DualToken-ViT: Position-aware Efficient Vision Transformer with Dual Token Fusion

Zhenzhen Chu [*] Jiayu Chen [†] Cen Chen [*] [‡] Chengyu Wang [†] Ziheng Wu [†]
Jun Huang [†] Weining Qian [*]

Abstract

Self-attention-based vision transformers (ViTs) have emerged as a highly competitive architecture in computer vision. Unlike convolutional neural networks (CNNs), ViTs are capable of global information sharing. With the development of various structures of ViTs, ViTs are increasingly advantageous for many vision tasks. However, the quadratic complexity of self-attention renders ViTs computationally intensive, and their lack of inductive biases of locality and translation equivariance demands larger model sizes compared to CNNs to effectively learn visual features. In this paper, we propose a light-weight and efficient vision transformer model called DualToken-ViT that leverages the advantages of CNNs and ViTs. DualToken-ViT effectively fuses the token with local information obtained by convolution-based structure and the token with global information obtained by self-attention-based structure to achieve an efficient attention structure. In addition, we use position-aware global tokens throughout all stages to enrich the global information, which further strengthening the effect of DualToken-ViT. Position-aware global tokens also contain the position information of the image, which makes our model better for vision tasks. We conducted extensive experiments on image classification, object detection and semantic segmentation tasks to demonstrate the effectiveness of DualToken-ViT. On the ImageNet-1K dataset, our models of different scales achieve accuracies of 75.4% and 79.4% with only 0.5G and 1.0G FLOPs, respectively, and our model with 1.0G FLOPs outperforms LightViT-T using global tokens by 0.7%.

Keywords: Vision Transformers, Attention

1 Introduction

In recent years, vision transformers (ViTs) have emerged as a powerful architecture for various vision tasks such as image classification [9] and object detection [3]. This is due to the ability of self-attention to capture global information from the image, providing sufficient and useful visual features, while convolutional neural networks (CNNs) are limited by the size of convolutional kernel and can only extract local information. As the model size of ViTs and the dataset size increase, there is still no sign of a saturation in performance, which is an advantage that CNNs do not have for large models as well as for large datasets [9]. However, CNNs are more advantageous than ViTs in light-weight models due to certain inductive biases that ViTs lack. Since the quadratic complexity of self-attention, the computational cost of ViTs can also be high. Therefore, it is challenging to design light-

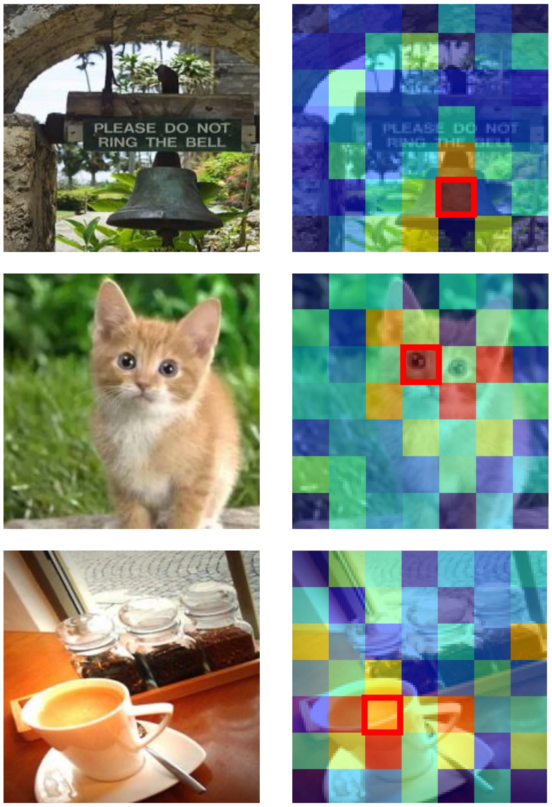

Figure 1: Visualization of the attention map of position-aware global tokens and the key token (the most important part of the image for the image classification task). In each row, the first image is the input of our model, and the second image represents the correlation between the red-boxed portion and each token in the position-aware global tokens containing 7×7 tokens, where the red-boxed portion is the key token of the first image.

[*] East China Normal University. {51215903091}@stu.ecnu.edu.cn and {cenchen, wnqian}@dase.ecnu.edu.cn

[†] Alibaba Group. {yunji.cjy, chengyu.wcy, zhoulou.wzh, huangjun.hj}@alibaba-inc.com

[‡] Cen Chen is the corresponding author.

weight-based efficient ViTs.

To design more efficient and light-weight ViTs, some works [30, 6] propose a pyramid structure that divides the model into several stages, with the number of tokens decreasing and the number of channels increasing by stage. And some works[22, 2] focus on reducing the quadratic complexity of self-attention by simplifying and improving the structure of self-attention, but they sacrifice the effectiveness of attention. Reducing the number of tokens involved in self-attention is also a common approach, e.g., PVTv1 [30], PVTv2 [31] and LITv2 [24] downsample the key and value in self-attention. Some works [18, 8] based on locally-grouped self-attention reduce the complexity of the overall attention part by performing self-attention on grouped tokens separately, but such methods may damage the sharing of global information. Some works also add a few additional learnable parameters to enrich the global information of the backbone, e.g., LightViT [13] and Mobile-Former [5] add a branch of global tokens that throughout all stages. This method can supplement global information for local attention (such as locally-grouped self-attention based and convolution-based structures). These existing methods using global tokens, however, consider only global information and ignore positional information that is very useful for vision tasks.

