MAFORMER: A TRANSFORMER NETWORK WITH MULTI-SCALE ATTENTION FUSION FOR VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

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 Dosovitskiy et al. (2020), 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 Li et al. (2021a) 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 Liu et al. (2021) 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. Shuffle Transformer Huang et al. (2021) revisits the ShuffleNet Ma et al. (2018) 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., ResNet He et al. (2016)), 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 Zhang et al. (2020) transforms the linear projection in the self-attention block into convolution projection. CoatNet Yan et al. (2021) merges depth-wise convolution with self-attention via simple relative attention and stacks convolution and attention layers in a principled way. DS-Net Mao et al. (2021) 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 Mao et al. (2021), convolution and attention hold intrinsically conflicting properties that might cause ambiguity in training. For instance, the

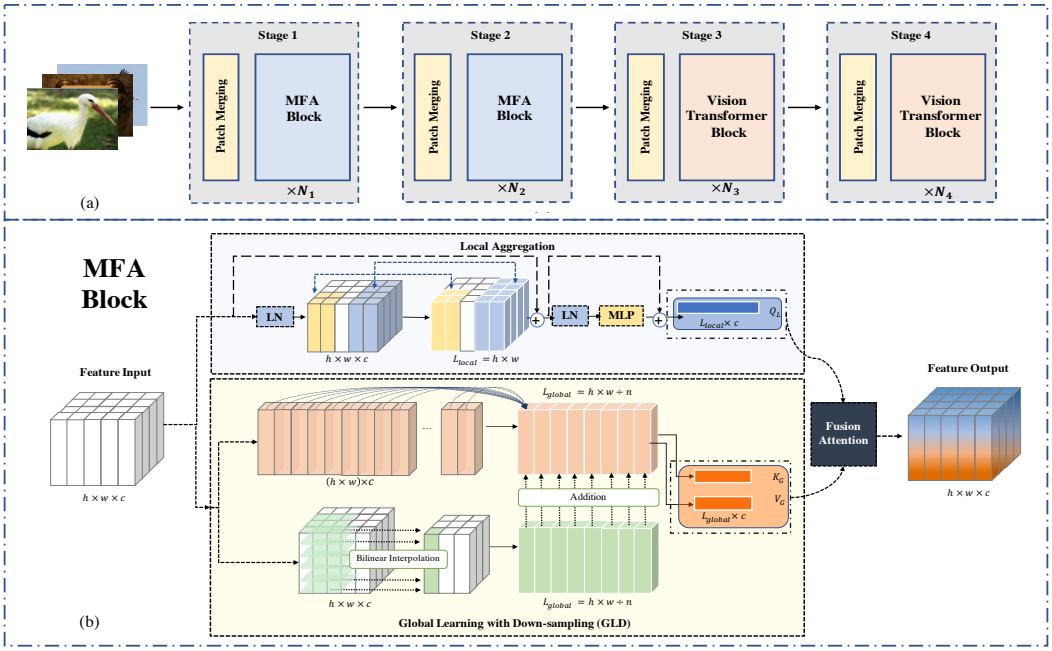

Figure 1: Architecture of MAFormer. We utilize the MAF block in the first two stages, which incorporates a Local Aggregation branch and a Global Learning with Down Sampling (GLD) branch. Both streams are fed into a fusion module to improve the capability of feature representation.

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 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

2.1 VISION TRANSFORMERS

Self-attention based architectures, in particular Transformers Vaswani et al. (2017), have become the dominant model for Natural Language Processing (NLP). Motivated by the success, ViT Dosovitskiy et al. (2020) applies a pure-transformer architecture to images by splitting an image into patches and equating them with tokens (words), which shows strong performance on image classification tasks Deng et al. (2009). Many efforts have been devoted to applying ViT for various vision tasks since, including object detection Carion et al. (2020); Zhu et al. (2020); Roh et al. (2021), semantic segmentation Cheng et al. (2021); Strudel et al. (2021); Xie et al. (2021) , pose estimation Li et al. (2021b); Yang et al. (2020); Yuan et al. (2021b), re-identification He et al. (2021), and low-level image processing Chen et al. (2021b). These results further validate the outstanding generality of the transformer as a visual backbone.

2.2 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 Transformers Liu et al. (2021); Vaswani et al. (2021) adopt the local window self-attention mechanism Ramachandran et al. (2019) and its shifted/haloed version that adds the interaction across different windows. To enlarge the receptive field, axial self-attention Ho et al. (2019) and criss-cross attention Huang et al. (2019) propose calculating attention within stripes along horizontal or/and vertical axis instead of fixing local windows as squares. The method Dong et al. (2021) presents the Cross-Shaped Window self-attention,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.

