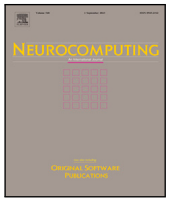
Neurocomputing 595 (2024) 127828

Contents lists available at ScienceDirect

Neurocomputing

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



MAFormer: A transformer network with multi-scale attention fusion for visual recognition

Huixin Sun [a,1], Yunhao Wang [b,1], Xiaodi Wang [b], Bin Zhang [b], Ying Xin [b], Baochang Zhang [c,d,e], Xianbin Cao [a], Errui Ding [b], Shumin Han [b,*,1]

[a] School of Electronic Information Engineering, Beihang University, Beijing, 100191, China
[b] Department of Computer Vision Technology (VIS), Baidu Inc, Beijing, 100024, China
[c] Institute of Artificial Intelligence, Beihang University, Beijing, 100191, China
[d] Hangzhou Research Institute, Beihang University, Hangzhou, 310052, China
[e] Zhongguancun Laboratory, Beijing, 100190, China

ARTICLE INFO

Communicated by X. Gu

Keywords:
Vision transformer
Multi-scale attention fusion

ABSTRACT

Vision Transformer and its variants have demonstrated great potential in various computer vision tasks. However conventional vision transformers often focus on global dependency at a coarse level, which results in a learning challenge on global relationships and fine-grained representation at a token level. In this paper, we introduce Multi-scale Attention Fusion into transformer (**MAFormer**), which explores local aggregation and global feature extraction in a dual-stream framework for visual recognition. We develop a simple but effective module to explore the full potential of transformers for visual representation by learning fine-grained and coarse-grained features at a token level and dynamically fusing them. Our Multi-scale Attention Fusion (MAF) block consists of: i) a local window attention branch that learns short-range interactions within windows, aggregating fine-grained local features; ii) global feature extraction through a novel Global Learning with Down-sampling (GLD) operation to efficiently capture long-range context information within the whole image; iii) a fusion module that self-explores the integration of both features via attention. Our MAFormer achieves state-of-the-art results on several common vision tasks. In particular, MAFormer-L achieves 85.9% Top-1 accuracy on ImageNet, surpassing CSWin-B and LV-ViT-L by 1.7% and 0.6% respectively. On MSCOCO, MAFormer outperforms the prior art CSWin by 1.7% mAPs on object detection and 1.4% on instance segmentation with similar-sized parameters. With the performance, MAFormer demonstrates the ability to generalize across various visual benchmarks and prospects as a general backbone for different self-supervised pre-training tasks in the future.

1. Introduction

Convolutional neural networks (CNNs) have dominated computer vision, excelling in image recognition, object detection, and semantic segmentation tasks. Initiated by the success of AlexNet [1], CNN architectures have continually evolved, enhancing their performance through scaling, enhanced connections, and varied convolutional techniques. These advancements have established CNNs as fundamental frameworks for diverse vision tasks, significantly advancing the field.

On the other hand, network architecture evolution in Natural Language Processing (NLP) has focused the Transformer model [2]. The Transformer represents input words as a sequence of tokens and achieves notable performance for using attention to model long-range dependencies in the data. Motivated by its achievements in NLP, Dosovitskiy et al. [3] introduced the Vision Transformer (ViT), a groundbreaking approach for detecting complex patterns and relationships within images, which has recently demonstrated promising results on certain tasks, including image recognition [1], semantic segmentation [4], image inpainting [5–8], and super resolution [9].

Despite the progress, the global self-attention mechanism in line with ViT [3] has a quadratic computation complexity to the input image size, which is insufferable for high-resolution scenes. To reduce the complexity, several variants have been introduced to replace global self-concern with local self-concern. Swin Transformer [10] with a hierarchical architecture partitions input features into non-overlapping windows and shifts the window positions by layer. After that various window partition mechanisms are designed for better

* Corresponding author.
E-mail addresses: sunhuixin@buaa.edu.cn (H. Sun), wangyunhao02@baidu.com (Y. Wang), hanshumin@baidu.com (S. Han).
[1] Equal contributions.

https://doi.org/10.1016/j.neucom.2024.127828
Received 12 December 2023; Received in revised form 13 March 2024; Accepted 5 May 2024
Available online 10 May 2024

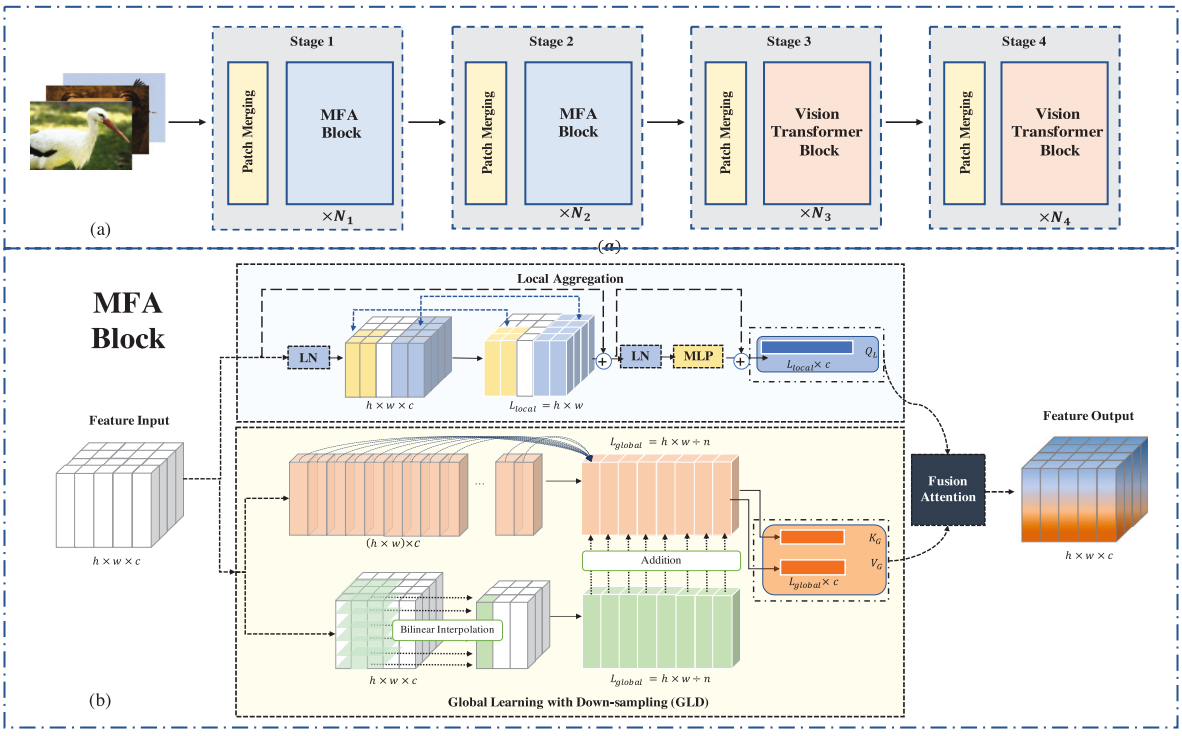


Fig. 1. Architecture of MAFormer. We utilize the MAF block in the first two stages, which incorporates a Local Aggregation branch and a Global Learning with Down Sampling (GLD) branch. Both streams are fed into a fusion module to improve the capability of feature representation.