In this paper, we propose a light-weight and efficient vision transformer model called DualToken-ViT. Our proposed model features a more efficient attention structure designed to replace self-attention. We combine the advantages of convolution and self-attention, leveraging them to extract local and global information respectively, and then fuse the outputs of both to achieve an efficient attention structure. Although window self-attention [18] is also able to extract local information, we observe that it is less efficient than the convolution on our light-weight model. To reduce the computational complexity of self-attention in global information broadcasting, we downsample the feature map that produces key and value by step-wise downsampling, which can retain more information during the downsampling process. Moreover, we use position-aware global tokens throughout all stages to further enrich the global information. In contrast to the normal global tokens [13, 5, 34], our position-aware global tokens are also able to retain position information of the image and pass it on, which can give our model an advantage in vision tasks. As shown in Figure 1, the key token in the image generates higher correlation with the corresponding tokens in the position-aware global tokens, which demonstrates the effectiveness of our position-aware global tokens. In summary, our contributions are as follows:

- We design a light-weight and efficient vision transformer model called DualToken-ViT, which combines the advantages of convolution and self-attention to achieve an efficient attention structure by fusing local and global tokens containing local and global information, respectively.
- We further propose position-aware global tokens that contain the position information of the image to enrich the global information.
- Among vision models of the same FLOPs magnitude, our DualToken-ViT shows the best performance on the tasks of image classification, object detection and semantic segmentation.

2 Related work

Efficient Vision Transformers. ViTs are first proposed by [9], which applies transformer-based structures to computer vision. PVTv1 [30] applies the pyramid structure to ViTs, which will incrementally transform the spatial information into the rich semantic information. To achieve efficient ViTs, some works are beginning to find suitable alternatives to self-attention in computer vision tasks, such as [22, 2], which make the model smaller by reducing the complexity of self-attention. Some works [30, 31, 24] reduce the required computational resources by reducing the number of tokens involved in self-attention. Some works [18, 8] use locally-grouped self-attention based methods to reduce the complexity of the overall attention part. There are also some works that combine convolution into ViTs, for example, PVTv2 [31] uses convolution-based FFN (feed-forward neural network) to replace the normal FFN, LITv2 [24] uses more convolution-based structure in the shallow stages of the model and more transformer-based structure in the deep stages of the model. Moreover, there are also many works that use local information extracted by convolution or window self-attention to compensate for the shortcomings of ViTs, such as [21, 23].

Efficient Attention Structures. For local attention, convolution works well for extracting local information in vision tasks, e.g., MobileViTv1 [21] and EdgeViT [23] add convolution to model to aggregate local information. Among transformer-based structures, locally-grouped self-attention [18, 8] can also achieve local attention by adjusting the window size, and their complexity will be much less than that of self-attention. For global attention, self-attention proposed in [29] has a strong ability to extract global information, but on light-weight models, it may not be able to extract visual features well due to the lack of model size. Methods [13, 5, 34] using global tokens can also aggregate global information. They use self-attention to update global tokens and broadcast global information. Since the number of tokens in global tokens will not be set very large, the complexity will not be very high. Some works [23, 13, 24, 5] achieve a more efficient attention structure by combining both local and global attention. In this paper, we implement an efficient attention structure by combining convolution-based lo-

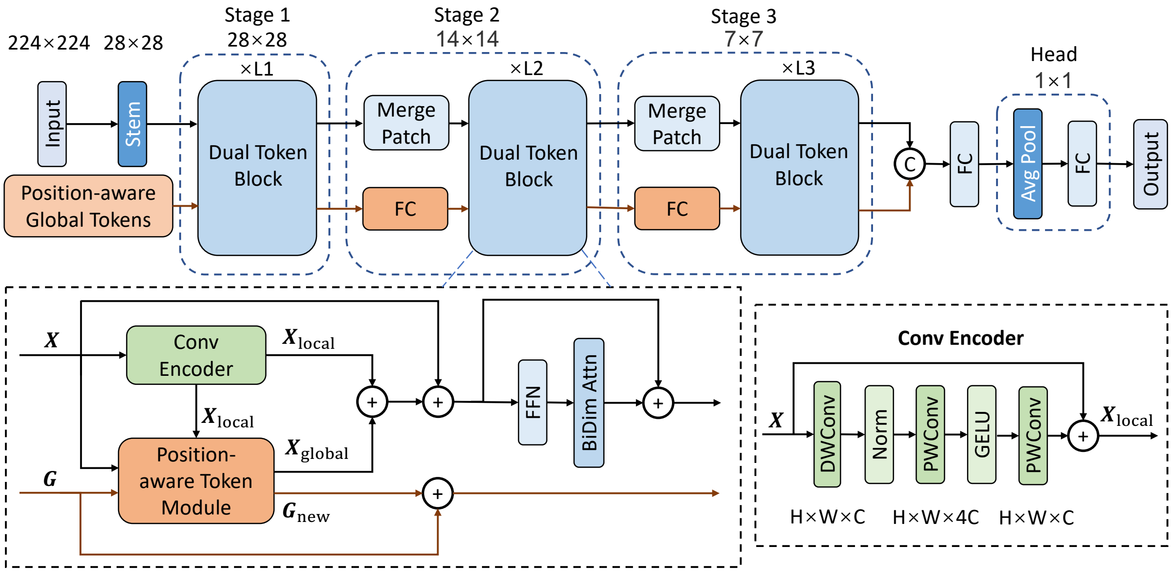

Figure 2: The architecture of DualToken-ViT. ⊕ represents element-wise addition. Ⓒ represents concatenation in the token axis.

cal attention and self-attention-based global attention, and use another branch of position-aware global tokens for the delivery of global and position information throughout the model, where position-aware global tokens are an improvement over global tokens [13, 5, 34].