2.3 CONVOLUTION IN TRANSFORMERS

According to recent analysis Peng et al. (2021); Dai et al. (2021), 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 Peng et al. (2021) proposes the Feature Coupling Unit to fuse convolutional local features with transformer-based global representations in an interactive fashion. CvT Wu et al. (2021) designs convolutional token embedding and convolutional transformer block for capturing more precise local spatial context. Apart from incorporating explicit convolution, some works Liu et al. (2021); Dong et al. (2021); Yuan et al. (2021a); Wang et al. (2021) 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 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

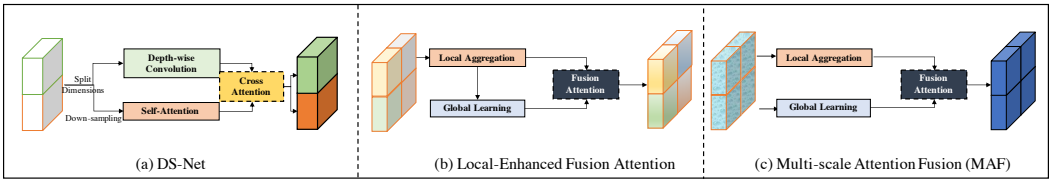

Figure 2: Different designs in dual-stream multi-scale representations.

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

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 Raghu et al. (2021), 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 Dai et al. (2021); Li et al. (2022) 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:

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

where 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.

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

where X_G^l denotes the global branch output of l-th 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, 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, the MAF block can capture the correlations between each local-global token pair, enabling the local features to adapt to the global representations.

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:

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

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:

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

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 Deng et al. (2009) for classification, MSCOCO 2017 Lin et al. (2014) with Mask R-CNN He et al. (2017) and Cascade R-CNN Cai & Vasconcelos (2018) for object detection with instance segmentation, and ADE20K Zhou et al. (2017) for semantic segmentation, where we employ the semantic FPN Kirillov et al. (2019) and UPerNet Xiao et al. (2018) 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 Huang et al. (2019); Liu et al. (2021); Ho et al. (2019). In the MAF block, we compare the original work on window partition Liu et al. (2021) and its recent variant cross-shaped window based self-attention Dong et al. (2021). 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 Liu et al. (2021)	None	81.3
CSWin-T	23M	Cross-shaped Dong et al. (2021)	None	82.7
MAFormer-S	23M	Shifted Window Liu et al. (2021)	GLD	83.4
MAFormer-S	23M	Cross-shaped Dong et al. (2021)	Convolution	83.4
MAFormer-S	**23M**	**Cross-shaped Dong et al. (2021)**	**GLD**	**83.7**

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

Framework	Params(M)	Dual Stream Design	Top1 (%)
DS-Net Mao et al. (2021)	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 Mao et al. (2021). 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 Deng et al. (2009) and compare the proposed architecture with the previous state-of-the-arts. MAFormer follows Jiang et al. (2021) by default and is trained with Token Labeling Jiang et al. (2021). Dropout regularization rate Srivastava et al. (2014) 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 Dong et al. (2021) and LV-ViT-L Jiang et al. (2021) by 1.7% and 0.6% respectively. MAFormer variants also outperform the prior art hybrid architectures Dai et al. (2021); Mao et al. (2021) and local window-attention-based transformers Huang et al. (2021); Chen et al. (2021a); Liu et al. (2021) by large margins with a fair amount of computation.

4.3 Object Detection and Instance Segmentation on MSCOCO

According to recent studies Raghu et al. (2021), 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.

Settings. To demonstrate the merits of MAFormer on downstream tasks, we evaluate the model on COCO object detection task Lin et al. (2014). We first utilize the typical framework Mask R-

Table 4: Comparison with the state-of-the-art on ImageNet-1K. † indicates with Token Labeling Jiang et al. (2021).