local feature capturing. CSWin Transformer [11] splits features into horizontal and vertical stripes in parallel, aiming to enlarge the window receptive field. However, the method aggregates information within local windows but neglects dependencies across them. Shuffle Transformer [12] revisits the ShuffleNet [13] and embeds the spatial shuffle in local windows to intensify their connections. While these local window-based attention methods have achieved excellent performance, even better than the convolutional neural network (CNN) counterparts (e.g., ResNets [14,15]), they can suffer from insufficient learning on the global relationship across windows, limiting their ability to generalize across various visual tasks.

Another line of research efforts focuses on combining CNNs with transformers, which focus on the learning synergies between local and global feature representations. CvT [16] transforms the linear projection in the self-attention block into convolution projection. CoatNet [17] merges depth-wise convolution with self-attention via simple relative attention and stacks convolution and attention layers in a principled way. DS-Net [18] proposes a dual-stream framework that fuses convolution and self-attention via cross-attention, where each form of scale learns to align with the other. However, as shown in DS-Net [18], convolution and attention hold intrinsically conflicting properties that might cause ambiguity in training. For instance, the long-range information captured by global self-attention could perturb the neighboring details of convolution in high-resolution feature maps, compromising both global and local feature representations.

In this paper, we develop a Multi-scale Attention Fusion transformer (**MAFormer**), which explores local aggregation and global feature extraction in a dual-stream transformer framework. To avoid the risk of incompatibility between convolution and self-attention in previous designs, we apply local window attention to extract fine-grained feature representation within local windows while extending their reduced computational complexity advantage in our framework. We also design a Global Learning with Down-sampling (GLD) module to extract global features, which captures coarse-grained features based on the full-sized input. We further encode token-level location information of the input into global representations via positional embeddings. Moreover, we describe two dual-stream architectures based on different fusion strategies, particularly the Multi-scale Attention Fusion (MAF) scheme that can fully explore the potential of both features. Its effectiveness can be explained by the fact that MAF block can enhance the interaction between each local–global token pair, where local features and global features are co-trained in a unified framework, formulating more comprehensive and informative feature representations. For this paper, the main contributions are as follows:

1. A MAFormer network is introduced to extract and fuse fine-grained and coarse-grained features at a token level, which can self-explore the integration of both features via attention to improve the representation capacity for the input image.
2. A local window attention branch is first introduced to learn the short-range interactions within local windows. We further introduce a Global Learning with Down-sampling (GLD) module on the dual branch, which efficiently captures the long-range context information within the whole image.
3. We develop two dual-stream architectures based on different fusion strategies, particularly the Multi-scale Attention Fusion (MAF) scheme that can fully explore the potential of both features.
4. Without bells and whistles, the proposed MAFormer outperforms prior vision Transformers by large margins in terms of recognition performance. We also achieve state-of-the-art results over the previous best CSWin for object detection and instance segmentation with similar parameters.

2. Related work

CNNs. CNNs have become the standard architecture in the field of computer vision [4,19,20]. The concept of CNNs has existed for several decades, with significant milestones such as the development described by LeCun et al. [21]. The advent of AlexNet [22] catalyzed the widespread adoption and success of CNNs in mainstream applications. Following the breakthrough, convolutional network models

were developed, including VGG [23], GoogleNet [24], ResNet [14], DenseNet [25], and HRNet [26]. Despite CNNs' dominance, recent developments highlight the promising capabilities of Transformer-like architectures for bridging the gap between visual and language models [27]. Our work achieves notable performance on several basic visual recognition tasks, and we hope it will contribute to a modeling shift.

Vision transformers. Self-attention-based architectures, in particular Transformers [2], have become the dominant model for Natural Language Processing (NLP). Motivated by its success in NLP, ViT [3] innovatively applies a pure-transformer architecture to images by splitting an image into patches and equating them with tokens (words). The method shows the competitive effect on image recognition [1], semantic segmentation, image inpainting [5–8], and super resolution [9]. Many efforts have been devoted to applying ViT for various vision tasks, including object detection [28–32], semantic segmentation [33–35], pose estimation [36–38], re-identification [39,40], face recognition [41], and low-level image processing [42,43]. These results verify the outstanding ability of the transformer as a general visual backbone. However, the self-attention mechanism is inefficient in encoding low-level features, hindering their high potential for efficient representation learning.

Local window attention-based transformers. Vision transformers demonstrate a high capability in modeling long-range dependencies, which is especially helpful for handling high-resolution inputs in downstream tasks. However, such methods adopt the original full self-attention and their computational complexity is quadratic to the image size. To reduce the cost, some recent vision Transformers [10, 44] adopt the local window self-attention mechanism [45] and its shifted/haloed version that adds the interaction across different windows. To enlarge the receptive field, axial self-attention [46] and criss-cross attention [47] propose calculating attention within stripes along horizontal or/and vertical axis instead of fixing local windows as squares. Inspired by axial self-attention and criss-cross attention, the method [11] presents the Cross-Shaped Window self-attention. CSWin performs the self-attention calculation in the horizontal and vertical stripes in parallel, with each stripe obtained by splitting the input feature into stripes of equal width.

Convolution in transformers. According to recent study [48,49], convolution networks and transformers hold different merits. While the convolution operation guarantees a better generalization and fast convergence, thanks to its inductive bias, attention forms networks with higher model capacity. Therefore, combining convolutional and attention layers can joint these advantages and achieve better generalization and capacity at the same time. Some existing transformers explore the hybrid architecture to incorporate both operations for better visual representation. Comformer [48] proposes the Feature Coupling Unit to fuse convolutional local features with transformer-based global representations in an interactive fashion. CvT [50] designs convolutional token embedding and convolutional transformer block for capturing more precise local spatial context. Followingly, CoatNet [49] merges depthwise convolution into attention layers with simple relative attention. Apart from incorporating explicit convolution, some works [10,11,51, 52] try to incorporate some desirable properties of convolution into the Transformer backbone.

