DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition

Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Andy J. Ma,
Yaowei Wang, *Member, IEEE*, and Wei-Shi Zheng

***Abstract*—As a *de facto* solution, the vanilla Vision Transformers (ViTs) are encouraged to model long-range dependencies between arbitrary image patches while the global attended receptive field leads to quadratic computational cost. Another branch of Vision Transformers exploits local attention inspired by CNNs, which only models the interactions between patches in small neighborhoods. Although such a solution reduces the computational cost, it naturally suffers from small attended receptive fields, which may limit the performance. In this work, we explore effective Vision Transformers to pursue a preferable trade-off between the computational complexity and size of the attended receptive field. By analyzing the patch interaction of global attention in ViTs, we observe two key properties in the shallow layers, namely locality and sparsity, indicating the redundancy of global dependency modeling in shallow layers of ViTs. Accordingly, we propose Multi-Scale Dilated Attention (MSDA) to model *local* and *sparse* patch interaction within the sliding window. With a pyramid architecture, we construct a Multi-Scale Dilated Transformer (DilateFormer) by stacking MSDA blocks at low-level stages and global multi-head self-attention blocks at high-level stages. Our experiment results show that our DilateFormer achieves state-of-the-art performance on various vision tasks. On ImageNet-1 K classification task, DilateFormer achieves comparable performance with 70% fewer FLOPs compared with existing state-of-the-art models. Our DilateFormer-Base achieves 85.6% top-1 accuracy on ImageNet-1 K classification task, 53.5% box mAP/46.1% mask mAP on COCO object detection/instance segmentation task and 51.1% MS mIoU on ADE20 K semantic segmentation task. The code is available at https://isee-ai.cn/˜jiaojiayu/DilteFormer.html.**

***Index Terms*—Vision transformer.**

I. Introduction

In THE past years, Convolution Neural Networks (CNNs) have dominated a wide variety of vision tasks such as classification [1], [2], [3], [4], [5], [6], [7], object detection [8], [9], [10], [11], [12] and semantic segmentation [13], [14], [15], attributing to the inductive bias of convolution operations, i.e., local connections and weight sharing. However, convolution only models local dependencies of pixels, which ignores the dependency modeling between distant pixels to some extent [16]. Inspired by sequence modeling tasks [17], [18] in natural language processing (NLP) [18], [19], [20], pioneer works [21], [22], [23], [24], [25] introduce Transformers with long-range dependency modeling ability into computer vision, achieving exciting results in various vision tasks.

With global attention, the vanilla Vision Transformers (ViTs) [21], [22] can conduct dependency modeling between arbitrary image patches. However, the **global** attended receptive field of ViTs leads to quadratic computational cost, and modeling dependencies among all patches may be redundant for mainstream vision tasks. To reduce the computational cost and redundancy of global attention, some works [26], [27], [28], [29], [30] introduce inductive bias explored in CNNs, performing **local** attention only in small neighborhoods. However, local attention naturally suffers from small attended receptive fields, which results in a lack of capability to model long-range dependencies.

In this work, we explore an effective Vision Transformer to pursue a preferable trade-off between the computational complexity and the size of the attended receptive field. By analyzing the patch interaction of global attention in ViTs [21], [22], we find that the attention matrix in shallow layers has two key properties, namely *locality* and *sparsity*. As shown in Fig. 2, in the third attention block of ViT-Small, relevant patches are sparsely distributed in the neighborhood of the query patch. Such a locality and sparsity property indicates that distant patches in shallow layers are mostly irrelevant in semantics modeling for mainstream vision tasks, and thus there is much redundancy to be reduced in the costly global attention module.

Based on the above analysis, we propose a Sliding Window Dilated Attention (SWDA) operation, which performs self-attention among patches sparsely selected in the surrounding field. To make further use of the information within the attended receptive field, we propose Multi-Scale Dilated Attention (MSDA), which simultaneously captures semantic dependencies at different scales. MSDA sets different dilation rates for different heads, enabling the ability of multi-scale representation

Manuscript received 22 September 2022; revised 16 January 2023; accepted 28 January 2023. Date of publication 9 February 2023; date of current version 12 December 2023. This work was supported in part by NSFC under Grants U21A20471, U1911401, and U1811461, in part by Guangdong NSF Project under Grants 2020B1515120085 and 2018B030312002, in part by Peng Cheng Laboratory Key Research Project under Grant PCL2021A07, in part by Guangzhou Research Project under Grant 201902010037, and in part by the Key-Area Research and Development Program of Guangzhou under Grant 202007030004. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Jianlong Fu. *(Jiayu Jiao and Yu-Ming Tang contributed equally to this work.) (Corresponding author: Wei-Shi Zheng.)*

Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, and Andy J. Ma are with the School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: jiaojy6@mail2.sysu.edu.cn; tangym9@mail2.sysu.edu.cn; linky5@mail2.sysu.edu.cn; gaoyp23@mail2.sysu.edu.cn; majh8@mail.sysu.edu.cn).

