ImageNet-1K	84.6
11	SepViT (Base)[50]	Depthwise Separable Self-attention	82.3	13.1	308	ImageNet-1K	84.0
12	DAT (Base)[51]	Deformable Attention	88	15.8	-	ImageNet-1K	84.0
13	MaxViT (Base)[52]	Multi-axis Attention	120	23.4	133.6	ImageNet-1K	84.95
14	BOAT (Base)[53]	Bilateral Local Attention	90	17.5	-	ImageNet-1K	84.7
15	NAT (Base)[54]	Neighborhood Attention	90	13.7	783	ImageNet-1K	84.3
16	LITv2 (Base)[55]	HiLo Attention	87	13.2	602	ImageNet-1K	83.6
17	DilateFormer (Base)[56]	Multi-Scale Dilated Attention	47	10.0	-	ImageNet-1K	84.4
18	QuadTree (Base)[57]	QuadTree Attention	64.2	11.5	-	ImageNet-1K	84.0
19	BiFormer (Base)[58]	Bi-Level Routing Attention	57	9.8	-	ImageNet-1K	84.3

*Notes: Except for ViT-Base, all other models are trained and evaluated on 224 × 224 resolution.
"-" Indicates that the data is not given in the original paper.

maps are split into multiple pale-shaped regions, and the attention calculations are performed within each region, where each region consists of the same number of interleaving rows and columns in the feature map, and the spacing between neighboring rows or columns is equal, as shown in Fig. 6-7. Wu et al. [47] proposed PSA to enable any token to interact directly with other tokens in the same pale, and richer contextual information is captured.

Seventhly, Cross-Shaped Window Self-Attention(CSWin), In CSWin, the self-attention is performed in a cross-shaped window in parallel, where each window consists of horizontal and vertical stripes, and each stripe obtained by splitting the input feature into stripes of equal width, as shown in Fig. 6-8. Dong et al. [41] proposed CSWin, with the deepening of the network, the window width is increased to associate more areas, and global self-attention is achieved more effectively by expanding the attention area.

Eighthly, Regional-to-Local Attention(R2L), In R2L, the regional

tokens and local tokens are generated by using different patch sizes from images, where each regional token is associated with a set of local tokens based on the spatial position. The global information is extracted by regional self-attention among all regional tokens, and then the information is exchanged by local self-attention among one regional token and the associated local tokens, as shown in Fig. 6-9. Chen et al. [48] proposed R2L, which can still receive global information although the scope is limited to a local region.

Ninthly, Dual Attention(DA), DA is composed of spatial window attention and Channel group attention. In spatial window attention, the feature maps are divided into different Windows in the spatial dimension, and self-attention calculation is performed in the spatial window; In Channel group attention, the tokens are grouped into multi-groups in the channel dimension, and self-attention calculation is performed in each channel group, as shown in Fig. 6-10. Ding et al. [49] proposed DA, the local features are extracted within the window by spatial window attention, and the global features are learned by channel group attention, so as to efficient global modeling is realized through the alternate use of two kinds of attention.

Tenthly, Depthwise Separable Self-attention, Depthwise Separable Self-attention is composed of Depthwise Self-Attention(DWA) and Pointwise Self-Attention(PWA). In DWA, the feature maps are divided into different Windows, and a window token is created to serve as a global representation for each window, attention is calculated for each window and its corresponding window token; In PWA, the feature maps and window tokens are extracted from the output of DWA, and the window token is used to model the attention relationship between Windows, as shown in Fig. 6-11. Li et al. [50] designed Depthwise Separable Self-attention inspired by depthwise separable convolution to capture long-distance visual dependencies of multiple windows by promoting information interaction within and between Windows.

Eleventhly, Deformable Attention(DA), In DA, a set of reference points are uniformly generated on the feature map, and the corresponding offsets for all reference points are generated by the offset network, so that key-value pairs are moved to important areas, which enhances the flexibility and efficiency of the self-attention module, and thus more information features are captured. as shown in Fig. 6-12. Xia et al. [51] proposed DA to select the position of key-value pairs in a data-dependent manner, focus on relevant areas, and capture more information.

Twelfthly, Multi-axis Attention(MaxA), MaxA is composed of Block Attention and Grid Attention. In Block Attention, the feature maps are divided into non-overlapping local Windows, and self-attention is performed within each window; In Grid Attention, the feature maps are divided into fixed grid blocks, and self-attention is performed at corresponding positions in each grid block, as shown in Fig. 6-13. Tu et al. [52] proposed MaxA to sequentially superimpose two types of attention, and allow global-local spatial interaction at any input resolution.