3. Method

3.1. Overall architecture

The Multi-scale Attention Fusion mechanism is proposed to extract fine-grained and coarse-grained features at a token level and fuse them dynamically, which formulates a general vision transformer backbone, dubbed as MAFormer, improving the performance in various visual tasks. Fig. 1(a) shows the overall architecture of MAFormer. It takes an image $\mathcal{X} \in R^{H\times W\times 3}$ as input, where W and H where represent the width and height of the input image, and employs a hierarchical design. By decreasing the resolution of feature maps, the network captures multi-scale features across different stages. We partition an input image into patches and perform patch merging, receiving $\frac{H}{4}\times\frac{W}{4}$ visual tokens with C feature channels. From there, the tokens flow through two stages of MAF Blocks and the two stages of the original Vision Transformer Blocks. Within each stage, MAFormer adopts a patch merging layer by convention which downsamples the spatial size of the feature map by 2×, while the feature channel dimension is increased.

According to recent studies into feature representations [53], visual transformers like the ViT attend locally and globally in its lower layers but primarily focus on global information in higher layers. In light of the pattern, we incorporate multi-scale feature representations in the first two stages of MAFormer, while in the last two stages, the original vision transformer block is utilized, where the resolution of the features is reduced and the computational cost of full attention becomes affordable.

3.2. Multi-scale attention fusion block

In this section, we elaborate the details of our Multi-scale Attention Fusion (MAF) block. As shown in Fig. 1(b), the MAF block includes a Local Aggregation branch and a Global Learning with Down Sampling (GLD) branch, generating token-level fine-grained and coarse-grained features respectively. Both streams are fed into a fusion module to improve the capability of feature representation.

Local aggregation. Previous hybrid networks [49,54] utilize CNNs to extract local features, which are further integrated into a Transformer branch. Yet, such approaches risk the mismatch between convolution and self-attention. In MAF, we avoid incompatibility and explore the usage of local window-based multi-head attention mechanisms as the fine-grained representation. Considering an input $X \in \mathbf{R}^{H\times W\times C}$, the local aggregation X_L^l is defined as follows:

$$\begin{aligned} X_L^l &= \text{Local-Window-Attention}\left(\text{LN}\left(X^{l-1}\right)\right) + X^{l-1}, \\ X_L^l &= \text{MLP}\left(X_L^l\right) + X_L^l, \end{aligned} \tag{1}$$

where Local-Window-Attention(·) applies local window-based attention mechanism on the layer-normalized (LN) input from the previous layer, MLP(·) indicates the multi-layer perceptron that processes the output of the local window attention mechanism, and X^l denotes the output of lth Transformer block.

Global feature extraction. Although local window self-attention methods have achieved excellent performance, they can only capture window-wise information and fail to explore the dependencies across them. Also, existing methods are still challenged in global dependency extraction due to insufficient usage of coarse-grained contextual information. As such, efficient capture of the global dependencies is constitutive of model representation.

To address these issues, we introduce a Global Learning with Downsampling (GLD) module to extract global information from a large-sized input. To this end, we first utilize a single neuron layer that is fully connected to the feature input. Without cutting out any dimensions, it outputs a down-sampled contextual abstraction that is dynamically learned. As illustrated in Fig. 1(c), the input $X \in \mathbf{R}^{H\times W\times C}$ is first flattened to $X_G \in \mathbf{R}^{C\times L}$, where L is equal to $H\times W$. Then $X_G \in \mathbf{R}^{C\times L}$ is globally extracted by a fully connected layer, downsized to scaling ratio N. During experiments, we have tuned several values of N, and 0.5 is optimal, which is set as the default in MAFormer. Further, we encode the token-level location information of the input into global representations via positional embeddings. As illustrated in the Fig. 1(c), the *Pos*

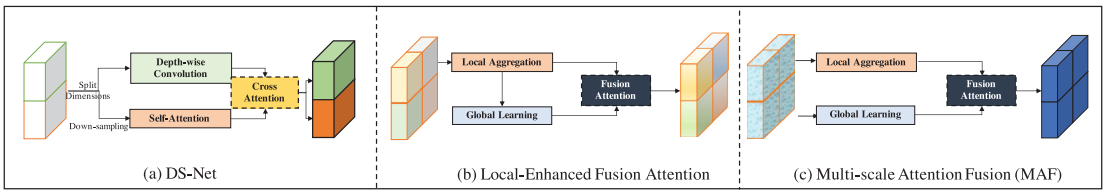


Fig. 2. Different designs in dual-stream multi-scale representations.

Algorithm 1 Multi-scale Attention Fusion (MAFormer)

Input: Image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ with width W and height H
Partition image into patches and perform patch merging
Transform patches into visual tokens with dimension $\frac{H}{4} \times \frac{W}{4} \times C$
for each stage in MAFormer **do**
  Compute local features X_L^l with Eq. (1)
  Compute global features X_G^l with Eq. (2)
  Multi-scale Attention Fusion (MAF):
    Calculate queries Q_L from local features with Eq. (3)
    Calculate keys K_G and values V_G from global features with Eq. (3)
    Calculate MAF features $= \text{softmax}\left(\frac{Q_L K_G^\top}{\sqrt{d_k}}\right) V_G$
end for
Output: Enhanced MAF features for visual tasks

operation utilizes a layer-wise bilinear interpolation as the measure and FC represents as the full connection,

$$X_G^l = Pos\left(X_G^{l-1}\right) + FC(X_G^{l-1}), \tag{2}$$

where X_G^l denotes the global branch output of lth Transformer block.

Multi-scale attention fusion (MAF). We develop two types of dual-stream multi-scale representations, as shown in Fig. 2. First, we extract global dependencies on top of local representations as an enhancement, aiming to provide information flow across local windows. As shown in Fig. 2(b), the GLD module takes the output of local window attention and fuses the global representations back with local. However, this approach can only capture the global correlations between local attributes, not from the input. Therefore, we propose the Multi-scale attention fusion (MAF) measure, extracting the local and global scales of input directly and separately. Both streams of information are fed into a fusion block via attention, as shown in Fig. 2(c). In this way, our MAF block can capture the correlations between each local–global token pair and prompts the local features to adaptively explore their relationship with global representation, enabling themselves to be more representative and informative.

