

MAFormer: A Transformer Network with Multi-scale Attention Fusion for Visual Recognition

Yunhao Wang^{1,†}, Huixin Sun^{2,†}, Xiaodi Wang¹, Bin Zhang¹, Chao Li¹,
Ying Xin¹, Baochang Zhang^{2,*}, Errui Ding¹, Shumin Han^{1,*}

¹Department of Computer Vision Technology (VIS), Baidu Inc

²Beihang University, Beijing, China

*Corresponding author, email: bczhang@buaa.edu.cn, hanshumin@baidu.com

[†]Equal contributions

Abstract

Vision Transformer and its variants have demonstrated great potential in various computer vision tasks. But conventional vision transformers often focus on global dependency at a coarse level, which suffer from 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 performance 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, demonstrating the potential to be a general backbone network.

1 Introduction

Transformers have prevailed in computation vision since the breakthrough of ViT [11], attaining excellent results in various visual tasks, including image recognition, object detection, and semantic segmentation. Despite these progress, the global self-attention mechanism in line with ViT [11, 21] 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 [24] 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 local feature capturing. CSWin Transformer [10] splits features into horizontal and vertical stripes in parallel, aiming to enlarge the window receptive field. However, it only focuses on the information within windows, leaving the dependencies across windows unexplored. Shuffle Transformer [16] revisits the ShuffleNet [26] 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 [13, 35]), they suffer from a learning challenge on the global relationship that is indispensable for a better feature representation.

Another line of research efforts focuses on combining CNNs with transformers, which are trade-offs between local patterns and global patterns. CvT [52] transforms the linear projection in the self-attention block into convolution projection. CoatNet [46] merges depth-wise convolution with self-attention via simple relative attention and stacks convolution and attention layers in a principled way. DS-Net [27] 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 [27], 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 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, we apply local window attention to extract fine-grained feature representation. 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 a more ample and informative representation. The contributions of this work are concluded 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

Vision transformers. Self-attention-based architectures, in particular Transformers [39], have become the dominant model for Natural Language Processing (NLP). Motivated by its success in NLP, ViT [11] innovatively applies a pure-transformer architecture to images by splitting an image into patches and equating them with tokens (words). ViT shows the competitive effect on classification tasks [9]. Many efforts have been devoted to applying ViT for various vision tasks, including object detection [2, 55, 49, 40, 31], semantic segmentation [5, 33, 44], pose estimation [22, 48, 51], re-identification [14], and low-level image processing [4]. These results verify the outstanding ability of the transformer as a general visual backbone. However, the self-attention mechanism is inefficient to encode low-level features, hindering their high potential for efficient representation learning.

Local window attention-based transformers. Vision transformers demonstrate a high capability in modeling the 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