Thirteenthly, Bilateral Local Attention(BOA), In BOA, a Balanced hierarchical clustering method is used to divide patches into multiple clusters of uniform size and self-attention is calculated within each cluster, as shown in Fig. 6-14. Yu et al. [53] proposed BOA, local attention in feature space and local attention in image space are combined, and it effectively captures the connection between the far apart but related patches in the image.

Fourteenthly, Neighborhood Attention(NA), NA localizes attention to a neighborhood around each token, introducing local inductive biases, maintaining translational equivariance, and allowing receptive field growth, as shown in Fig. 6-15. Hassani et al. [54] proposed NA introduces local inductive bias to maintain translation invariance.

Fifteenthly, HiLo Attention, which consists of High-frequency Attention(Hi-Fi) and Low-frequency attention(Lo-Fi). In Hi-Fi, the input feature map is divided into non-overlapping Windows, and the high-frequency information is encoded by performing self-attention in the local window; In Lo-Fi, firstly, the average pooling is applied to each window to obtain the low-frequency signal in the input image, then the average pooling feature map is mapped to obtain the key K and value V, the query Q is still from the original feature map, finally, the standard attention is applied to capture the low-frequency information, as shown in Fig. 6-16. Pan et al. [55] proposed HiLo Attention, with high frequency capturing local details and low frequency focusing on global structure can help achieve higher efficiency on high-resolution images.

Sixteenthly, Multi-Scale Dilated Attention(MSDA), In MSDA, the channels of the feature maps are split into different headers, and different dilation rates are used in different headers to perform self-attention for patches around the query patch, as shown in Fig. 6-17. Jiao et al. [56] proposed MSDA to capture contextual semantic dependencies of different scales at the same time, so as to realize the ability of multi-scale representation learning.

Seventeenthly, QuadTree Attention, In QuadTree Attention, the feature maps are recursively subdivided into 4 regions, top-k regions with the highest attention scores are selected for each query, so that in the next level, the attention evaluation is only performed in the relevant sub-areas corresponding to these top-k regions, as shown in Fig. 6-18. Tang et al. [57] proposed Quadtree attention to compute attention in a coarse to fine manner, according to the results at the coarse level, irrelevant image regions are skipped quickly at the fine level, this design achieves less information loss while maintaining high efficiency.

Eighteenthly, Bi-Level Routing Attention(BLRA), In BLRA, the input feature maps are divided into multiple non-overlapping regions, the degree of semantic correlation between each two regions is found by constructing a region-to-region adjacency matrix, and the fine-grained token-to-token attention is applied to only the top-k related areas for each region, as shown in Fig. 6-19. Zhu et al. [58] proposed BLRA to achieve efficient allocation of computation in a dynamic and query-aware manner by filtering out most of the irrelevant key-value pairs at the coarse region level

3.5. Network structure

Due to its excellent performance, Transformer is widely used in the field of medical imaging and achieves good results in the computer-aided diagnosis of lung cancer, breast cancer, skin diseases, cardiovascular and cerebrovascular diseases. since the limited performance of a single network, more and more studies find that the ensemble of Transformer with other network models is an important development direction. In the field of medical imaging, there are 5 application fields for Transformer, such as Transformer combined with U-Net is used for medical image segmentation, Transformer combined with GAN is used for fusion, Transformer combined with YOLO is used for detection, Transformer combined with ResNet and DenseNet is used for classification and recognition, the work can provide help for Transformer application.

3.5.1. Transformer combined with U-Net

Medical image segmentation is that the desired lesion regions are segmented from the medical image, such as organs and tissues, the segmented lesions are visualized by digital image processing technology, and the operation is guided by graphic processing technology [59]. Precision lesion segmentation plays a good guiding role in medical image classification. Before 2020, most of the mainstream medical image segmentation methods are improved based on the U-Net [60] model. Although realizing a good segmentation effect, U-Net lacks long-distance relationship modeling due to the locality of convolutional operations, and it is difficult to learn global semantic information. Transformer can capture global information through the attention mechanism to establish long-distance dependencies and extract more feature information. In recent years, many researchers combine Transformer and U-Net for medical image segmentation, helping to improve segmentation precision. In the field of medical image segmentation, The combination of Transformer and U-Net mainly includes TransUNet, TransClaw U-Net, LeViT-UNet, Teeth U-Net, and UTNet.

Firstly, Chen et al. [61] proposed TransUNet. In this network, the hybrid structure of CNN and Transformer is used as an encoder, the feature maps in encoder are upsampled by decoder, and high-resolution CNN feature maps are combined to realize precision localization.