Wei-Shi Zheng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, also with the Peng Cheng Laboratory, Shenzhen 518066, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, Guangzhou 510275, China (e-mail: wszheng@ieee.org).

Yaowei Wang is with the Pengcheng Laboratory, ShenZhen 518066, China (e-mail: wangyw@pcl.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMM.2023.3243616, provided by the authors.

Digital Object Identifier 10.1109/TMM.2023.3243616

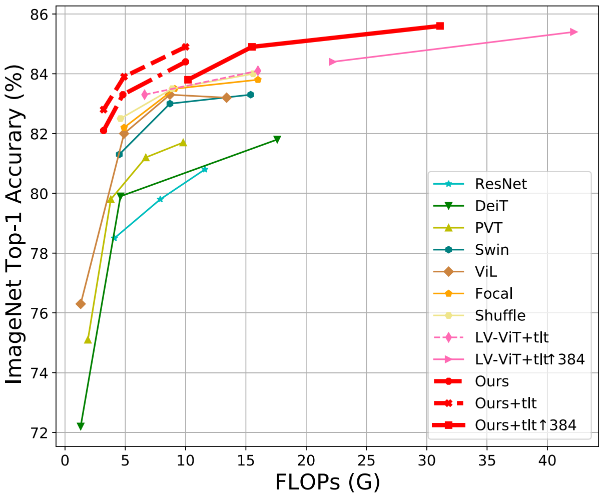


Fig. 1. Performance comparisons with respect to FLOPs on ImageNet-1 K classification. Without extra training data, our DilateFormer variants achieve comparable or even better performance with fewer FLOPs.

learning. Following PVT [31] and Swin [26], we adopt a pyramid architecture to develop a new effective Transformer model, namely Multi-Scale Dilated Transformer (DilateFormer), which stacks MSDA in shallow stages to capture low-level information and global Multi-Head Self-Attention [21], [22] in deeper stages to model high-level interaction.

For model evaluation, we design variants of DilateFormer with different capacities and apply them to different vision tasks. Experimental results show that our proposed DilateFormer outperforms state-of-the-art Vision Transformers [21], [22], [26], [27], [28], [30] on various datasets across different model sizes. As depicted in Fig. 1, we demonstrate the performance of our DilateFormers on ImageNet-1 K classification task. Without extra training data, our Dilate-S (4.8 GFLOPs) achieves comparable performance with Swin-B (15.4 GFLOPs) [26] on ImageNet-1 K using only 1/3 FLOPs. With the assistance of Token Labeling [32], our DilateFormers achieve better performance than LV-ViTs [32] at different model sizes. Specifically, our Dilate-S* (4.9 GFLOPs) and our Dilate-B* (10.0 GFLOPs) achieve 83.9% and 84.9% respectively, surpassing LV-ViT-S [32] (6.6 GFLOPs) and LV-ViT-M [32] (16 GFLOPs). Besides, our Dilate-B achieves 85.6% top-1 accuracy on ImageNet-1 K classification [33] task, 53.5% box mAP/46.1% mask mAP on COCO [34] object detection/instance segmentation task and 51.1% MS mIoU on ADE20K [35] semantic segmentation task.

II. Related Work

A comparison of technical details with various models is shown in Table I. We summarize and classify our DilateFormer and related vision transformer models from the perspectives of overlapping tokenizer/downsampler, positional embedding, attention type and multi-scale. In the following section, we detail some related works.