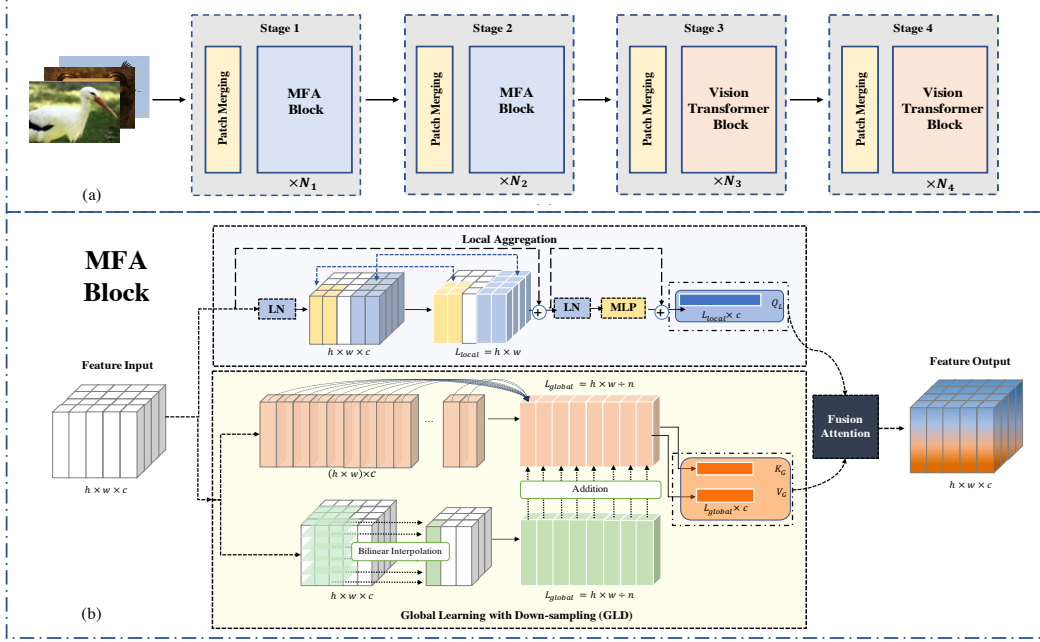


Figure 1: Architecture of MAFormer. We utilize the MAF block in the first two stages, which incorporates a Local Aggregation branch and a Global Learning with Down Sampling (GLD) branch. Both streams are fed into a fusion module to improve the capability of feature representation.

Transformers [24, 38] adopt the local window self-attention mechanism [30] and its shifted/haloed version that adds the interaction across different windows. To enlarge the receptive field, axial self-attention [15] and criss-cross attention [17] 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 [10] 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 analysis [28, 8], convolution networks and transformers hold different merits. While the convolution operation guarantees a better generalization and fast convergence, thanks to its inductive bias, attention formulate 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 [28] proposes the Feature Coupling Unit to fuse convolutional local features with transformer-based global representations in an interactive fashion. CvT [42] designs convolutional token embedding and convolutional transformer block for capturing more precise local spatial context. Followingly, CoatNet [8] merges depth-wise convolution into attention layers with simple relative attention. Apart from incorporating explicit convolution, some works [24, 10, 50, 41] 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 represents 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

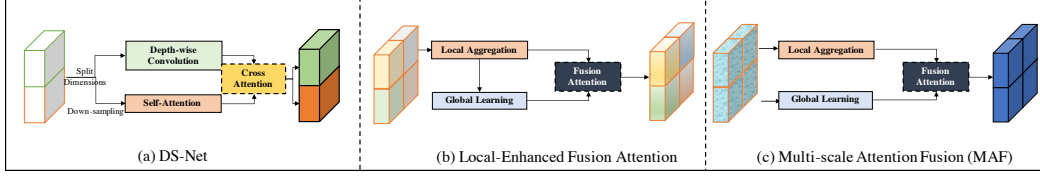


Figure 2: Different designs in dual-stream multi-scale representations.

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 $2\times$, while the feature channel dimension is increased.

According to recent studies into feature representations [29], 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 [8, 20] 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 the 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:

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

where X^l denotes the output of l -th 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 for model representation.

To address these issues, we introduce a Global Learning with Down-sampling (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 output a down-sampled contextual abstraction that is dynamically learned. As illustrated in the 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 operation utilizes a layer-wise bilinear interpolation as the measure and FC represents as the full connection.

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

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

Table 1: Detailed settings of MAFormer of different model sizes and their performance on ImageNet-1k validation set. 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

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, such approach can only capture the global correlations between local attributes, not from input. Therefore, we propose the Multi-scale attention fusion (MAF) measure, extracting the local and global scales of input directly and separately. Both stream 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:

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

where W_Q^{local} , W_K^{global} , W_V^{global} are learning hyper-parameter matrix. Then we calculate the Multi-scale Attention Fusion (MAF) between every pair of X_L and X_G :

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

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 [9] for classification, MSCOCO 2017 [23] with Mask R-CNN [12] and Cascade R-CNN [1] for object detection with instance segmentation, and ADE20K [54] for semantic segmentation, where we employ the semantic FPN [19] and UPerNet [43] as the basic framework. All experiments are conducted on V100 GPUs.

4.1 Ablation Study and Analysis

The multi-scale attention fusion (MAF) module in MAFormer network is mainly a composition of three: the Local Aggregation mechanism, the Global Learning with Down-sampling (GLD) module, and the dynamic fusion module. In the following experiments, we explore the best-performed structure of MAFormer by substituting and ablating different parts of the network. We set MAFormer-S as the baseline and all experiments are conducted on the image classification dataset ImageNet-1K.