Secondly, Chang et al. [62] proposed TransClaw U-Net. In this network, the encoder adopts a hybrid structure that feature maps are extracted by convolution operation and enhanced by Transformer later, and the decoder part is a two-way design: one way is directly upsampled, and the other way is skipping connections at the same time of upsampling, which effectively realizes the segmentation of medical images.

Thirdly, Xu et al. [63] proposed LeViT-UNet. In this network, LeViT [64] is used as an encoder, Transformer blocks in LeViT and multi-scale feature maps of convolutional blocks are transmitted to the decoder through skip connections, and spatial feature is reused effectively.

Fourthly, Due to blurred boundaries between teeth that make it difficult to segment teeth from panoramic dental X-ray images, Hou et al. [65] proposed Teeth U-Net, in this network, a dilated hybrid self-attentive block is designed for captured dental feature information in a larger field of perception.

Fifthly, Gao et al. [66] propose UTNet, in which self-attention is integrated into CNN to enhance medical image segmentation, and self-attention modules are used in both encoders and decoders to capture long-range dependencies.

3.5.2. Transformer combined with GAN

Due to the diversity and complexity of diseases, it is difficult to diagnose disease types and locate lesions through a single modal medical image. There are richer features and more comprehensive information in the fusion image obtained through multi-modal medical image fusion, which can assist medical images to better serve clinical applications. GAN [67] is the most common deep learning technology in the cross-modal reconstruction of medical images, Through an adversarial learning mechanism, significant information in medical images can be modeled. Transformer not only can process long sequence information, but also can learn the relationship among different modes, and can effectively extract features in multi-model medical images. Therefore, Transformer and GAN are combined to improve the effectiveness and precision of medical image fusion. In the field of medical image synthesis, the combination of Transformer and GAN mainly includes RTCGAN, D-ESRGAN, MedViTGAN, CT-GAN, and TCGAN.

Firstly, due to the local mismatch between MR and CT images in pelvic soft tissue, Zhao et al. [68] proposed a new GAN model, RTCGAN, which utilizes CNN and Transformer to extract multi-level features from MR and CT images, CNN can perceive local texture features and Transformer can perceive global relevance.

Secondly, Wang et al. [69] proposed D-ESRGAN, a generative adversarial network of super-resolution dual encoders, in this network, the loss of texture information in iris images is compensated while the newly generated texture features are kept more natural. D-ESRGAN not only integrates a residual CNN encoder to extract local features, but also uses a ViT encoder to capture global association information.

Thirdly, Li et al. [70] proposed MedViTGAN based on Transformer architecture, which synthetic histopathological images are generated for image enhancement in an end-to-end manner.

Fourthly, Pan et al. [71] proposed CT-GAN, which can better predict Alzheimer's disease by integrating functional information contained in resting-state functional Magnetic Resonance Imaging (rs-fMRI) and structural information contained in Diffusion Tensor Imaging (DTI).

Fifthly, Li et al. [72] proposed TCGAN, which uses a dual generator architecture to fuse PET and CT, the dual generator is combined CNN generator and Transformer generator by series connect.

3.5.3. Transformer combined with YOLO

The computer is used to realize medical image detection, which can help experts to control the disease more accurately. Different from natural images, the edge of most lesion detection targets in medical detection images is often fuzzy and irregular, and the number of pixels is small. It is very difficult to accurately locate lesions by conventional methods, which often leads to problems of missing detection and misdetection. In recent years, object detection methods based on deep neural networks are widely used in medical images. YOLO [73] is a single-stage deep learning detection method with characteristics of real-time and accuracy. Transformer can effectively extract lesion position information by modeling global features in images. Therefore, the Transformer mechanism added to the YOLO backbone network can better extract complex features of lesions and improve detection accuracy. In the field of medical image detection, the combination of Transformer and YOLO mainly includes: CL-YOLOv5, YOLO-LOGO, RDFNet, Improved YOLO, and CCGL-YOLOV5.

Firstly, Zhou et al. [74] proposed CL-YOLOv5. In this network, a cosine reweighting computing Transformer is designed to efficiently learn global feature relations and interactively enhance the ability of the network to extract lesions.

Secondly, Su et al. [75] Proposed YOLO-LOGO for the detection of breast tumors in digital mammograms. In this network, the target detection model YOLOV5L6 is used to locate and cut breast tumors, and the whole image and the cropped image are trained on the global and local Transformer branches respectively.