3 Methodology

As shown in Figure 2, DualToken-ViT is designed based on the 3-stage structure of LightViT [13]. The structure of stem and merge patch block in our model is the same as the corresponding part in LightViT. FC refers to fully connected layer. There are two branches in our model: image tokens and position-aware global tokens. The branch of image tokens is responsible for obtaining various information from position-aware global tokens, and the branch of position-aware global tokens is responsible for updating position-aware global tokens through the branch of image tokens and passing it on. In the attention part of each Dual Token Block, we obtain information from the position-aware global tokens and fuse local and global information. We also add BiDim Attn (bi-dimensional attention) proposed in LightViT after the FFN. In this section, we mainly introduce two important parts: the fusion of local and global information and position-aware global tokens.

3.1 Fusion of Local and Global Information In the attention part of each Dual Token Block, we extract the local and global information through two branches, Conv Encoder (convolution encoder) and Position-aware Token Module, respectively, and then fuse these two parts.

Local Attention. We use Conv Encoder for local information extraction in each block of our model, since for lightweight models, local information extraction with convolution will perform better than window self-attention. Conv Encoder has the same structure as the ConvNeXt block [19], which is represented as follows:

$$\boldsymbol{X}_{\text{local}} = \boldsymbol{X} + \text{PW}_2(\text{GELU}(\text{PW}_1(\text{LN}(\text{DW}(\boldsymbol{X}))))) \tag{3.1}$$

where $\boldsymbol{X}$ is the input image tokens of size H×W×C, DW is the depth-wise convolution, PW_1 and PW_2 are point-wise convolution, LN is the layer norm, and $\boldsymbol{X}_{\text{local}}$ containing local information is the output of Conv Encoder.

Position-aware Token Module. This module is responsible for extracting global information, and its structure is shown in Figure 3(b). In order to reduce the complexity of extracting global information, we first downsample $\boldsymbol{X}_{\text{local}}$ containing local information and aggregate the global information. Position-aware global tokens are then used to enrich global information. We end up broadcasting this global information to image tokens. The detailed process is as follows:

(1) Downsampling. If the size of $\boldsymbol{X}_{\text{local}}$ is large and does not match the expected size, then it is downsampled twice first. After that, local information is extracted by convolution and downsampled twice, and the process is repeated M times until the feature map size reaches the expected size. Compared with the one-step downsampling method, this step-wise downsampling method can reduce the loss of information during the downsampling process and retain more useful information. The entire step-wise downsampling process is represented as follows:

$$\boldsymbol{X}_{\text{ds}} = \phi(\text{DS}(\boldsymbol{X}_{\text{local}})) \tag{3.2}$$

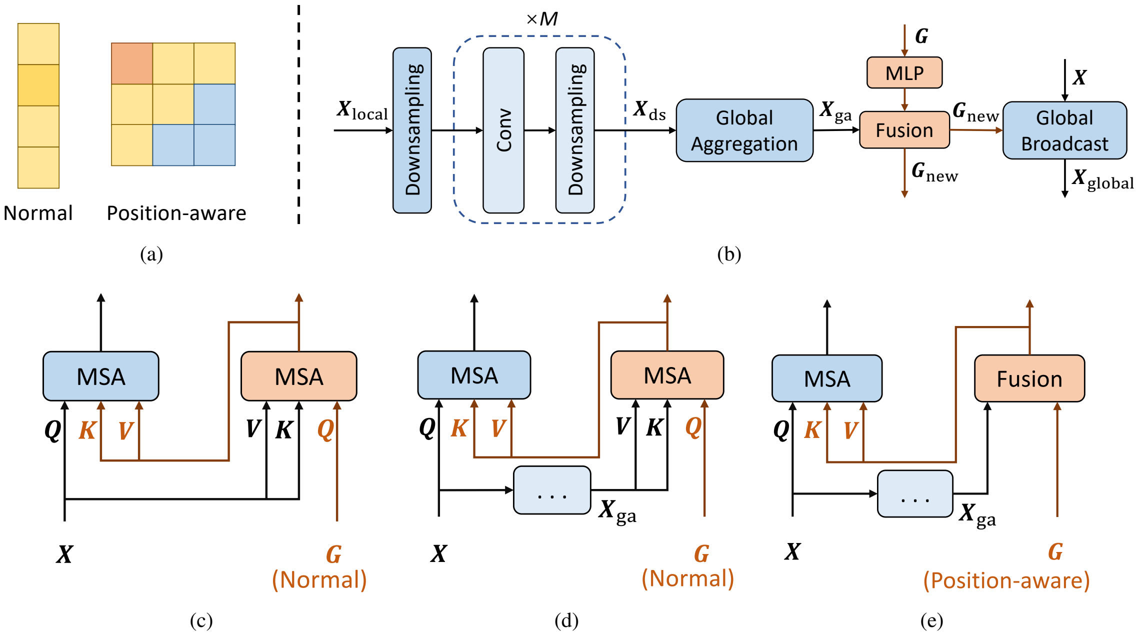

Figure 3: (a) shows normal global tokens and position-aware global tokens. (b) shows the structure of Position-aware Token Module using position-aware global tokens. (c), (d) and (e) show different methods of applying global tokens and show only the interaction of $\boldsymbol{X}$ and $\boldsymbol{G}$, omitting the processing of $\boldsymbol{X}$ and $\boldsymbol{G}$. MSA represents multi-head self-attention.