Local aggregation. The selection of attention method in the Local Aggregation module is very flexible, which could be substituted by different approaches on window based self-attention [17, 24, 15]. In the MAF block, we compare the original work on window partition [24] and its recent variant cross-shaped window based self-attention [10]. As shown as Table 2, the experiments demonstrate that MAFormer-S using the cross-shaped window based self-attention outperforms shifted window-based self-attention by +0.2% top-1 accuracy on ImageNet 1K, which is set as the default approach.

Table 2: Ablation study of different local aggregation and global feature representation modules.

Method	Params	Attention in Local Aggregation	Global Feature Extraction	Top1 (%)
Swin-T	29M	Shifted Window [24]	None	81.3
CSWin-T	23M	Cross-shaped [10]	None	82.7
MAFormer-S	23M	Shifted Window [24]	GLD	83.4
MAFormer-S	23M	Cross-shaped [10]	Convolution	83.4
MAFormer-S	23M	Cross-shaped [10]	GLD	83.7

Table 3: Accuracy of MAFormer-S using different structure design.

Framework	Params(M)	Dual Stream Design	Top1 (%)
DS-Net [27]	23M	Co-Attention from Convolution and Self-attention	82.3
MAFormer-S	23M	Local-Enhanced Fusion Attention	83.5
MAFormer-S	23M	Multi-scale Fusion Attention	83.7

Global feature extraction. Global information is vital to feature representation. We show in Table 2 that MAFormer-S with GLD yields +1% top-1 accuracy than methods without global information on ImageNet-1K. We also compare GLD with other measures that extract global information and down-sample the input at the same time. As shown, GLD brings +0.3% accuracy than basic-configured convolution, demonstrating that the detailed information from global tokens can be extracted in a learnable and dynamic manner using GLD, with local positional information encoded.

Fusion structure analysis. The implementations of different connection modules are compared in Table 3. As shown, our proposed Multi-scale Fusion Attention is more efficient than the previous local/global dual-stream architecture [27]. Also, MAF is validated in our experiments with +0.2% superiority over the local enhanced fusion measure. Instead of fixing the fusion, cross-scale information transfers are automatically determined by feature themselves, making the combined more effective.

4.2 Image Classification on ImageNet-1K

Settings. In this section, we conduct experiments of MAFormer on ImageNet-1K classification [9] and compare the proposed architecture with the previous state-of-the-arts. MAFormer follows [18] by default and is trained with Token Labeling [18]. Dropout regularization rate [32] 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 4 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 [10] and LV-ViT-L [18] by 1.7% and 0.6% respectively. MAFormer variants also outperform the prior art hybrid architectures [8, 27] and local window-attention-based transformers [16, 3, 24] by large margins with a fair amount of computation.

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 [29], 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.

Table 4: Comparison with the state-of-the-art on ImageNet-1K. † indicates with Token Labeling [18].

Models	Train Size	Test Size	Params(M)	FLOPs(G)	Top1(%)
DeiT-S [36]	224 ²	224 ²	22	4.6	79.8
Swin-T [24]	224 ²	224 ²	29	4.5	81.3
CrossViT-15 [3]	224 ²	224 ²	27	5.8	81.5
CoAtNet-0 [8]	224 ²	224 ²	25	4.6	81.6
Focal-T [47]	224 ²	224 ²	29	4.9	82.2
DS-Net-S [27]	224 ²	224 ²	23	3.5	82.3
Shuffle-T [16]	224 ²	224 ²	29	4.6	82.5
CSWin-T [10]	224 ²	224 ²	23	4.3	82.7
MAFormer-S	224 ²	224 ²	23	4.5	83.0
LV-ViT-S† [18]	224 ²	224 ²	26	6.6	83.3
MAFormer-S†	224 ²	224 ²	23	4.5	83.7
CrossViT-18 [3]	224 ²	224 ²	44	9.5	82.8
Swin-S [24]	224 ²	224 ²	50	8.7	83.0
DS-Net-B [27]	224 ²	224 ²	49	8.4	83.1
Twins-SVT-B [6]	224 ²	224 ²	56	8.3	83.2
CoAtNet-1 [8]	224 ²	224 ²	42	8.4	83.3
Shuffle-S [16]	224 ²	224 ²	50	8.9	83.5
Focal-S [47]	224 ²	224 ²	51	9.1	83.5
CSWin-S [10]	224 ²	224 ²	35	8.9	83.6
LV-ViT-M† [18]	224 ²	224 ²	56	16	84.1
MAFormer-B†	224 ²	224 ²	53	9.8	85.0
DeiT-B [36]	224 ²	224 ²	86	17.5	81.8
CrossViT-B [3]	224 ²	224 ²	105	21.2	82.2
Swin-B [24]	224 ²	224 ²	88	15.4	83.5
Focal-B [47]	224 ²	224 ²	90	16.0	83.8
Shuffle-B [16]	224 ²	224 ²	88	15.6	84
CSWin-B [10]	224 ²	224 ²	78	15.0	84.2
CoAtNet-3 [8]	224 ²	224 ²	168	34.7	84.5
CaiT-M36 [37]	224 ²	384 ²	271	247.8	85.1
LV-ViT-L† [18]	288 ²	288 ²	150	59.0	85.3
MAFormer-L†	224 ²	224 ²	105	22.6	85.9