Thirdly, because of the lack of current research on caries detection, Jiang et al. [76] proposed the RDFNet. In this network, an improved Transformer encoder module is added based on the original SPP structure to improve the network's ability to extract caries features and realize rapid caries detection.

Fourthly, Qi et al. [77] proposed an improved YOLO network, in this network, CBAM (Convolutional Block attention Module) and Multi-Head Self-Attention mechanisms are combined with the yolov3 network for the detection of pulmonary nodules in chest CT images.

Fifthly, Zhou et al. [78] Proposed CCGL-YOLOV5, which is used for Lung Tumor Detection, in this network, a Cross-Modal Fusion Transformer Module is designed for multi-modal feature fusion.

3.5.4. Transformer combined with ResNet

Deep learning-based medical image classification plays a key role in computer-aided diagnosis, such as speeding up film reading, shortening patient wait times, and reducing the burden on imaging physicians. Before the vision Transformer, CNN based on deep learning is used to classify medical images. The classic CNN models include ResNet and DenseNet. Among them, ResNet [79] can effectively alleviate the gradient disappearance and network degradation caused by the increase of network depth through skip connection. The introduction of self-attention mechanisms in Transformer into ResNet allows the model to better model long-distance dependencies, helping to capture global context information in images. In the field of medical image classification and recognition, the combination of Transformer and ResNet mainly includes: Trans-ResNet, ASI-DBNet, DCET-Net, RMT-Net, and REC-ResNet.

Firstly, Li et al. [80] proposed Trans-ResNet network for Alzheimer's disease classification. In this network, ResNet-18 is used to extract local semantic information from input images, then the generated feature maps are divided and fed into Transformer network for classification.

Secondly, Zhou et al. [81] proposed ASI-DBNet for rapid and accurate classification of brain cancer. In this network, an adaptive sparse interaction block is designed to realize the interaction between ResNet branches and ViT branches, which makes the feature maps transmitted during the interaction more beneficial.

Thirdly, Zou et al. [82] proposed DCET-Net, which is a dual-flow network based on two backbone networks of CNN and Transformer for breast cancer histopathological image classification. The network uses CNN to capture the local depth features of histopathological images, and Transformer to enhance the global information of depth features, presents a more discriminating feature.

Fourthly, Ren et al. [83] proposed a new deep learning network

RMT-Net based on the combination of ResNet-50 and Transformer. In this network, Transformer is used to capture long-distance feature information, and local features are obtained by deep convolution.

Fifthly, Zhou et al. [84] proposed a COVID-19 auxiliary diagnosis model REC-ResNet, which uses ResNet50 as the main trunk network, and three feature enhancement strategies are introduced to improve the feature extraction capability of the model.

3.5.5. Transformer combined with DenseNet

Compared with ResNet, DenseNet [85] proposed a dense connection mechanism: all layers are interconnected, and each layer is concatenated with all previous layers in the channel dimension as the input of the next layer, which can not only achieve feature reuse, but also improve efficiency. Transformer uses self-attention mechanism to calculate all locations so that each location in the network can interact with each other to further improve feature reuse. The combination of DenseNet and Transformer can capture more global context information and improve feature reuse, thus, the performance of the model is improved. In the field of medical image classification and recognition, the combination of Transformer and DenseNet mainly includes: DRLTransformer, DDSF-Net, and DPE-BoTNet.

Firstly, Zhou et al. [86] proposed DRLTransformer. In this network, heavy reference dense blocks and hierarchical Transformerfor are designed for COVID-19 recognition in CT images.

Secondly, Zhou et al. [87] proposed DDSF-Net for pneumonia diagnosis. In this network, Transformer is used to learn the global context semantic information, the convolutional layer is used to extract local features, and the dense connection method is used to realize the deep and shallow layer feature fusion of the two information flows.

Thirdly, Nakai et al. [88] proposed DPE-BoTNet for skin disease classification based on DenseNet201. By combining Transformer and DenseNet, both local interaction and global dependence can be modeled to improve skin disease classification performance.

4. Conclusion

ViT makes breakthrough progress in the field of deep learning, in ViT, there are four main steps, which are “four secrets”, such as patch division, token selection, position encoding addition, attention calculation, the existing research on transformer in computer vision mainly focuses on the above four steps. Therefore, “how to divide patch?”, “how to select token?”, “how to add position encoding?”, and “how to calculate attention?” are crucial to improve ViT performance. But so far, most of the review literatures are summarized from the perspective of application, and there is no corresponding literature to comprehensively summarize these four steps from the technology perspective, which restricts the further development of ViT in some degree. Therefore, aiming at the above problems, this paper makes a comprehensive summary of the 4 mechanisms and 5 applications of ViT, the main contributions are as follows:

Firstly, aiming to “how to divide patch?”, 5 key technologies of patch division mechanism are summarized: (1) from single-size division to multi-size division; (2) from fixed number division to adaptive number division; (3) from non-overlapping division to overlapping division; (4) from semantic segmentation division to semantic aggregation division; (5) from original image division to feature map division.