Given extracted local features $X_L \in \mathbf{R}^{C\times L_{local}}$ and global features $X_G \in \mathbf{R}^{C\times L_{global}}$, the Multi-scale Attention Fusion is defined as follows:

$$\begin{aligned} Q_L &= X_L W_Q^{local}, \\ K_G &= X_G W_K^{global}, \\ V_G &= X_G W_V^{global}, \end{aligned} \tag{3}$$

where W_Q^{local}, W_K^{global}, W_V^{global} are learning hyper-parameter matrix and Q_L, K_G, and V_G represent the query, key, and value matrices for local and global features respectively. C is the channel dimension, L_{local} and L_{global} denote the length of local and global feature sequences respectively. Then we calculate the Multi-scale Attention Fusion (MAF) between every pair of X_L and X_G.

$$\text{MAF}\left(Q_L, K_G, V_G\right) = \text{softmax}\left(\frac{Q_L K_G^T}{\sqrt{d}}\right) V_G. \tag{4}$$

The operation softmax$(\cdot)$ computes attention weights, where d represents the scaling factor, typically the dimensionality of the key vectors, to control the softmax gradient. This ensures a balanced attention mechanism by comparing the similarity of local queries and global keys.

Table 1
Detailed settings of MAFormer of different model sizes and their performance on ImageNet-1k validation set. The model size is calculated by summing the product of the number of elements and the size of each element for all parameter in the model. FLOPs are calculated based on the model's MACs (Multiply-Accumulate operations), accumulated at operator-level. In all configurations, the expansion ratio of each MLP is set as 4.

Models	Dim	Blocks	Params (M)	FLOPs (G)	Top1 (%)
MAFormer-S	[64, 128, 320, 512]	[3, 5, 8, 3]	23	4.5	83.7
MAFormer-B	[64, 128, 320, 512]	[3, 8, 20, 7]	53	9.8	85.0
MAFormer-L	[128, 192, 448, 640]	[3, 8, 24, 7]	104	22.6	85.9

Table 2
Evaluating the components and different local aggregation methods using MAFormer-S.

Method	Params (M)	Local aggregation		GLD	Top1
		Eq. (1)	Local-window-attention	Eq. (2)	(%)
Swin-T	29M	–	Shifted window [10]	–	81.3
CSWin-T	23M	–	Cross-shaped [11]	–	82.7
MAFormer-S	23M	✓	Shifted window [10]	–	82.1
	23M	✓	Cross-shaped [11]	–	82.9
	23M	✓	Shifted window [10]	✓	83.4
	23M	✓	**Cross-shaped** [11]	✓	**83.7**

Table 3
Evaluating different global feature extraction operations.

Method	Local aggregation		Global extraction	Top1
	Eq. (1)	Local-window-attention		(%)
MAFormer-S	✓	Cross-shaped [11]	–	82.9
			Conv	83.4
			FC	80.3
			Pos	83.1
			GLD (Eq. (2))	**83.7**

4. Experiment

In this section, we first provide ablation studies of the MAF block. Then, we give the experimental results of MAFormer in three settings: image classification, object detection with instance segmentation and semantic segmentation. Specifically, we use ImageNet-1K [55] for classification, MSCOCO 2017 [56] with Mask R-CNN [57] and Cascade R-CNN [58] for object detection with instance segmentation, and ADE20K [59] for semantic segmentation, where we employ the semantic FPN [60] and UPerNet [46] as the basic framework. All experiments are conducted on V100 GPUs.

4.1. Ablation study

The multi-scale attention fusion (MAF) module includes three components: a Local Aggregation branch, the Global Learning with Down-sampling (GLD) module, and the Multi-scale Attention Fusion (MAF)

Table 4
Accuracy of MAFormer-S using different fusion structures.

Framework	Fusion structure design	Top1 (%)
DS-Net [18]	Co-Attention from convolution and self-attention	82.3
MAFormer-S	Local-enhanced fusion attention	83.5
	Multi-scale fusion attention (Eq. (4))	**83.7**

Table 5
Comparison with the state-of-the-art on ImageNet-1K.

Models	Train size	Test size	Params (M)	FLOPs (G)	Top1 (%)
DeiT-S [61]	224^2	224^2	22	4.6	79.8
Swin-T [10]	224^2	224^2	29	4.5	81.3
CrossViT-15 [62]	224^2	224^2	27	5.8	81.5
CoAtNet-0 [49]	224^2	224^2	25	4.6	81.6
Focal-T [63]	224^2	224^2	29	4.9	82.2
DS-Net-S [18]	224^2	224^2	23	3.5	82.3
Shuffle-T [12]	224^2	224^2	29	4.6	82.5
CSWin-T [11]	224^2	224^2	23	4.3	82.7
DaViT-Tiny [64]	224^2	224^2	28	4.5	82.8
MAFormer-S	224^2	224^2	23	4.5	**83.0**
LV-ViT-S[a] [65]	224^2	224^2	26	6.6	83.3
MAFormer-S[a]	224^2	224^2	23	4.5	**83.7**
CrossViT-18 [62]	224^2	224^2	44	9.5	82.8
Swin-S [10]	224^2	224^2	50	8.7	83.0
DS-Net-B [18]	224^2	224^2	49	8.4	83.1
Twins-SVT-B [66]	224^2	224^2	56	8.3	83.2
CoAtNet-1 [49]	224^2	224^2	42	8.4	83.3
Shuffle-S [12]	224^2	224^2	50	8.9	83.5
Focal-S [63]	224^2	224^2	51	9.1	83.5
CSWin-S [11]	224^2	224^2	35	8.9	83.6
LV-ViT-M[a] [65]	224^2	224^2	56	16	84.1
DaViT-Small [64]	224^2	224^2	49.7	8.8	84.2
MAFormer-B[a]	224^2	224^2	53	9.8	**85.0**
DeiT-B [61]	224^2	224^2	86	17.5	81.8
CrossViT-B [62]	224^2	224^2	105	21.2	82.2
Swin-B [10]	224^2	224^2	88	15.4	83.5
Focal-B [63]	224^2	224^2	90	16.0	83.8
Shuffle-B [12]	224^2	224^2	88	15.6	84
CSWin-B [11]	224^2	224^2	78	15.0	84.2
CoAtNet-3 [49]	224^2	224^2	168	34.7	84.5
DaViT-Base [64]	224^2	224^2	88	15.5	84.6
CaiT-M36 [67]	224^2	384^2	271	247.8	85.1
LV-ViT-L[a] [65]	288^2	288^2	150	59.0	85.3
MAFormer-L[a]	224^2	224^2	105	22.6	**85.9**

[a] Indicates with Token Labeling [65].

structure. In the following experiments, we conduct ablative experiments to examine the impact of these components based on the MAFormer-S network. All experiments are conducted on the image classification dataset ImageNet-1K.