where DS represents twice the downsampling using average pooling, ϕ represents that if the feature map size does not match the expected size, then several convolution and downsampling operations are performed, with each operation represented by DS(Conv(·)), and $\boldsymbol{X}_{\text{ds}}$ represents the result after step-wise downsampling.

(2) Global Aggregation. Aggregation of global information using multi-head self-attention for the $\boldsymbol{X}_{\text{ds}}$ output in the previous step:

$$\boldsymbol{X}_{\text{ga}} = \text{MSA}(\boldsymbol{Q}_{\text{ds}}, \boldsymbol{K}_{\text{ds}}, \boldsymbol{V}_{\text{ds}}) \tag{3.3}$$

where $\boldsymbol{Q}_{\text{ds}}$, $\boldsymbol{K}_{\text{ds}}$ and $\boldsymbol{V}_{\text{ds}}$ are produced by $\boldsymbol{X}_{\text{ds}}$ through linear projection, and then $\boldsymbol{X}_{\text{ga}}$ containing global information is obtained.

(3) Enrich the global information. Use position-aware global tokens $\boldsymbol{G}$ to enrich $\boldsymbol{X}_{\text{ga}}$'s global information:

$$\boldsymbol{G}_{\text{new}} = \text{Fuse}(\boldsymbol{G}, \boldsymbol{X}_{\text{ga}}) \tag{3.4}$$

where Fuse is how the two are fused, which will be explained later along with position-aware global tokens.

(4) Global Broadcast. The global information in $\boldsymbol{G}_{\text{new}}$ is broadcast to the image tokens using self-attention. This process is represented as follows:

$$\boldsymbol{X}_{\text{global}} = \text{MSA}(\boldsymbol{Q}_{\text{image}}, \boldsymbol{K}_{\text{g}}, \boldsymbol{V}_{\text{g}}) \tag{3.5}$$

where $\boldsymbol{Q}_{\text{image}}$ is produced by image tokens through linear projection, $\boldsymbol{K}_{\text{g}}$ and $\boldsymbol{V}_{\text{g}}$ are produced by $\boldsymbol{G}_{\text{new}}$ through linear projection.

Fusion. Fusing the two tokens, which contain local and global information respectively:

$$\boldsymbol{X}_{\text{new}} = \boldsymbol{X}_{\text{local}} + \boldsymbol{X}_{\text{global}} \tag{3.6}$$

3.2 Position-aware Global Tokens Global Aggregation is able to extract global information, but its scope is only in a block. For this reason, we employ the position-aware global tokens $\boldsymbol{G}$, which throughout all stages, to fuse with the $\boldsymbol{X}_{\text{ga}}$ to obtain $\boldsymbol{G}_{\text{new}}$. $\boldsymbol{G}_{\text{new}}$ has richer global information and can be used to enrich the global information and function as new position-aware global tokens to the next block after adding the identical mapping. In addition to global information, position information in position-aware global tokens is also delivered.

Global Tokens with Position Information. Figure 3(a) shows the normal global tokens [13, 5, 34] and our position-aware global tokens. The one-dimensional global tokens contain global information, and our two-dimensional position-aware global tokens additionally contain location information. The normal global tokens use the way in Figure 3(c) to fuse $\boldsymbol{X}$ and $\boldsymbol{G}$ via multi-head self-attention and broadcast the global information. Figure 3(e) is our Position-aware Global Tokens, which we set to the same number of tokens as in $\boldsymbol{X}_{\text{ga}}$, and use weighted summation to fuse them:

$$\boldsymbol{G}_{\text{new}} = \text{Fuse}(\boldsymbol{G}, \boldsymbol{X}_{\text{ga}}) = \alpha\text{MLP}(\boldsymbol{G}) + (1-\alpha)\boldsymbol{X}_{\text{ga}} \tag{3.7}$$

where $\alpha \in [0, 1]$ is a weight that is set in advance. Although

Table 1: Macro structures of two DualToken-ViT variants. B, C and H represent the number of blocks, channels and attention heads in multi-head self-attention, respectively.

Stage	Stride	DualToken-ViT-T	DualToken-ViT-S
Stage 1	1/8	B=2 C=48 H=2	B=2 C=64 H=2
Stage 2	1/16	B=6 C=96 H=4	B=6 C=128 H=4
Stage 3	1/32	B=4 C=192 H=8	B=6 C=256 H=8

our position-aware global tokens will cause the parameters to increase due to the increase in the number of tokens, it will perform better than the normal global tokens.