A. Global Attention in Vision Transformers

Inspired by the success in NLP [17], [19], [45], the vanilla Vision Transformers (ViTs) [21], [22] directly apply self-attention mechanisms to patches split from images. By utilizing sufficient training data [21], [46] and strong data augmentation strategies [22], [47], [48], [49], [50], [51], Transformer-based methods [25], [52], [53], [54], [55], [56], [57], [58] achieve exciting performance improvements on various vision tasks, i.e., image classification [22], [26], [41], [44], [59], [60], object detection [25], [28], [30], [60], [61], [62], [63], [64], semantic segmentation [29], [42], [58], [65], [66], [67], [68], and re-identification [69], [70], [71]. Since the computational complexity of the self-attention mechanism is quadratic *w.r.t.* the number of patches, global attention is difficult to apply in high-resolution image encoding. Furthermore, according to our analysis in Section I, the long-range modeling capability of the global attention mechanism in shallow layers of ViTs is redundant. To reduce the redundancy and computational cost of the self-attention mechanism, some works [29], [31], [36] introduce sub-sampling operations in self-attention blocks while preserving the global receptive field. Such sub-sampling operations require complex designs and introduce extra parameters or computational cost. Different from these works, our Sliding Window Dilated Attention (SWDA) is easy to implement for reducing the redundancy of self-attention mechanism in a dilated manner.

B. Local Attention in Vision Transformers

In order to make the self-attention mechanism applicable for high-resolution image encoding, some works [26], [60], [72] apply the self-attention mechanism to patches in a fixed local region to reduce computational cost. For example, Swin [26] applies self-attention to the patches within fixed windows and then adopts a window-shifting strategy in the next layer for information exchange between the patches in different windows. CSwin [60] improves the window-fixed setting in Swin [26], performing self-attention to cross-shaped windows. Other works [37], [38], [39] use grouped sampling or spatial shuffling operation for information exchange between different local windows. Inspired by the convolution operation in CNNs [1], [3], [4], [5], [73], ViL [28] and NAT [30] propose sliding window attention, which models dependencies only with neighboring patches in the window centering each query patch. Moreover, some works [40], [41], [74], [75], [76], [77], [78] combine CNNs and Transformers for introducing the locality prior, and they usually design hand-crafted and complex modules for interaction between CNNs and Transformers features, leading to a lack of scalability to large-scale parameters [79], [80]. However, some works [26], [28], [30], [60] above only consider the locality of the self-attention mechanism and lack consideration of the sparsity. Although some works [36], [37], [38], [39] above perform self-attention in a sparse and uniform manner, they are designed to approximate the global attended receptive field. In comparison, our Sliding Window Dilated Attention (SWDA) takes both the locality and sparsity of self-attention mechanism into consideration. Our SWDA introduces a prior to reduce the redundancy of self-attention mechanism, which

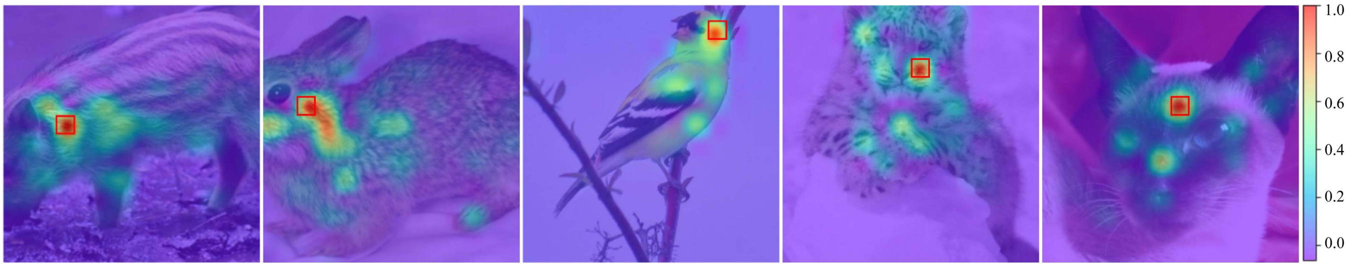


Fig. 2. Visualization of attention maps of the third Multi-Head Self-Attention block of ViT-Small.[1] We visualize the activations in attention maps of the query patches (in the red box). The attention maps show that patches with high attention scores sparsely scatter around the query patch, and other patches have low attention scores.