Local aggregation. Firstly, we examine the Local Aggregation branch using different local-window-attention based methods. The selection of the local-window-attention methods in Eq. (2) is flexible [10,46,47]. We examine two approaches, the shifted window-based local-window-attention of Swin-T [10] and the cross-shaped local-window-attention of CSWin-T [11]. The two architectures achieve 81.3% and 82.7% top-1 accuracy on the imagenet respectively. Table 2 exhibits that incorporating the local-window-attention methods in the Local Aggregation branch brings consistent performance improvement, boosting the Swin-T [10] and CSWin-T [11] baseline by 0.8% and 0.2% top-1 accuracy respectively. The improvement validates our local aggregation design and model configurations.

As shown in Table 2, MAFormer-S using the cross-shaped approach outperforms the shifted window-based local-window-attention by 0.7% top-1 accuracy without GLD and 0.3% top-1 with GLD, which is set as the default method in the Local Aggregation branch.

Global feature extraction. Global information is vital to feature representation. We first exhibit the effectiveness of the GLD module. Table 2 shows that MAFormer-S with GLD yields 83.7% top-1 accuracy with the cross-shaped local-window-attention approach and 83.4% top-1 accuracy with the shifted window-based approach, bringing +0.8% and 1.2% improvements respectively, which is significant. The improvements validate the contribution of the GLD module.

Moreover, we examine different global feature extraction operations in Table 3. We use MAFormer-S with only the Local Aggregation branch as the baseline. As shown, GLD achieves +3.4% accuracy than utilizing only the FC and +0.6% than utilizing only the Pos operation. The results underscore that learning the global tokens coarsely or encoding local positional information alone cannot well capture the global dependencies. By combining these operations, our GLD module obtains more accurate learning and achieves optimal performance. Further, GLD achieves +0.5% accuracy than basic convolution, demonstrating that our method learns the detailed information from global tokens more effectively.

Fusion structure analysis. We explore different fusion structures in Table 4. As shown, our proposed Multi-scale Fusion Attention is more efficient than the previous local/global dual-stream architecture [18]. Also, our multi-scale attention fusion structure brings +0.2% increment over the local enhanced fusion measure (Fig. 2(b)). This indicates that learning local and global features independently achieves more comprehensive feature representations than learning the global dependencies between local attributes.

4.2. Image classification on ImageNet-1K

Settings. In this section, we conduct experiments of MAFormer on ImageNet-1K classification [55] and compare the proposed architecture with the previous state-of-the-arts. MAFormer follows [65] by default and is trained with Token Labeling [65]. Dropout regularization rate [68] is set as 0.1/0.3/0.4 for MAFormer-S/B/L respectively, as shown in Table 1. The learning rate of MAFormer-S and MAFormer-B are 1.6e−3, while for MAFormer-L it is 1.2e−3. All experiments are conducted on V100 GPUs.

Results. As shown in Table 1, MAFormer-S with only 23M parameters can achieve a top-1 accuracy of 83.7% on ImageNet-1k. Increasing the embedding dimension and network depth can further boost the performance. Table 5 shows in details that MAFormer outperforms the previous state-of-the-art vision transformers. Specifically, MAFormer-L achieves 85.9% Top-1 accuracy with 22.6G FLOPs, surpassing CSWin-B [11] and LV-ViT-L [65] by 1.7% and 0.6% respectively. MAFormer variants also outperform the prior art hybrid architectures [18,49] and local window-attention-based transformers [10,12,62] by large margins with a fair amount of computation.

We visualize comparisons of Params and FLOPs between models in Table 5 to demonstrate the efficiency of our method. As shown in Fig. 3(a), MAFormer-S achieves the optimal performance, a top-1 accuracy of 83.7% on ImageNet-1k, with less FLOPs than all other models and requiring less memory than all models except DeiT-S [61].

4.3. Object detection and instance segmentation on MSCOCO

Pre-training models on image classification resources and adapting them to downstream tasks has become the standard approach in most vision works. However, the data volume of downstream tasks is much lower than classification benchmarks, the ImageNet for instance. According to recent studies [53], the lower layers of attention-based networks perform poorly on aggregating local correlations when trained a small amount of data, given the lack of inductive bias. As a result, state-of-the-art transformer backbones on the ImageNet provide no significant improvement to downstream subtasks. MAFormer, on the other hand, utilize local window based attention in the lower layers and strategically encode global information with it. In this way, local patterns are easier to acquire when the training data is not sufficient, making it a general and efficient visual backbone.

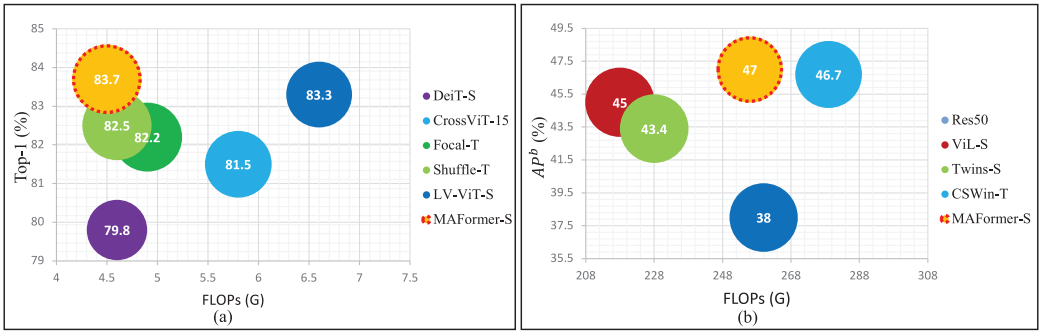


Fig. 3. The comparison of Params and FLOPs between different models with (a) image classification on ImageNet-1K and (b) object detection and instance segmentation on the COCO val2017 with the Mask R-CNN framework. The FLOPs are represented on the x-axis, the size of the circle represents Params, and the y-axis represents model performance.

Table 6
Object detection and instance segmentation performance on the COCO val2017 with the Mask R-CNN framework. The FLOPs (G) are measured at resolution 800 × 1280, and the models are pretrained on the ImageNet-1K.