Secondly, aiming to “how to select token?”, 3 key technologies of token selection mechanism are summarized: (1) token selection based on score; (2) token selection based on merge; (3) token selection based on convolution and pooling.

Thirdly, aiming to “how to add position encoding?”, 5 key technologies of position encoding mechanism are summarized: (1) absolute position encoding; (2) relative position encoding; (3) conditional position encoding; (4) locally-enhanced position encoding; (5) zero-padding position encoding.

Fourthly, aiming to “how to calculate attention?”, 18 attention mechanisms are summarized based on the time axis, such as Focal Self-attention, W-MSA and SW-MSA, Cross-Attention, Spatially Separable Self-Attention, Long Short Distance Attention, Pale-Shaped Self-attention, Cross-Shaped Window Self-Attention, Regional-to-Local Attention, Dual Attention, Depthwise Separable Self-attention, Deformable Attention, Multi-axis Attention, Bilateral Local Attention, Neighborhood Attention, HiLo Attention, Multi-Scale Dilated Attention, QuadTree Attention, Bi-Level Routing Attention.

Fifthly, the extensive applications of ViT in medical image processing are discussed by combining with U-Net, GAN, YOLO, ResNet, and DenseNet.

5. Future work

Although ViT makes breakthrough progress in the field of computer vision and plays a substantial role, it is important to design a reasonable network model with a good generalization effect for the study of ViT, patch division mechanism, token selection mechanism, position encoding mechanism and attention mechanism are worth further discussion and improvement, and building a unified framework for multi-tasks, reducing high-dimensional data computation, realizing small sample learning, and having good interpretability of model structure are all the future development directions of ViT.

Firstly, the research of ViT patch division mechanism. The future directions of patch division mechanism are as follows: (1)non-square patch division, such as rectangle, circle, triangle et al, which is used to extract target features with complex shapes; (2)multiple patch division, which is used to process multi-scale features in parallel; (3)dynamical patch division, patch size is determined dynamically according to the size of targets in the image, which is used to better adapt to different targets in the image.

Secondly, the research of ViT token selection mechanism. The future directions of token selection mechanism are as follows: (1) adaptive scoring strategy, the model adaptively scores for each token based on the characteristics of each token and the interrelationship among all tokens; (2) adaptive merge strategy, how to merge redundant tokens is very necessary to ViT token selection; (3) selection strategy combined with other evolutionary algorithms, such as genetic algorithm, ant colony algorithm, and other methods, to research the optimal token search strategy.

Thirdly, the research of ViT position encoding mechanism. This is a mechanism that is based on position encoding generators, such as sine function, cosine function, Gaussian function, polynomial function, etc.

Fourthly, the research of ViT attention mechanism. The future directions of attention mechanism are as follows: (1) the selection of attention mechanism, how to select appropriate attention mechanism and improve the model performance are important for ViT; (2) the problem of selecting the area for attention computation; (3) the multi-modal attention mechanism, focusing on visual, speech, text, and other modal information at the same time, and improving the generalization ability of the model.

Fifthly, the research of building a unified framework for multi-tasking. Traditional multimodal models take different processing methods for different data types, so it is inevitable that patterns cannot be aligned when feature splicing. Not only the model structure is complex, but also the processing results of different data types are not good. At present, Transformer achieves great success in text, image, video, language, and other aspects. Transformer self-attention mechanism has strong feature extraction and mode alignment capabilities. Therefore, how to build a unified framework to capture the internal relationship between multi-modal data will be the future development trend.

Sixthly, the research of reducing high-dimensional data calculation. Due to the large number of parameters and high computational complexity, the existing vision Transformer model takes a long time to train and reason, which requires a lot of computing resources and time as well as strong hardware support. Even though the performance of

hardware devices is constantly improving, Transformer still cannot meet the requirements of computing efficiency. Therefore, how to improve the computing efficiency of Transformer model is a hot and difficult problem in future work.