TABLE I
COMPARISON OF TECHNICAL DETAILS WITH OTHER MODELS

Model	Overlapping Tokenizer	Overlapping Downsampler	Positional Embedding	Attention			Multi-scale	
				Local	Global	Sparse	Stage-level	Block-level
ViT [22]/DeiT [72]	×	-	APE	×	✓	×	×	×
PVT [76]	×	×	APE	×	✓	×	✓	×
Swin [54]	×	×	RPE	✓	×	×	✓	×
Twins [13]	×	×	CPE	✓	✓	×	✓	×
GG [94]	×	×	RPE/APE	×	×	✓	✓	×
Shuffle [38]	✓	×	RPE	✓	×	✓	✓	×
MaxViT [74]	✓	✓	CPE	✓	×	✓	✓	×
CrossFormer [77]	✓	✓	RPE	✓	×	✓	✓	✓
ViL [104]	×	×	RPE/APE	✓	✓	×	✓	×
NAT [32]	✓	✓	RPE	✓	×	×	✓	×
Mobile-Former [10]	-	-	CPE	×	✓	×	✓	×
Conformer [60]	✓	-	-	×	✓	×	✓	×
Shunted [65]	✓	✓	CPE	×	✓	×	✓	✓
MPViT [65]	✓	✓	CPE	×	✓	×	✓	✓
ViTAE [88]	✓	-	APE	×	✓	×	✓	✓
UniFormer [46]	✓	✓	CPE	×	✓	×	✓	×
Focal [90]	×	×	RPE	✓	✓	×	✓	✓
DilateFormer (ours)	✓	✓	CPE	✓	✓	✓	✓	✓

"-" Indicates these modules do not exist. For overlapping tokenizer/downsampler, "✓" and "×" indicate whether these modules are overlapping or not. For positional embedding, "APE," "RPE" and "CPE" indicate absolute positional embedding, relative positional embedding and convolutional positional embedding, respectively. For other technical details, "✓" and "×" indicate these modules are used or not.

performs self-attention in a dilated window centered on query patch.

C. Multi-Scale Vision Transformer

The vanilla Vision Transformer [21], [22] is a "columnar" structure for visual tasks. Since multi-scale information [1], [3], [4], [81], [82], [83], [84], [85] is beneficial for dense prediction tasks such as object detection, instance and semantic segmentation, recent works [25], [26], [28], [30], [31], [37], [39], [42], [60], [65], [86], [87] introduce multi-scale modeling capability by using a pyramid structure to design their transformer backbones. Several works [27], [39], [40], [41], [42], [43], [65], [88] introduce multi-scale information in patch embedding layers [39] or self-attention blocks [27], [42], [65] or add extra branches [40], [41], [43] to perform convolution operation. CrossFormer [39] utilizes different convolution operations or different patch sizes for designing patch embedding. Shunted Transformer [42] uses multi-scale token aggregation for obtaining keys and values of various sizes. MPViT [65] consists of multi-scale patch embedding and multi-path transformer blocks. Conformer [41], Mobile-Former [40] and ViTAE [43] design additional convolution branches outside or inside the self-attention blocks to integrate multi-scale information. The above methods all require complex design, which inevitably introduce additional parameters and computational cost. Our Multi-Scale Dilated Attention (MSDA) extracts multi-scale features by setting different dilation rates, which is simple and does not need to introduce extra parameters and computational cost.

D. Dilated Convolution

Traditional Convolution-based networks [1], [3], [4], [5] usually use downsampling or convolution with a large stride

[1] We use the official checkpoint from https://github.com/google-research/vision_transformer

to increase the receptive field and reduce computational cost. However, these approaches [1], [3], [4], [5] result in reduced resolution of feature maps, affecting model performance in many tasks such as object detection [8], [9], [10], [11], [12] and semantic segmentation [13], [14], [15]. Therefore, Cohen et al. [89], [90] propose dilated convolution [91], [92], which increases the receptive field without reducing the resolution and extracts the information of the feature map at different scales by setting different dilation rates. Dilated convolution with dynamic weights [93], namely Dynamic Dilated Convolution (DDC), uses the entire feature map to generate the kernel parameter of convolution, which is data-specific at the feature-map level.

Different from existing works, we propose a simple yet effective Dilated Attention operation by introducing various dilation rates at the same semantic level into a single self-attention operation, which more flexibly models multi-scale interaction. Although ours is a sliding window based dilated attention, ours differs from DDC because our modelling performs self-attention on keys and values sparsely selected in a sliding window centered on the query patch, which is data-specific at the token level. In addition, we also notice a concurrent work, DiNAT [94], which uses a single-scale and fixed dilation rate in each block of the same stage, lacking multi-scale interaction. In contrast, our DilateFormer uses a multi-scale strategy in each block i.e., setting different dilation rates for different heads, which can capture and fuse multi-scale semantic feature.