Backbone	Params	FLOPs	Mask R-CNN 1x schedule					
	(M)	(G)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 [14]	44	260	38.0	58.6	41.4	34.4	55.1	36.7
PVT-S [52]	44	245	40.4	62.9	43.8	37.8	60.1	40.3
ViL-S [69]	45	218	44.9	67.1	49.3	41.	64.2	44.1
TwinsP-S [66]	44	245	42.9	65.8	47.1	40.4	62.7	42.9
Twins-S [66]	44	228	43.4	66.0	47.3	40.3	63.2	43.4
Swin-T [10]	48	264	42.2	64.6	46.2	39.1	64.6	42.0
CSWin-T [11]	42	279	46.7	68.6	51.3	42.2	65.6	45.4
MAFormer-S	41	256	**47.0**	**69.5**	**51.6**	**42.7**	**66.5**	**46.1**
Res101 [14]	63	336	40.4	61.1	44.2	36.4	57.7	38.8
X101-32 [70]	63	340	41.9	62.5	45.9	37.5	59.4	40.2
PVT-M [52]	64	302	42.0	64.4	45.6	39.0	61.6	42.1
ViL-M [69]	60	261	43.4	–	–	39.7	–	–
TwinsP-B [66]	64	302	44.6	66.7	48.9	40.9	63.8	44.2
Twins-B [66]	76	340	45.2	67.6	49.3	41.5	64.5	44.8
Swin-S [10]	69	354	44.8	66.6	48.9	40.9	63.4	44.2
CSWin-S [11]	54	342	47.9	70.1	52.6	43.2	67.1	46.2
MAFormer-B	71	354	**49.6**	**71.4**	**54.7**	**44.6**	**68.6**	**48.4**
X101-64 [70]	101	493	42.8	63.8	47.3	38.4	60.6	41.3
PVT-L [52]	81	364	42.9	65.0	46.6	39.5	61.9	42.5
ViL-B [69]	76	365	45.1	–	–	41.0	–	–
TwinsP-L [66]	81	364	45.4	–	–	41.5	–	–
Twins-L [66]	111	474	45.9	–	–	41.6	–	–
Swin-B [10]	107	496	46.9	–	–	42.3	–	–
CSWin-B [11]	97	526	48.7	70.4	53.9	43.9	67.8	47.3
MAFormer-L	122	609	**50.7**	**72.4**	**55.6**	**45.4**	**69.7**	**49.2**

Table 7
Object detection and instance segmentation performance on the COCO val2017 with the Cascade R-CNN framework. The FLOPs (G) are measured at resolution 800 × 1280, and the models are pretrained on the ImageNet-1K.

Backbone	Params	FLOPs	Cascade R-CNN 3x schedule					
	(M)	(G)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 [14]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T [10]	86	745	50.5	69.3	54.9	43.7	66.6	47.1
CSWin-T [11]	80	757	52.5	**71.5**	57.1	45.3	68.8	48.9
MAFormer-S	80	733	**52.6**	71.3	**57.3**	**45.7**	**68.9**	**49.8**
X101-32 [70]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S [10]	107	838	51.8	70.4	56.3	44.7	67.9	48.5
CSWin-S [11]	92	820	53.7	72.2	58.4	46.4	69.6	50.6
MAFormer-B	109	833	**54.4**	**72.8**	**59.2**	**46.8**	**70.4**	**51.0**
X101-64 [70]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B [10]	145	982	51.9	70.9	56.5	45.0	68.4	48.7
CSWin-B [11]	135	1005	53.9	72.6	58.5	46.4	70.0	50.4
MAFormer-L	160	1088	**54.7**	**73.2**	**59.4**	**47.3**	**71.2**	**51.3**

Settings. To demonstrate the merits of MAFormer on downstream tasks, we evaluate the model on COCO object detection task [56]. We first utilize the typical framework Mask R-CNN [57], where we configure 1x schedule with 12 epochs training schedules. In details, the shorter side of the image is resized to 800 while keeping the longer side no more than 1333. We utilize the same AdamW [71] optimizer with initial learning rate of 1e−4, decayed by 0.1 at epoch 8 and 11(1x schedule), and weight decay of 0.05. We set stochastic drop path regularization of 0.2 for MAFormer-S backbone, and 0.3 for MAFormer-B and MAFormer-L backbone, referred in Table 1.

To extend our research, we evaluate MAFormer in another typical framework Cascade R-CNN [58]. For Cascade R-CNN, we adopt 3x schedule with 36 epochs training schedules and the multi-scale training strategy [28,72] to randomly resize the shorter side between 480 to 800. We utilize the same AdamW [71] optimizer with initial learning rate of 1e−4, decayed by 0.1 at epoch 27 and 33, and weight decay of 0.05. We set stochastic drop path regularization of 0.2, 0.3, and 0.4 for MAFormer-S, MAFormer-B, and MAFormer-L backbone, respectively.

We compare MAFormer with various works: typical CNN backbones ResNet [14], ResNeXt [70], and competitive Transformer backbones PVT [52], Twins [66], Swin [10] and CSWin [11].

Results. Table 6 reports box mAP (AP^b) and mask mAP (AP^s) of the Mask R-CNN framework with 1x training schedule. It shows that the MAFormer variants notably outperform all the CNN and Transformer counterparts. Our MAFormer-S, MAFormer-B, and MAFormer-L achieve 47.0%, 49.6%, and 50.7% box mAP for object detection, surpassing the previous best CSWin Transformer by +0.3%, +1.7%, and +2.0%. Besides, our models present consistent improvement in instance segmentation, with +0.5%, +1.4%, and +1.5% mask mAP higher than the previous best backbone. We further visualize comparisons of Params and FLOPs between models. As shown in Fig. 3(b), MAFormer-S achieves the optimal performance with 47.0% box mAP for object detection, outperforming CSWin-T with far less parameters.

Table 7 contains the box mAP (AP^b) and mask mAP (AP^m) results from the Cascade R-CNN framework with 3x training schedule. It shows that MAFormer variants outperform all the CNN and Transformer counterparts in great margin. Specifically, MAFormer-S, MAFormer-B, and MAFormer-L achieve 52.6%, 54.4%, and 54.7% box mAP for object detection, surpassing the previous best CSWin Transformer by +0.1%, +0.7%, and +0.8%. Besides, our variants also have consistent improvement on instance segmentation, which are +0.3%, +0.4%, and +0.9% mask mAP higher than the previous best backbone. It shows with a stronger framework, MAFormer still surpass the counterparts by promising margins under different configurations.