Settings. To demonstrate the merits of MAFormer on downstream tasks, we evaluate the model on COCO object detection task [23]. We first utilize the typical framework Mask R-CNN [12], 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 [25] 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 [1]. For Cascade R-CNN, we adopt 3x schedule with 36 epochs training schedules and the multi-scale training strategy [2, 34] to randomly resize the shorter side between 480 to 800. We utilize the same AdamW [25] 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 [13], ResNeXt [45], and competitive Transformer backbones PVT [41], Twins [6], Swin [24] and CSWin [10].

Results. Table 5 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%,

Table 5: Object detection and instance segmentation performance on the COCO val2017 with the Mask R-CNN framework. The FLOPs (G) are measured at resolution 800x1280, and the models are pretrained on the ImageNet-1K.

Backbone	Params (M)	FLOPs (G)	Mask R-CNN 1x schedule					
			AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m	AP_{75}^m
Res50 [13]	44	260	38.0	58.6	41.4	34.4	55.1	36.7
PVT-S [41]	44	245	40.4	62.9	43.8	37.8	60.1	40.3
ViL-S [53]	45	218	44.9	67.1	49.3	41.	64.2	44.1
TwinsP-S [6]	44	245	42.9	65.8	47.1	40.4	62.7	42.9
Twins-S [6]	44	228	43.4	66.0	47.3	40.3	63.2	43.4
Swin-T [24]	48	264	42.2	64.6	46.2	39.1	64.6	42.0
CSWin-T [10]	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 [13]	63	336	40.4	61.1	44.2	36.4	57.7	38.8
X101-32 [45]	63	340	41.9	62.5	45.9	37.5	59.4	40.2
PVT-M [41]	64	302	42.0	64.4	45.6	39.0	61.6	42.1
ViL-M [53]	60	261	43.4	—	—	39.7	—	—
TwinsP-B [6]	64	302	44.6	66.7	48.9	40.9	63.8	44.2
Twins-B [6]	76	340	45.2	67.6	49.3	41.5	64.5	44.8
Swin-S [24]	69	354	44.8	66.6	48.9	40.9	63.4	44.2
CSWin-S [10]	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 [45]	101	493	42.8	63.8	47.3	38.4	60.6	41.3
PVT-L [41]	81	364	42.9	65.0	46.6	39.5	61.9	42.5
ViL-B [53]	76	365	45.1	—	—	41.0	—	—
TwinsP-L [6]	81	364	45.4	—	—	41.5	—	—
Twins-L [6]	111	474	45.9	—	—	41.6	—	—
Swin-B [24]	107	496	46.9	—	—	42.3	—	—
CSWin-B [10]	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

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. Notably, MAFormer-B outperforms CSWin-S and Swin-S with far less parameters.

Table 6 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 [54] 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 the semantic FPN [19] and UPerNet [43] as the basic framework. All experiments are conducted on 8 V100 GPUs. For fair comparison, we train Semantic FPN [19] 80k iterations with batch size as 16, and UPerNet [43] 160k iterations with the batch size as 16 and the image resolution is 512x512.

Results. In Table 7, we provide the experimental results in terms of mIoU and Multi-scale tested mIoU (MS mIoU).