Models	Train Size	Test Size	Params(M)	FLOPs(G)	Top1(%)
DeiT-S Touvron et al. (2021a)	224^2	224^2	22	4.6	79.8
Swin-T Liu et al. (2021)	224^2	224^2	29	4.5	81.3
CrossViT-15 Chen et al. (2021a)	224^2	224^2	27	5.8	81.5
CoAtNet-0 Dai et al. (2021)	224^2	224^2	25	4.6	81.6
Focal-T Yang et al. (2021)	224^2	224^2	29	4.9	82.2
DS-Net-S Mao et al. (2021)	224^2	224^2	23	3.5	82.3
Shuffle-T Huang et al. (2021)	224^2	224^2	29	4.6	82.5
CSWin-T Dong et al. (2021)	224^2	224^2	23	4.3	82.7
MAFormer-S	224^2	224^2	23	4.5	**83.0**
LV-ViT-S† Jiang et al. (2021)	224^2	224^2	26	6.6	83.3
MAFormer-S†	224^2	224^2	23	4.5	**83.7**
CrossViT-18 Chen et al. (2021a)	224^2	224^2	44	9.5	82.8
MixFormer-B4 Chen et al. (2022)	224^2	224^2	35	3.6	83.0
Swin-S Liu et al. (2021)	224^2	224^2	50	8.7	83.0
DS-Net-B Mao et al. (2021)	224^2	224^2	49	8.4	83.1
Twins-SVT-B Chu et al. (2021b)	224^2	224^2	56	8.3	83.2
CoAtNet-1 Dai et al. (2021)	224^2	224^2	42	8.4	83.3
Shuffle-S Huang et al. (2021)	224^2	224^2	50	8.9	83.5
Focal-S Yang et al. (2021)	224^2	224^2	51	9.1	83.5
CSWin-S Dong et al. (2021)	224^2	224^2	35	8.9	83.6
LV-ViT-M† Jiang et al. (2021)	224^2	224^2	56	16	84.1
MAFormer-B†	224^2	224^2	53	9.8	**85.0**
DeiT-B Touvron et al. (2021a)	224^2	224^2	86	17.5	81.8
CrossViT-B Chen et al. (2021a)	224^2	224^2	105	21.2	82.2
Swin-B Liu et al. (2021)	224^2	224^2	88	15.4	83.5
Focal-B Yang et al. (2021)	224^2	224^2	90	16.0	83.8
Shuffle-B Huang et al. (2021)	224^2	224^2	88	15.6	84
CSWin-B Dong et al. (2021)	224^2	224^2	78	15.0	84.2
CoAtNet-3 Dai et al. (2021)	224^2	224^2	168	34.7	84.5
CaiT-M36 Touvron et al. (2021b)	224^2	384^2	271	247.8	85.1
LV-ViT-L† Jiang et al. (2021)	288^2	288^2	150	59.0	85.3
MAFormer-L†	224^2	224^2	105	22.6	**85.9**

CNN He et al. (2017), 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 Loshchilov & Hutter (2017) 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 Cai & Vasconcelos (2018). For Cascade R-CNN, we adopt 3x schedule with 36 epochs training schedules and the multi-scale training strategy Carion et al. (2020); Sun et al. (2021) to randomly resize the shorter side between 480 to 800. We utilize the same AdamW Loshchilov & Hutter (2017) 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 He et al. (2016), ResNeXt Xie et al. (2017), and competitive Transformer backbones PVT Wang et al. (2021), Twins Chu et al. (2021b), Swin Liu et al. (2021) and CSWin Dong et al. (2021).

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

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^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 He et al. (2016)	44	260	38.0	58.6	41.4	34.4	55.1	36.7
PVT-S Wang et al. (2021)	44	245	40.4	62.9	43.8	37.8	60.1	40.3
ViL-S Zhang et al. (2021)	45	218	44.9	67.1	49.3	41.	64.2	44.1
TwinsP-S Chu et al. (2021b)	44	245	42.9	65.8	47.1	40.4	62.7	42.9
Twins-S Chu et al. (2021b)	44	228	43.4	66.0	47.3	40.3	63.2	43.4
Swin-T Liu et al. (2021)	48	264	42.2	64.6	46.2	39.1	64.6	42.0
CSWin-T Dong et al. (2021)	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 He et al. (2016)	63	336	40.4	61.1	44.2	36.4	57.7	38.8
X101-32 Xie et al. (2017)	63	340	41.9	62.5	45.9	37.5	59.4	40.2
PVT-M Wang et al. (2021)	64	302	42.0	64.4	45.6	39.0	61.6	42.1
ViL-M Zhang et al. (2021)	60	261	43.4	–	–	39.7	–	–
TwinsP-B Chu et al. (2021b)	64	302	44.6	66.7	48.9	40.9	63.8	44.2
MixFormer-B4 Chen et al. (2022)	53	243	45.1	67.1	49.2	41.2	64.3	44.1
Twins-B Chu et al. (2021b)	76	340	45.2	67.6	49.3	41.5	64.5	44.8
Swin-S Liu et al. (2021)	69	354	44.8	66.6	48.9	40.9	63.4	44.2
CSWin-S Dong et al. (2021)	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 Xie et al. (2017)	101	493	42.8	63.8	47.3	38.4	60.6	41.3
PVT-L Wang et al. (2021)	81	364	42.9	65.0	46.6	39.5	61.9	42.5
ViL-B Zhang et al. (2021)	76	365	45.1	–	–	41.0	–	–
TwinsP-L Chu et al. (2021b)	81	364	45.4	–	–	41.5	–	–
Twins-L Chu et al. (2021b)	111	474	45.9	–	–	41.6	–	–
Swin-B Liu et al. (2021)	107	496	46.9	–	–	42.3	–	–
CSWin-B Dong et al. (2021)	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 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^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 He et al. (2016)	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T Liu et al. (2021)	86	745	50.5	69.3	54.9	43.7	66.6	47.1
CSWin-T Dong et al. (2021)	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 Xie et al. (2017)	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S Liu et al. (2021)	107	838	51.8	70.4	56.3	44.7	67.9	48.5
CSWin-S Dong et al. (2021)	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 Xie et al. (2017)	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B Liu et al. (2021)	145	982	51.9	70.9	56.5	45.0	68.4	48.7
CSWin-B Dong et al. (2021)	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**