4.4. Experiments of semantic segmentation with semantic FPN and UPerNet on ADE20K

Settings. ADE20K [59] is a widely used semantic segmentation dataset, covering a broad range of 150 semantic categories. It has 25K images in total, with 20K for training, 2K for validation, and another 3K for testing. We further investigate the capability of MAFormer for semantic segmentation on the ADE20K dataset. Here we employ

Table 8
Comparison with previous best results on ADE20K semantic segmentation. UPerNet: learning rate of 6×10^{-5}, a weight decay of 0.01, a scheduler that uses linear learning rate decay, and a linear warmup of 1500 iterations. Semantic FPN: learning rate of 2×10^{-4}, a weight decay of 1×10^{-4}, a scheduler that uses Cosine Annealing learning rate decay, and a linear warmup of 1000 iterations. The FLOPs are measured at resolution 2048 × 512.

Models	Semantic FPN 80K			UPerNet 160k			
	Params (M)	FLOPs (G)	mIoU (%)	Params (M)	FLOPs (G)	mIoU (%)	MS mIoU (%)
Res50 [14]	29	183	36.7	–	–	–	–
Twins-S [73]	28	144	43.2	54	901	46.2	47.1
TwinsP-S [73]	28	162	44.3	55	919	46.2	47.5
HRNet-w48 [26]	–	–	–	–	664	71	45.7
Swin-T [10]	32	182	41.5	60	945	44.5	45.8
Focal-T [63]	–	–	–	62	998	45.8	47.0
Shuffle-T [12]	–	–	–	60	949	46.6	47.6
MAFormer-S	28	170	**47.9**	52	929	**48.3**	**48.6**
Res101 [14]	48	260	38.8	86	1029	–	44.9
TwinsP-B [73]	48	220	44.9	74	977	47.1	48.4
Twins-B [73]	60	261	45.3	89	1020	47.7	48.9
Swin-S [10]	53	274	45.2	81	1038	47.6	49.5
Focal-S [63]	–	–	–	85	1130	48.0	50.0
Shuffle-S [12]	–	–	–	81	1044	48.4	49.6
Swin-B [10]	91	442	46.0	121	1188	48.1	49.1
MAFormer-B	55	274	**49.8**	82	1031	**51.1**	**51.6**

the semantic FPN [60]and UPerNet [46] as the basic framework. All experiments are conducted on 8 V100 GPUs. For fair comparison, we train Semantic FPN [60] 80k iterations with batch size as 16, and UPerNet [46] 160k iterations with the batch size as 16 and the image resolution is 512 × 512.

Results. Table 8, we provide the experimental results in terms of mIoU and Multi-scale tested mIoU (MS mIoU).

It shows that MAFormer-S, MAFormer-B achieve 47.9, 49.8 with the semantic FPN framework, 6.4 and 2.6 higher mIoU than the Swin-Transformer [10]. Also, MAFormer-S, MAFormer-B achieve 49.8, 51.1 with the UPerNet framework, 3.9, 3.0 higher mIoU than the Swin-Transformer [10].

5. Conclusion

In this paper, we introduce a general vision transformer backbone MAFormer, which integrates local and global features in tokens. MAFormer can improve the information interaction between local windows, where both local and global features are deployed with a linear operation to ensure the consistency of features distribution. Our MAFormer achieves state-of-the-art results on common vision tasks. In particular, MAFormer-L achieves 85.9% Top-1 accuracy on ImageNet, surpassing CSWin-B and LV-ViT-L by 1.7% and 0.6% respectively. On MSCOCO, MAFormer outperforms the prior art CSWin by 1.7% mAPs on object detection and 1.4% on instance segmentation with similar-sized parameters. The MAFormer's performance across visual recognition tasks underscores its effectiveness and versatility. Further, the MAFormer demonstrates promising potential as a general backbone network for future exploration, particularly in self-supervised pre-training tasks.

We also address the MAFormer's limitations, focusing on computational efficiency and domain-specific adaptability. Despite its state-of-the-art results on common vision tasks, MAFormer's computational efficiency is an area for improvement, especially in real-time processing or deployment on resource-constrained devices. Additionally, while the MAFormer demonstrates robust performance across on common vision tasks, its adaptability to specialized domains such as medical imaging or real-time video analysis remains essential for broadening its applicability.

Future work on the MAFormer can focus on enhancing computational efficiency, expanding its adaptability across diverse domains, and reducing dependence on large pre-training datasets. Addressing these limitations and pursuing these research avenues not only will refine the MAFormer's capabilities but also push the boundaries of vision transformer architectures in real-world applications.

CRediT authorship contribution statement

Huixin Sun: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Conceptualization. **Yunhao Wang:** Writing – original draft, Methodology. **Xiaodi Wang:** Software. **Bin Zhang:** Software. **Ying Xin:** Software. **Baochang Zhang:** Writing – review & editing, Writing – original draft. **Xianbin Cao:** Supervision, Project administration. **Errui Ding:** Supervision, Project administration. **Shumin Han:** Writing – review & editing, Writing – original draft.

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

Data will be made available on request.

Acknowledgments

The work was supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029). This research was also supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016, "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268, Beijing Municipal Science and Technology Commission, Administrative Commission of Zhongguancun Science Park Grant No. Z231100005923035, Taiyuan City "Double hundred Research action" 2024TYJB0127.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[4] M. Gao, F. Zheng, J.J. Yu, C. Shan, G. Ding, J. Han, Deep learning for video object segmentation: a review, Artif. Intell. Rev. 56 (1) (2023) 457–531.
[5] Y. Chen, R. Xia, K. Yang, K. Zou, MFMAM: Image inpainting via multi-scale feature module with attention module, Comput. Vis. Image Underst. 238 (2024) 103883.
[6] Y. Chen, R. Xia, K. Yang, K. Zou, GCAM: lightweight image inpainting via group convolution and attention mechanism, Int. J. Mach. Learn. Cybern. (2023) 1–11.
[7] Y. Chen, R. Xia, K. Yang, K. Zou, DGCA: high resolution image inpainting via DR-GAN and contextual attention, Multimedia Tools Appl. (2023) 1–21.
[8] Y. Chen, R. Xia, K. Yang, K. Zou, DARGS: Image inpainting algorithm via deep attention residuals group and semantics, J. King Saud Univ.-Comput. Inf. Sci. 35 (6) (2023) 101567.
[9] Y. Chen, R. Xia, K. Yang, K. Zou, MICU: Image super-resolution via multi-level information compensation and U-net, Expert Syst. Appl. 245 (2024) 123111.
[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[11] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, 2021, arXiv preprint arXiv:2107.00652.
[12] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, B. Fu, Shuffle transformer: Rethinking spatial shuffle for vision transformer, 2021, arXiv preprint arXiv:2106.03650.