MLP. Before fusion, we use MLP for position-aware global tokens, which allows for a better fusion of the two. The formula of MLP is as follows:

$$G' = (\text{Linear}(\text{GELU}(\text{Linear}(G))) \quad (3.8)$$

Since the normal MLP is only in the channel dimension, we also attempt to use token-mixing MLP [27] to additionally extract the information in the spatial dimension:

$$G' = \text{Transpose}(\text{Linear}(\text{Transpose}(\text{Linear}(G)))) \quad (3.9)$$

where Transpose represents the transposition of spatial and channel axis. We refer to this process as MixMLP.

3.3 Architectures

We design two DualToken-ViT models of different scales, and their macro structures are shown in Table 1. For the task of image classification on the ImageNet-1k [7] dataset, we default the size of the image after data augment is 224×224. To prevent the complexity of the model from being too large, we set the size of position-aware global tokens to 7×7. In this way, the M of the first two stages are set to 1 and 0 respectively, and the size of X_{ga} is exactly 7×7. In the third step, the feature map size of the image tokens is exactly 7×7, this eliminates the need for local information extraction and downsampling, and allows these steps to be skipped directly. Furthermore, the convolutional kernel size of depth-wise convolution in the Conv Encoder of the first two stages is 5×5 and 7×7 respectively, and the convolutional kernel sizes in the step-wise downsampling are all 3×3. In addition, if the size of the input image changes (as in the object detection and semantic segmentation tasks) and it is not possible to make X_{ga} the same size as the position-aware global tokens, we use interpolation to change the size of X_{ga} to the same size as the position-aware global tokens. In the fusion of G and X_{ga}, we set α to 0.1.

4 Experiments

4.1 Image Classification

Setting. We perform image classification experiments on the ImageNet-1k [7] dataset and validate the top-1 accuracy on its validation set. Our model is trained with 300 epochs and is based on 224×224 resolution images. For the sake of fairness of the experiment, we try to choose models with this setup and do not use extra datasets and pre-trained models to compare with our model. We employ the AdamW [20] optimizer with betas (0.9, 0.999), weight decay 4e-2, learning rate 1e-3 and batch size 1024. And we use Cosine scheduler with 20 warmup epoch. RandAugmentation (RandAug (2, 9)), MixUp (alpha is 0.2), CutMix (alpha is 1.0), Random Erasing (probability is 0.25), and drop path (rate is 0.1) are also employed.

Table 2: Image classification performance on ImageNet-1k. "mix" indicates that our model uses MixMLP instead of normal MLP.

Model	FLOPs (G)	Params (M)	Top-1 (%)
MobileNetV2 (1.4) [26]	0.6	6.9	74.7
MobileViTv1-XXS [21]	0.4	1.3	69.0
MobileViTv2-0.5 [22]	0.5	1.4	70.2
PVTv2-B0 [31]	0.6	3.4	70.5
EdgeViT-XXS [23]	0.6	4.1	74.4
DualToken-ViT-T (mix)	**0.5**	**5.8**	**75.4**
RegNetY-800M [25]	0.8	6.3	76.3
DeiT-Ti [28]	1.3	5.7	72.2
T2T-ViT-7 [35]	1.1	4.3	71.7
SimViT-Micro [14]	0.7	3.3	71.1
MobileViTv1-XS [21]	1.0	2.3	74.8
TNT-Ti [10]	1.4	6.1	73.9
LVT [33]	0.9	5.5	74.8
EdgeViT-XS [23]	1.1	6.7	77.5
XCiT-T12 [1]	1.3	6.7	77.1
LightViT-T [13]	0.7	9.4	78.7
DualToken-ViT-S (mix)	1.0	11.4	79.4
DualToken-ViT-S	**1.1**	**11.9**	**79.5**

Results. We compare DualToken-ViT to other vision models on two scales of FLOPs, and the experimental results are shown in Table 2, where our model performs the best on both scales. For example, DualToken-ViT-S (mix) achieves 79.4% accuracy at 1.0G FLOPs, exceeding the current SoTA model LightViT-T [13]. And we improved the accuracy to 79.5% after replacing MixMLP with normal MLP.

4.2 Object Detection and Instance Segmentation

Setting. We perform experiments on the MS-COCO [17] dataset and use RetinaNet [16] and Mask R-CNN [11] architectures with FPN [15] neck for a fair comparison. Since DualToken-ViT has only three stages, we modified the FPN neck using the same method as in LightViT [13] to make our model compatible with these two detection architectures. For the RetinaNet architecture, we employ the AdamW [20] optimizer for training, where betas (0.9,

Table 3: Object detection and instance segmentation performance by Mask R-CNN on MS-COCO. All the models are pretrained on ImageNet-1K.