Seventhly, the research of realizing small sample learning. Compared with CNN, Transformer model has more parameters, so its training usually relies on a larger number of training samples. The current training method is to pre-train the vision Transformer model on a large dataset, and then fine-tune the model with a small amount of data for the task type. However, for some visual tasks with sparse training samples, it is often difficult to obtain massive training data, and the quantity and quality of training data limit the training and performance improvement of the model. Therefore, how to realize small sample learning with prior knowledge is still a challenge.

Eighthly, the research of ViT interpretability. Compared with CNN and RNN, Transformer has a larger capacity and its architecture can support large-scale data training. However, the theoretical reasons are not clear. The attention of each layer in Transformer is mixed in a complex way in subsequent layers, thus, it is difficult to visualize the relative weight of input tokens to the final prediction. In order to better design and improve the Transformer model structure, it is necessary to deeply study and understand its operation mechanism and internal information interaction. Therefore, it is great of significance to study Transformer model interpretability.

CRediT authorship contribution statement

Tao Zhou: Writing – original draft. **Yuxia Niu:** Writing – original draft, Investigation. **Huiling Lu:** Supervision. **Caiyue Peng:** Formal analysis. **Yujie Guo:** Investigation. **Huiyu Zhou:** Writing – review & editing.

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.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant No. 62062003, Natural Science Foundation of Ningxia under Grant No.2023AAC03293.

References

[1] A. Vaswani, et al., Attention is all you need, Neural Inf. Process. Syst. (2017) 5998–6008.
[2] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: pretraining of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805.
[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI blog, 2018.
[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, 2019.
[5] T. Brown, et al., Language models are few-shot learners, Neural Inf. Process. Syst. (2020) 1877–1901.
[6] Y. Liu et al., "RoBERTa: a robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692.
[7] C. Raffel, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. (2020) 5485–5551.
[8] P. Battaglia et al., "Relational inductive biases, deep learning, and graph networks," 2018, arXiv:1806.01261.
[9] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.
[10] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2009, pp. 248-255.
[11] T.Y. Lin, et al., Microsoft coco: Common objects in context, in: Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, 2014, pp. 740–755.
[12] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 633–641.
[13] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, I. Sutskever, Generative pretraining from pixels, in: Proceedings of the International Conference on Machine Learning, 2020, pp. 1691–1703.
[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proceedings of the European Conference on Computer Vision, 2020, pp. 213-229.
[15] H. Wang, Y. Zhu, H. Adam, A. Yuille, L.C. Chen, Max-deeplab: end-to-end panoptic segmentation with mask transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5463–5474.
[16] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, H. Lu, Transformer tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8126–8135.
[17] Y. Jiang, S. Chang, Z. Wang, Transgan: Two pure transformers can make one strong gan, and that can scale up, Neural Inf. Process. Syst. (2021) 14745–14758.
[18] H. Chen, et al., Pre-trained image processing transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12299–12310.
[19] K. Han, et al., A survey on vision transformer, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, pp. 87–110.
[20] Y. Liu, et al., A survey of visual transformers, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–21.
[21] S. Khan, M. Naseer, M. Hayat, S.W. Zamir, F.S. Khan, M. Shah, Transformers in vision: a survey, ACM Comput. Surv. (2022) 1–41.
[22] C.F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.
[23] Y. Lee, J. Kim, J. Willette, S.J. Hwang, Mpvit: Multi-path vision transformer for dense prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7287–7296.
[24] Y. Wang, R. Huang, S. Song, Z. Huang, G. Huang, Not all images are worth 16×16 words: dynamic transformers for efficient image recognition, Neural Inf. Process. Syst. (2021) 11960–11973.
[25] W. Wang, et al., Pvt v2: improved baselines with pyramid vision transformer, Comput. Vis. Media (2022) 415–424 (Beijing).
[26] Z. Chen, Y. Zhu, C. Zhao, G. Hu, W. Zeng, J. Wang, M. Tang, Dpt: deformable patch-based transformer for visual recognition, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2899–2907.
[27] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 579–588.
[28] C. Liu, K. Hirota, Y. Dai, Patch attention convolutional vision transformer for facial expression recognition with occlusion, Inf. Sci. (2023) 781–794.
[29] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, C.J. Hsieh, Dynamicvit: efficient vision transformers with dynamic token sparsification, Neural Inf. Process. Syst. (2021) 13937–13949.
[30] 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," 2022, arXiv: 2202.07800.
[31] Y. Xu, et al., Evo-vit: slow-fast token evolution for dynamic vision transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2964–2972.
[32] H. Yin, A. Vahdat, J.M. Alvarez, A. Mallya, J. Kautz, P. Molchanov, A-ViT: adaptive tokens for efficient vision transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10809–10818.
[33] S. Kim, S. Shen, D. Thorsley, A. Gholami, W. Kwon, J. Hassoun, K. Keutzer, Learned token pruning for transformers, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 784–794.
[34] L. Yuan, et al., Tokens-to-token ViT: training vision transformers from scratch on ImageNet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.
[35] D. Bolya, C.Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: your ViT but faster," 2022, arXiv:2210.09461.
[36] W. Zeng, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, X. Wang, Not all tokens are equal: human-centric visual analysis via token clustering transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11101–11111.
[37] Z. Pan, B. Zhuang, J. Liu, H. He, J. Cai, Scalable vision transformers with hierarchical pooling, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 377–386.
[38] B. Chen et al., "PSViT: better vision transformer via token pooling and attention sharing," 2021, arXiv:2108.03428.
[39] Z. Liu, et al., Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[40] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," 2021, arXiv:2102.10882.