[13] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.

[14] 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, CVPR, 2016.

[15] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[16] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, Q. Sun, Feature pyramid transformer, in: European Conference on Computer Vision, Springer, 2020.

[17] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, C. Zhang, ConTNet: Why not use convolution and transformer at the same time? 2021, arXiv preprint arXiv:2104.13497.

[18] M. Mao, R. Zhang, H. Zheng, T. Ma, Y. Peng, E. Ding, B. Zhang, S. Han, et al., Dual-stream network for visual recognition, Adv. Neural Inf. Process. Syst. 34 (2021).

[19] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3688–3704.

[20] Z. Shao, J. Han, K. Debattista, Y. Pang, Textual context-aware dense captioning with diverse words, IEEE Trans. Multimed. (2023).

[21] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[22] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size, 2016, arXiv preprint arXiv:1602.07360.

[23] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[25] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, Densenet: Implementing efficient convnet descriptor pyramids, 2014, arXiv preprint arXiv:1404.1869.

[26] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.

[27] Z. Shao, J. Han, D. Marnerides, K. Debattista, Region-object relation-aware dense captioning via transformer, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, Springer, 2020.

[29] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.

[30] Z. Yao, J. Ai, B. Li, C. Zhang, Efficient detr: improving end-to-end object detector with dense prior, 2021, arXiv preprint arXiv:2104.01318.

[31] T. Wang, L. Yuan, Y. Chen, J. Feng, S. Yan, PnP-DETR: towards efficient visual analysis with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[32] B. Roh, J. Shin, W. Shin, S. Kim, Sparse DETR: Efficient end-to-end object detection with learnable sparsity, in: The Tenth International Conference on Learning Representations 2022, Virtual Event, April 25-29, 2022, 2021.

[33] B. Cheng, A. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, Adv. Neural Inf. Process. Syst. 34 (2021).

[34] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[35] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021).

[36] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, E. Zhou, Tokenpose: Learning keypoint tokens for human pose estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[37] S. Yang, Z. Quan, M. Nie, W. Yang, Transpose: Towards explainable human pose estimation by transformer, 2020, arXiv e-prints.

[38] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, J. Wang, HRFormer: High-resolution transformer for dense prediction, 2021, arXiv preprint arXiv:2110.09408.

[39] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[40] C. Yan, T. Teng, Y. Liu, Y. Zhang, H. Wang, X. Ji, Precise no-reference image quality evaluation based on distortion identification, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 17 (3s) (2021) 1–21.

[41] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, B. Zheng, Age-invariant face recognition by multi-feature fusionand decomposition with self-attention, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 18 (1s) (2022) 1–18.

[42] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-trained image processing transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[43] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 43 (4) (2020) 1445–1451.

[44] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, J. Shlens, Scaling local self-attention for parameter efficient visual backbones, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[45] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, Adv. Neural Inf. Process. Syst. 32 (2019).

[46] J. Ho, N. Kalchbrenner, D. Weissenborn, T. Salimans, Axial attention in multidimensional transformers, 2019, arXiv preprint arXiv:1912.12180.

[47] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[48] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[49] Z. Dai, H. Liu, Q. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, Adv. Neural Inf. Process. Syst. 34 (2021).

[50] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[51] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[52] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[53] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks? Adv. Neural Inf. Process. Syst. 34 (2021).

[54] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, Y. Qiao, UniFormer: Unifying convolution and self-attention for visual recognition, 2022, arXiv preprint arXiv:2201.09450.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2009.

[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014.

[57] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.

[58] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.

[59] 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, CVPR, 2017.

[60] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6399–6408.

[61] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[62] 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.

[63] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal self-attention for local-global interactions in vision transformers, 2021, arXiv preprint arXiv:2107.00641.

[64] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, L. Yuan, Davit: Dual attention vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 74–92.

[65] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, J. Feng, All tokens matter: Token labeling for training better vision transformers, Adv. Neural Inf. Process. Syst. 34 (2021).

[66] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting spatial attention design in vision transformers, in: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual, 2021.

[67] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, H. Jégou, Going deeper with image transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 32–42.

[68] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014).

[69] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, J. Gao, Multi-scale vision longformer: A new vision transformer for high-resolution image encoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
[70] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[71] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations 2019, 2017.
[72] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al., Sparse r-cnn: End-to-end object detection with learnable proposals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[73] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 9355–9366.

Huixin Sun received her B.S. degree from Beijing University of Posts and Telecommunications in 2021 and is now pursuing Ph.D. at Beihang University. Her research interests are small object detection and network compression.

Yunhao Wang received the M.Eng. degree with the School of Electronic Engineering, Xidian University. He is currently a Engineer with Baidu Online Network Technology (Beijing) Company Ltd. His research interests include multimodal large model, image classification, object detection and self-supervised visual feature learning.

Xiaodi Wang received her M.Eng. degree with Beihang University. She is currently a Engineer with Baidu Online Network Technology (Beijing) Company Ltd. His research interests include multimodal large model, image classification, object detection and self-supervised visual feature learning.

Bin Zhang received her M.Eng. degree from Beijing University Of Posts and Telecommunications in 2010. Her research interests are object detection , Large Model Pre-training and Multimodal Foundation Models.

Ying Xin received her M.Eng. degree from Electronic Information Engineering at Tianjin University in 2019. Her research interests are object detection and Large Model Pre-training.

Baochang Zhang is a professor at Beihang University. He received his Ph.D. degrees in Computer Science from the Harbin Institute of the Technology, His current research interests include deep learning, pattern recognition, object detection, and wavelets.

Xianbin Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer applications and information science from Anhui University, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree in information science from the University of Science and Technology of China, Hefei, in 1996. He is currently a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China.

His current research interests include intelligent transportation systems, air traffic management, and intelligent computation.

Errui Ding received the Ph.D. degree from Xidian University in 2008. He is currently the Director of the Computer Vision Technology Department (VIS) and the Augmented Reality Department (AR), Baidu Inc. He serves as a member for the Special Committee of China Society of Image and Graphics.

Shumin Han received the M.S. degree in computer science from Peking University, in 2012. He is currently a Senior Engineer with Baidu Online Network Technology (Beijing) Company Ltd. His research interests include computer vision: image classification, object detection, object segmentation algorithms and applications, and self-supervised visual feature learning.