Backbone	**FLOPs**	**Params**	**Mask R-CNN 1x schedule**						**Mask R-CNN 3x + MS schedule**					
	(G)	**(M)**	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet-18 [12]	207	31	34.0	54.0	36.7	31.2	51.0	32.7	36.9	57.1	40.0	33.6	53.9	35.7
ResNet-50 [12]	260	44	38.0	58.6	41.4	34.4	55.1	36.7	41.0	61.7	44.9	37.1	58.4	40.1
ResNet-101 [12]	493	101	40.4	61.1	44.2	36.4	57.7	38.8	42.8	63.2	47.1	38.5	60.1	41.3
PVTv1-T [30]	208	33	36.7	59.2	39.3	35.1	56.7	37.3	39.8	62.2	43.0	37.4	59.3	39.9
PVTv1-S [30]	245	44	40.4	62.9	43.8	37.8	60.1	40.3	43.0	65.3	46.9	39.9	62.5	42.8
PVTv2-B0 [31]	196	24	38.2	60.5	40.7	36.2	57.8	38.6	-	-	-	-	-	-
LightViT-T [13]	187	28	37.8	60.7	40.4	35.9	57.8	38.0	41.5	64.4	45.1	38.4	61.2	40.8
DualToken-ViT-S (mix)	**191**	**30**	**41.1**	**63.5**	**44.7**	**38.1**	**60.5**	**40.5**	**43.7**	**65.8**	**47.4**	**39.9**	**62.7**	**42.8**

Table 4: Object detection performance by RetinaNet on MS-COCO. All the models are pretrained on ImageNet-1K.

Backbone	**FLOPs (G)**	**Params (M)**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-18 [12]	181	21.3	31.8	49.6	33.6	16.3	34.3	43.2
ResNet-50 [12]	239	37.7	36.3	55.3	38.6	19.3	40.0	48.8
PVTv1-T [30]	221	23.0	36.7	56.9	38.9	22.6	38.8	50.0
PVTv2-B0 [31]	178	13.0	37.2	57.2	39.5	23.1	40.4	49.7
ConT-M [32]	217	27.0	37.9	58.1	40.2	23.0	40.6	50.4
MF-508M [5]	168	17.9	38.0	58.3	40.3	22.9	41.2	49.7
DualToken-ViT-S	**170**	**20.0**	**40.3**	**61.2**	**42.8**	**25.5**	**43.7**	**55.2**

0.999), weight decay 1e-4, learning rate 1e-4 and batch size 16. And we use the training schedule of 1× from the MMDetection library. For the Mask R-CNN architecture, we employ the AdamW optimizer for training, where betas (0.9, 0.999), weight decay 5e-2, learning rate 1e-4 and batch size 16. We use the 1× and 3× training schedules from the MMDetection library, respectively. We use all the standard metrics for object detection and instance segmentation of the MS-COCO dataset.

Results. We compare the performance of our model with other models on Mask R-CNN and RetinaNet architectures, and the experimental results are shown in Table 3 and Table 4, respectively. Although our backbone has only three stages, DualToken-ViT-S without the maximum resolution stage still performs well in a model of the same FLOPs magnitude. In particular, in the experiments of Mask R-CNN architecture using the training schedule of 1×, our backbone achieves 41.1% AP^b and 38.1% AP^m at 191G FLOPs, which far exceeds LightViT-T [13] with similar FLOPs. This may be related to our position-aware global tokens, which we will explain in detail later.

4.3 Semantic Segmentation

Setting. We perform experiments on ADE20K [37] dataset at 512×512 resolution and use DeepLabv3 [4] and PSPNet [36] architectures for a fair comparison. For training, we employ the AdamW [20] optimizer, where betas (0.9, 0.999), weight decay 1e-4, learning rate 2e-4 and batch size 32.

Results. We compare the performance of our model with other models on DeepLabv3 and PSPNet architectures, and the experimental results are shown in Table 5. DualToken-ViT-S performs best among models of the same FLOPs magnitude on both architectures.

4.4 Ablation Study

MLPs. We compare two MLPs performed on position-aware global tokens: normal MLP and MixMLP. The experimental results on DualToken-ViT-S are shown in Table 2. The normal MLP is 0.1% more accurate than MixMLP, but it adds a little extra FLOPs and parameters. This is because MixMLP extracts information in the spatial dimension, it may damage some positional information on the position-aware global tokens.

Different methods of applying global tokens. We compare three different methods of applying global tokens. The method [13, 5, 34] in Figure 3(c) is the most common. Figure 3(e) shows our method that uses weighted summation to fuse $\boldsymbol{X}_{\text{ga}}$ and $\boldsymbol{G}$. Figure 3(d) combines the previous two methods, replacing the weighted summation based fusion in our method with the multi-head self-attention based fusion. We perform experiments on DualToken-ViT-S. In the implementation, because the complexity of the methods using multi-head self-attention based fusion is greater, we set the number of global tokens to 8, which is the same number as LightViT-T [13]. The experimental results are shown in Table 6, which show that our position-aware-based method performs the best and has 1.1M less parameters than the Normal method, with only 0.06G more FLOPs. Since the other two methods employ multi-head self-attention based fusion that requires many parameters, whereas our method employs weighted summation based fusion, our method has the smallest parameters. This demonstrates the superiority of position-aware global tokens.

The number of tokens in position-aware global tokens. We performed ablation study on the number of tokens in position-aware global tokens on ImageNet-1k [7] dataset at 224×224 resolution. In our model, the number of tokens in position-aware global tokens is set to 7×7. In order to

Table 5: Semantic segmentation performance by DeepLabv3 and PSPNet on ADE20K dataset. All the models are pretrained on ImageNet-1K.