[41] X. Dong, et al., CSWin transformer: a general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.
[42] M.A. Islam, S. Jia, and N.D. Bruce, "How much position information do convolutional neural networks encode," 2020, arXiv:2001.08248.
[43] V. Mnih, N. Heess, A. Graves, Recurrent models of visual attention, Neural Inf. Process. Syst. (2014) 2204–2212.
[44] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," 2021, arXiv:2107.00641.
[45] X. Chu, et al., Twins: revisiting the design of spatial attention in vision transformers, Neural Inf. Process. Syst. (2021) 9355–9366.
[46] W. Wang et al., "CrossFormer: a versatile vision transformer based on cross-scale attention," 2021, arXiv:2108.00154.
[47] S. Wu, T. Wu, H. Tan, G. Guo, Pale transformer: A general vision transformer backbone with pale-shaped attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 2731–2739.
[48] C.F. Chen, R. Panda, and Q. Fan, "RegionViT: Regional-to-local attention for vision transformers," 2021, arXiv:2106.02689.
[49] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, L. Yuan, Davit: dual attention vision transformers, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 74–92.
[50] W. Li, X. Wang, X. Xia, J. Wu, X. Xiao, M. Zheng, and S. Wen, "SepViT: separable vision transformer," 2022, arXiv:2203.15380.
[51] Z. Xia, X. Pan, S. Song, L.E. Li, G. Huang, Vision transformer with deformable attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4794–4803.
[52] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, Y. Li, MaxViT: multi-axis vision transformer, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 459–479.
[53] T. Yu, G. Zhao, P. Li, and Y. Yu, "BOAT: bilateral local attention vision transformer," 2022, arXiv:2201.13027.
[54] A. Hassani, S. Walton, J. Li, S. Li, H. Shi, Neighborhood attention transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6185–6194.
[55] Z. Pan, J. Cai, B. Zhuang, Fast vision transformers with hilo attention, Neural Inf. Process. Syst. (2022) 14541–14554.
[56] J. Jiao, Y.M. Tang, K.Y. Lin, Y. Gao, J. Ma, Y. Wang, W.S. Zheng, DilateFormer: Multi-scale dilated transformer for visual recognition, IEEE Trans. Multimedia (2023) 1–14.
[57] S. Tang, J. Zhang, S. Zhu, and P. Tan, "QuadTree attention for vision transformers," 2022, arXiv:2201.02767.
[58] L. Zhu, X. Wang, Z. Ke, W. Zhang, R.W. Lau, BiFormer: vision transformer with bi-level routing attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10323–10333.
[59] C.J. Yang, X. Yang, Abdominal CT image segmentation based on graph cuts and fast level set, CT Theory Appl. (2011) 291–300.
[60] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, 2015, pp. 234–241.
[61] J. Chen et al., "TransUNet: transformers make strong encoders for medical image segmentation," 2021, arXiv:2102.04306.
[62] Y. Chang, H. Menghan, Z. Guangtao, and Z. Xiao-Ping, "TransClaw U-Net: claw U-Net with transformers for medical image segmentation," 2021, arXiv:2107.05188.
[63] G. Xu, X. Wu, X. Zhang, and X. He, "Levit-unet: make faster encoders with transformer for medical image segmentation," 2021, arXiv:2107.08623.
[64] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, M. Douze, LeViT: a vision transformer in ConvNet's clothing for faster inference, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12259–12269.
[65] S. Hou, T. Zhou, Y. Liu, P. Dang, H. Lu, H. Shi, Teeth U-Net: a segmentation model of dental panoramic X-ray images for context semantics and contrast enhancement, Comput. Biol. Med. (2023) 106296.
[66] Y. Gao, M. Zhou, D.N. Metaxas, UTNet: a hybrid transformer architecture for medical image segmentation, in: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, 2021, pp. 61–71.