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. 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

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 2048×512.

Models	Semantic FPN 80K			UPerNet 160k			
	#Params(M)	FLOPs(G)	mIoU(%)	#Params(M)	FLOPs(G)	mIoU(%)	MS mIoU(%)
Res50 He et al. (2016)	29	183	36.7	-	-	-	-
Twins-S Chu et al. (2021a)	28	144	43.2	54	901	46.2	47.1
TwinsP-S Chu et al. (2021a)	28	162	44.3	55	919	46.2	47.5
Swin-T Liu et al. (2021)	32	182	41.5	60	945	44.5	45.8
Focal-T Yang et al. (2021)	-	-	-	62	998	45.8	47.0
Shuffle-T Huang et al. (2021)	-	-	-	60	949	46.6	47.6
MAFormer-S	28	170	**47.9**	52	929	**48.3**	**48.6**
Res101 He et al. (2016)	48	260	38.8	86	1029	-	44.9
TwinsP-B Chu et al. (2021a)	48	220	44.9	74	977	47.1	48.4
Twins-B Chu et al. (2021a)	60	261	45.3	89	1020	47.7	48.9
Swin-S Liu et al. (2021)	53	274	45.2	81	1038	47.6	49.5
Focal-S Yang et al. (2021)	-	-	-	85	1130	48.0	50.0
Shuffle-S Huang et al. (2021)	-	-	-	81	1044	48.4	49.6
Swin-B Liu et al. (2021)	91	442	46.0	121	1188	48.1	49.1
MAFormer-B	55	274	**49.8**	82	1031	**51.1**	**51.6**

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 Zhou et al. (2017) 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 Kirillov et al. (2019)and UPerNet Xiao et al. (2018) as the basic framework. All experiments are conducted on 8 V100 GPUs. For fair comparison, we train Semantic FPN Kirillov et al. (2019) 80k iterations with batch size as 16, and UPerNet Xiao et al. (2018) 160k iterations with the batch size as 16 and the image resolution is 512×512.

Results. In Table 7, 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 Liu et al. (2021). Also, MAFormer-S, MAFormer-B achieve 49.8, 51.1 with the UPerNet framework, 3.9, 3.0 higher mIoU than the Swin-Transformer Liu et al. (2021).

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

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 2020.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021a.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b.

Qiang Chen, Qiman Wu, Jian Wang, Qinghao Hu, Tao Hu, Errui Ding, Jian Cheng, and Jingdong Wang. Mixformer: Mixing features across windows and dimensions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5259, 2022.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021a.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. *arXiv e-prints*, 2021b.

Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Ieee, 2009.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint arXiv:2107.00652*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021.

Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6399–6408, 2019.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021a.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *Advances in Neural Information Processing Systems*, 32, 2019.

Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 2014.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu 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.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Haotian Yan, Zhe Li, Weijian Li, Changhu Wang, Ming Wu, and Chuang Zhang. Contnet: Why not use convolution and transformer at the same time? *arXiv preprint arXiv:2104.13497*, 2021.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Towards explainable human pose estimation by transformer. *arXiv e-prints*, 2020.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021a.

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021b.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *European Conference on Computer Vision*. Springer, 2020.

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng 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.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.