Backbone	DeepLabv3			PSPNet		
	FLOPs (G)	Params (M)	mIoU (%)	FLOPs (G)	Params (M)	mIoU (%)
MobileNetv2 [26]	75.4	18.7	34.1	53.1	13.7	29.7
MobileViTv2-1.0 [22]	56.4	13.4	37.0	40.3	9.4	36.5
DualToken-ViT-S	**68.4**	**26.3**	**39.0**	**58.3**	**21.7**	**38.8**

Table 6: Ablation study on the method of applying global tokens. Normal, Normal* and Position-aware represent the methods in Figure 3(c), Figure 3(d) and Figure 3(e), respectively.

Global Tokens	FLOPs (G)	Params (M)	Top-1 (%)
Normal	0.99	13.0	79.2
Normal*	0.98	13.4	79.0
Position-aware	**1.05**	**11.9**	**79.5**

Table 7: Ablation study on the number of tokens in position-aware global tokens.

Number	FLOPs (G)	Params (M)	Top-1 (%)
0	0.99	10.8	79.2
3×3	0.93	11.9	79.1
4×4	0.95	11.9	79.2
5×5	0.98	11.9	79.4
6×6	1.01	11.9	79.3
7×7	**1.05**	**11.9**	**79.5**
8×8	1.10	11.9	79.3

compare the impact of different numbers of tokens on our model, we experiment with various settings for the number of tokens. If the number of tokens is set to 0, then the position-aware global tokens are not used. Because the size of $\boldsymbol{X}_{\text{ga}}$ and the position-aware global tokens will not match when the number of tokens is not 7×7, we will use interpolation for $\boldsymbol{X}_{\text{ga}}$ to make the size of the two match. The experimental results on DualToken-ViT-S are shown in Table 7. The model with the number of tokens set to 7×7 has the best performance due to the sufficient number of tokens and does not damage the information by the interpolation method. Compared to the 0 token setting, our setting is 0.3% more accurate and will only increase by 0.06G FLOPs and 1.1M parameters, which demonstrates the effectiveness of our position-aware global tokens.

Local attention. We compare the role of Conv Encoder and window self-attention [18] in our model. And we set the window size of window self-attention to 7. The experimental results on DualToken-ViT-S (mix) are shown in Table 8. The model using Conv Encoder as local attention achieves better performance, with 0.8% more accuracy than when using window self-attention, and the number of FLOPs and parameters does not increase very much. The performance of Conv Encoder is superior for two reasons. On the one hand, the convolution-based structure will be more advantageous than the transformer-based structure for light-weight models. On the other hand, window self-attention damages the position information in the position-aware global tokens. This is because the transformer-based structure does not have the inductive bias of locality. And in window self-attention, the features in the edge part of the window will be damaged due to the feature map being split into several small parts.

Table 8: Ablation study on the method of local attention.

Local Attention	FLOPs (G)	Params (M)	Top-1 (%)
Window Self-attention	0.92	10.8	78.6
Conv Encoder	**1.04**	**11.4**	**79.4**

Table 9: Ablation study on the step-wise downsampling part of the Position-aware Token Module.

Downsampling	FLOPs (G)	Params (M)	Top-1 (%)
one-step	1.01	11.3	79.2
step-wise	**1.04**	**11.4**	**79.4**

Downsampling. We perform ablation study on the step-wise downsampling part of the position-aware token module. For the setup of one-step downsampling, we directly downsample $\boldsymbol{X}_{\text{local}}$ to get the desired size, and then input it to the Global Aggregation. The experimental results on DualToken-ViT-S (mix) are shown in Table 9. Step-wise downsampling is 0.2% more accurate than one-step downsampling, and FLOPs and parameters are only 0.03G and 0.1M more, respectively. The reason for this is that the method of step-wise can retain more information by convolution during the downsampling process.

4.5 Visualization

To get a more intuitive feel for the position information contained in position-aware global tokens, we visualize the attention map of the Global Broadcast for the last block in DualToken-ViT-S (mix), and the results are shown in Figure 4. In each row, the second and third images show that the key tokens in the first image generate higher correlation with the corresponding tokens in the position-aware global tokens. And in the second image in each row,