[67] T. Zhou, Q. Li, H. Lu, Q. Cheng, X. Zhang, GAN review: models and medical image fusion applications, Inf. Fusion (2023) 134–148.
[68] B. Zhao, et al., CT synthesis from MR in the pelvic area using residual transformer conditional GAN, Comput. Med. Imaging Graph. (2023) 102150.
[69] C. Wang, T. Lu, G. Wu, Y. Wang, Z. Sun, D-ESRGAN: a dual-encoder GAN with residual CNN and vision transformer for iris image super-resolution, in: Proceedings of the IEEE International Joint Conference on Biometrics, 2022, pp. 1–8.
[70] M. Li, C. Li, P. Hobson, T. Jennings, B.C. Lovell, MedViTGAN: end-to-end conditional GAN for histopathology image augmentation with vision transformers, in: Proceedings of the 26th International Conference on Pattern Recognition, 2022, pp. 4406–4413.
[71] J. Pan, and S. Wang, "Cross-modal transformer GAN: a brain structure-function deep fusing framework for Alzheimer's disease," 2022, arXiv:2206.13393.
[72] J. Li, Z. Qu, Y. Yang, F. Zhang, M. Li, S. Hu, TCGAN: a transformer-enhanced GAN for PET synthetic CT, Biomed. Opt. Express (2022) 6003–6018.
[73] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[74] T Zhou, X Ye, F Liu, H Lu, PET/CT lung tumor detection based on cross-modal lightweight YOLOv5 model, J. Electron. Inf. (2023) 1–9.
[75] Y. Su, Q. Liu, W. Xie, P. Hu, YOLO-LOGO: a transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms, Comput. Methods Progr. Biomed. (2022) 106903.
[76] H. Jiang, P. Zhang, C. Che, B. Jin, Rdfnet: A fast caries detection method incorporating transformer mechanism, Comput. Math. Methods Med. (2021).
[77] H. Qi, J. Jia, R. Zhang, Detection of CT pulmonary nodule based on improved YOLO using attention mechanism, in: Proceedings of the 6th International Conference on Computer Science and Artificial Intelligence, 2022, pp. 13–20.
[78] T. Zhou, F. Liu, X. Ye, H. Wang, H. Lu, CCGL-YOLOV5: a cross-modal cross-scale global-local attention YOLOV5 lung tumor detection model, Comput Biol Med (2023) 107387.
[79] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[80] C. Li, Y. Cui, N. Luo, Y. Liu, P. Bourgeat, J. Fripp, T. Jiang, Trans-ResNet: Integrating transformers and CNNs for Alzheimer's disease classification, in: Proceedings of the IEEE 19th International Symposium on Biomedical Imaging, 2022, pp. 1–5.
[81] X. Zhou, C. Tang, P. Huang, S. Tian, F. Mercaldo, A. Santone, ASI-DBNet: an adaptive sparse interactive ResNet-vision transformer dual-branch network for the grading of brain cancer histopathological images, Interdiscip. Sci. Comput. Life Sci. (2023) 15–31.
[82] Y. Zou, S. Chen, Q. Sun, B. Liu, J. Zhang, DCET-Net: dual-stream convolution expanded transformer for breast cancer histopathological image classification, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, 2021, pp. 1235–1240.
[83] K. Ren, G. Hong, X. Chen, Z. Wang, A COVID-19 medical image classification algorithm based on transformer, Sci. Rep. (2023) 5359.
[84] T Zhou, Y Liu, S Hou, X Ye, H Lu, REC-ResNet: a feature enhancement model for COVID-19 diagnosis, Opt. Precis. Eng. (2023) 2093–2110.
[85] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[86] T Zhou, X Ye, F Liu, H Lu, J Zhou, Y Du, A dense re-referencing lightweight transformer model for the identification of new coronary pneumonia in CT images, J. Electron. Inf. (2022) 1–9.
[87] T. Zhou, X Ye, H Lu, Y Liu, X Chang, A dense dual-flow focusing network-based model for pneumonia-assisted diagnosis, Opt. Precis. Eng. (2023) 1074–1084.
[88] K. Nakai, X.H. Han, DPE-BoTNeT: dual position encoding bottleneck transformer network for skin lesion classification, in: Proceedings of the IEEE 19th International Symposium on Biomedical Imaging, 2022, pp. 1–5.