Table 6: Object detection and instance segmentation performance on the COCO val2017 with the Cascade R-CNN framework. The FLOPs (G) are measured at resolution 800x1280, and the models are pretrained on the ImageNet-1K.

Backbone	Params (M)	FLOPs (G)	Cascade R-CNN 3x schedule					
			AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m	AP_{75}^m
Res50 [13]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T [24]	86	745	50.5	69.3	54.9	43.7	66.6	47.1
CSWin-T [10]	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 [45]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S [24]	107	838	51.8	70.4	56.3	44.7	67.9	48.5
CSWin-S [10]	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 [45]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B [24]	145	982	51.9	70.9	56.5	45.0	68.4	48.7
CSWin-B [10]	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

Table 7: 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 1,500 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 1,000 iterations. The FLOPs are measured at resolution 2048x512.

Models	Semantic FPN 80K			UPerNet 160k			
	#Params(M)	FLOPs(G)	mIoU(%)	#Params(M)	FLOPs(G)	mIoU(%)	MS mIoU(%)
Res50 [13]	29	183	36.7	-	-	-	-
Twins-S [7]	28	144	43.2	54	901	46.2	47.1
TwinsP-S [7]	28	162	44.3	55	919	46.2	47.5
Swin-T [24]	32	182	41.5	60	945	44.5	45.8
Focal-T [47]	-	-	-	62	998	45.8	47.0
Shuffle-T [16]	-	-	-	60	949	46.6	47.6
MAFormer-S	28	170	47.9	52	929	48.3	48.6
Res101 [13]	48	260	38.8	86	1029	-	44.9
TwinsP-B [7]	48	220	44.9	74	977	47.1	48.4
Twins-B [7]	60	261	45.3	89	1020	47.7	48.9
Swin-S [24]	53	274	45.2	81	1038	47.6	49.5
Focal-S [47]	-	-	-	85	1130	48.0	50.0
Shuffle-S [16]	-	-	-	81	1044	48.4	49.6
Swin-B [24]	91	442	46.0	121	1188	48.1	49.1
MAFormer-B	55	274	49.8	82	1031	51.1	51.6

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 [24]. Also, MAFormer-S, MAFormer-B achieve 49.8, 51.1 with the UPerNet framework, 3.9, 3.0 higher mIoU than the Swin-Transformer [24].

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. With an outstanding performance on image classification and dense downstream tasks, MAFormer has shown its promising potential in vision tasks. In the future, MAFormer can be utilized as a general backbone in the self-supervised pre-training tasks.

References

- [1] Z. Cai and 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.
- [2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 2020.
- [3] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.
- [4] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [5] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.
- [6] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. Twins: Revisiting spatial attention design in vision transformers. *arXiv e-prints*, 2021.
- [7] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021.
- [8] Z. Dai, H. Liu, Q. Le, and M. Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34, 2021.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009.
- [10] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.
- [11] 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. *arXiv preprint arXiv:2010.11929*, 2020.
- [12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [14] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [15] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.
- [16] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021.
- [17] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
- [18] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [19] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019.
- [20] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022.
- [21] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021.

- [22] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 2014.
- [24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [25] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [26] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.
- [27] M. Mao, R. Zhang, H. Zheng, T. Ma, Y. Peng, E. Ding, B. Zhang, S. Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34, 2021.
- [28] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [29] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.
- [30] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [31] B. Roh, J. Shin, W. Shin, and S. Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.
- [32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 2014.
- [33] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [34] 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.
- [35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [36] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.
- [37] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.
- [38] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and 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.
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [40] T. Wang, L. Yuan, Y. Chen, J. Feng, and S. Yan. Pnp-detr: towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [41] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

- [42] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [43] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [44] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [46] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, and C. Zhang. Contnet: Why not use convolution and transformer at the same time? *arXiv preprint arXiv:2104.13497*, 2021.
- [47] 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. *arXiv preprint arXiv:2107.00641*, 2021.
- [48] S. Yang, Z. Quan, M. Nie, and W. Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv e-prints*, 2020.
- [49] Z. Yao, J. Ai, B. Li, and C. Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.
- [50] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and 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.
- [51] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021.
- [52] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun. Feature pyramid transformer. In *European Conference on Computer Vision*. Springer, 2020.
- [53] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and 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.
- [54] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [55] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.