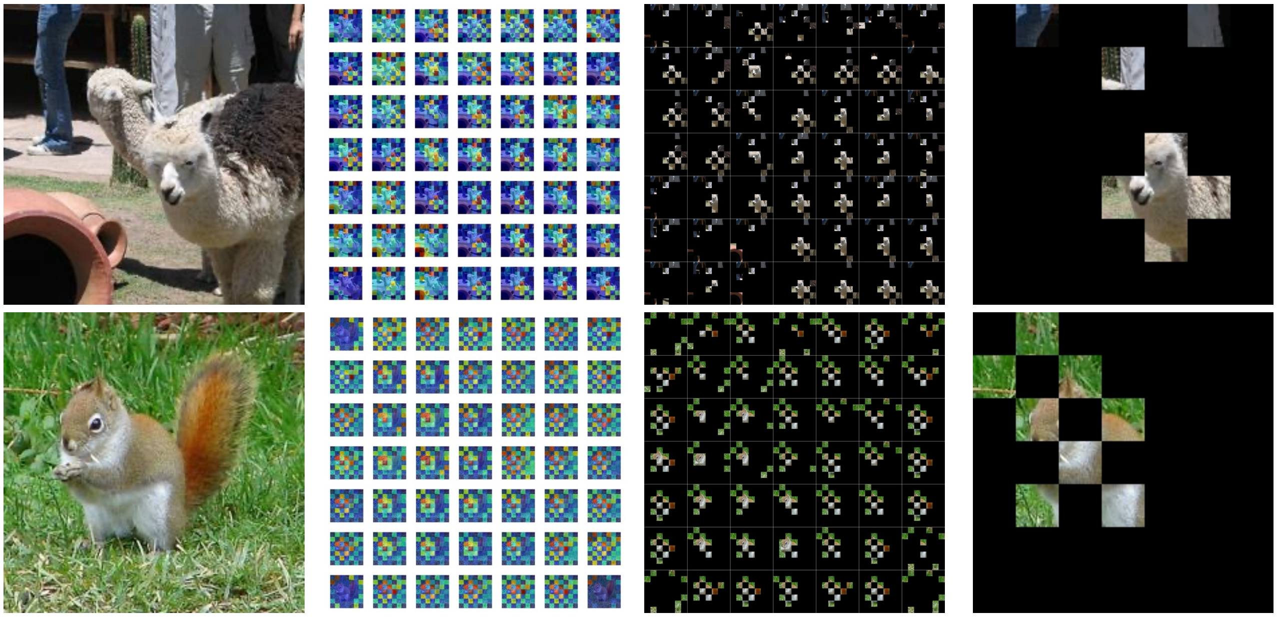

Figure 4: Visualization of the attention map of the Global Broadcast for the last block in our model. In each row, each subimage in the second image represents the correlation between this part of the first image and each token in the position-aware global tokens, and the third image shows the 8 tokens with the highest correlation in each subimage. The fourth image in each row represents the average of all subimages in the second image and shows the 8 tokens with the highest correlation.

the non-key tokens in the first image generate more uniform correlation with each part of the position-aware global tokens. The fourth image in each row shows that the overall position-aware global tokens have a higher correlation with the key tokens of the first image. These demonstrate that our position-aware global tokens contain position information.

5 Conclusion

In this paper, we propose a light-weight and efficient visual transformer model called DualToken-ViT. It achieves efficient attention structure by combining convolution-based local attention and self-attention-based global attention. We improve global tokens and propose position-aware global tokens that contain both global and position information. We demonstrate the effectiveness of our model on image classification, object detection and semantic segmentation tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant number 62202170, Alibaba Group through Alibaba Innovation Research Program, and the Open Research Fund of KLATASDS-MOE, ECNU.

References

[1] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek, et al., *Xcit: Cross-covariance image transformers*, Advances in neural information processing systems, 34 (2021), pp. 20014–20027.

[2] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, and J. Hoffman, *Hydra attention: Efficient attention with many heads*, in Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, Springer, 2023, pp. 35–49.

[3] 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, pp. 213–229.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking atrous convolution for semantic image segmentation*, arXiv preprint arXiv:1706.05587, (2017).

[5] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, *Mobile-former: Bridging mobilenet and transformer*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5270–5279.

[6] 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 (2021), pp. 9355–9366.

[7] 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, Ieee, 2009, pp. 248–255.

[8] 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*, in Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.

[9] 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).

[10] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, *Transformer in transformer*, Advances in Neural Information Processing Systems, 34 (2021), pp. 15908–15919.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask r-cnn*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[12] 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, 2016, pp. 770–778.

[13] T. Huang, L. Huang, S. You, F. Wang, C. Qian, and C. Xu, *Lightvit: Towards light-weight convolution-free vision transformers*, arXiv preprint arXiv:2207.05557, (2022).

[14] G. Li, D. Xu, X. Cheng, L. Si, and C. Zheng, *Simvit: Exploring a simple vision transformer with sliding windows*, in 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 1–6.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, *Feature pyramid networks for object detection*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[17] 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[18] 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, pp. 10012–10022.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, *A convnet for the 2020s*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

[20] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, arXiv preprint arXiv:1711.05101, (2017).

[21] S. Mehta and M. Rastegari, *Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer*, arXiv preprint arXiv:2110.02178, (2021).

[22] ———, *Separable self-attention for mobile vision transformers*, arXiv preprint arXiv:2206.02680, (2022).

[23] J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, and B. Martinez, *Edgevits: Competing light-weight cnns on mobile devices with vision transformers*, in Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI, Springer, 2022, pp. 294–311.

[24] Z. Pan, J. Cai, and B. Zhuang, *Fast vision transformers with hilo attention*, arXiv preprint arXiv:2205.13213, (2022).

[25] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, *Designing network design spaces*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10428–10436.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[27] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., *Mlp-mixer: An all-mlp architecture for vision*, Advances in neural information processing systems, 34 (2021), pp. 24261–24272.

[28] 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, PMLR, 2021, pp. 10347–10357.

[29] 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).

[30] 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, pp. 568–578.

[31] ———, *Pvt v2: Improved baselines with pyramid vision transformer*, Computational Visual Media, 8 (2022), pp. 415–424.

[32] 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).

[33] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, *Lite vision transformer with enhanced self-attention*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11998–12008.

[34] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, *Dual vision transformer*, arXiv preprint arXiv:2207.04976, (2022).

[35] 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, pp. 558–567.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, *Pyramid scene parsing network*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[37] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, *Semantic understanding of scenes through the ade20k dataset*, International Journal of Computer Vision, 127 (2019), pp. 302–321.