III. Multi-Scale Dilated Transformer

In this section, we introduce our proposed Multi-Scale Dilated Transformer (DilateFormer) in details. In Section III-A, we introduce our Sliding Window Dilated Attention (SWDA) operation, towards effective long-range dependency modeling in feature maps. In Section III-B, we design Multi-Scale Dilated Attention (MSDA), which simultaneously captures contextual semantic dependencies at different scales to make good use of the information inside the block. The overall framework and variants of the proposed Multi-Scale Dilated Transformer (DilateFormer) are illustrated in Section III-C.

A. Sliding Window Dilated Attention

According to the locality and sparsity properties observed in the global attention of shallow layers in vanilla Vision Transformers (ViTs), we propose a Sliding Window Dilated Attention (SWDA) operation, where the keys and values are *sparsely* selected in a sliding window centered on the query patch. Self-attention is then performed on these representative patches. Formally, our SWDA is described as follows:

$$X = \text{SWDA}(Q, K, V, r), \tag{1}$$

where Q, K and V represent the query, key and value matrix, respectively. Each row of the three matrices indicates a single query/key/value feature vector. For the query at location $(i,\, j)$ in the original feature map, SWDA sparsely selects keys and values to conduct self-attention in a sliding window of size $w \times w$ centered on $(i,\, j)$. Furthermore, we define a dilation rate $r \in \mathbb{N}^+$ to control the degree of sparsity. Particularly, for the position $(i,\, j)$, the corresponding component x_{ij} of the output X from SWDA operation is defined as follows:

$$\begin{aligned} x_{ij} &= \text{Attention}(q_{ij}, K_r, V_r), \\ &= \text{Softmax}\left(\frac{q_{ij}K_r^T}{\sqrt{d_k}}\right)V_r, \quad 1 \le i \le W, 1 \le j \le H, \end{aligned} \tag{2}$$

where H and W are the height and width of the feature map. K_r and V_r represent keys and values selected from the feature maps K and V. Given the query positioned at $(i,\, j)$, keys and values positioned at the following set of coordinate $(i,'\, j')$ will be selected for conducting self-attention:

$$\begin{aligned} &\left\{(i,'\, j') \,\middle|\, i' = i + p \times r, j' = j + q \times r\right\}, \\ &\qquad -\frac{w}{2} \le p,\, q \le \frac{w}{2}. \end{aligned} \tag{3}$$

Our SWDA conducts the self-attention operation for all query patches in a sliding window manner. For the query at the edge of the feature map, we simply use the zero padding strategy commonly used in convolution operations to maintain the size of the feature map. By sparsely selecting keys and values centered on queries, the proposed SWDA explicitly satisfies the locality and sparsity property and can model the long-range dependency effectively.

B. Multi-Scale Dilated Attention

To exploit the sparsity at different scales of the self-attention mechanism in block-level, we further propose a Multi-Scale Dilated Attention (MSDA) block to extract multi-scale semantic information. As shown in Fig. 4, given a feature map X, we obtain corresponding queries, keys and values by linear projection. After that, we divide the channels of the feature map to n different heads and perform multi-scale SWDA in different heads with different dilation rates. Specifically, our MSDA is formulated as follows:

$$h_i = \text{SWDA}(Q_i, K_i, V_i, r_i), \quad 1 \le i \le n, \tag{4}$$

$$X = \text{Linear}\left(\text{Concat}\left[h_1, \ldots, h_n\right]\right), \tag{5}$$

where r_i is the dilation rate of the i-th head and Q_i, K_i and V_i represent slices of feature maps fed into the i-th head. The outputs $\{h_i\}_{i=1}^n$ are concatenated together and then sent to a linear layer for feature aggregation.

By setting different dilation rates for different heads, our MSDA effectively aggregates semantic information at various scales within the attended receptive field and efficiently reduces the redundancy of self-attention mechanism without complex operations and extra computational cost.

C. Overall Architecture

With a pyramid structure, we propose the Multi-Scale Dilated Transformer (DilateFormer) as shown in Fig. 3. According to the locality and sparsity property of shallow layers in ViTs, the first two stages of DilateFormer use Multi-Scale Dilated Attention (MSDA) proposed in Section III-B while the latter two stages utilize ordinary Multi-Head Self-Attention (MHSA). What's more, we use the overlapping tokenizer [74] for patch

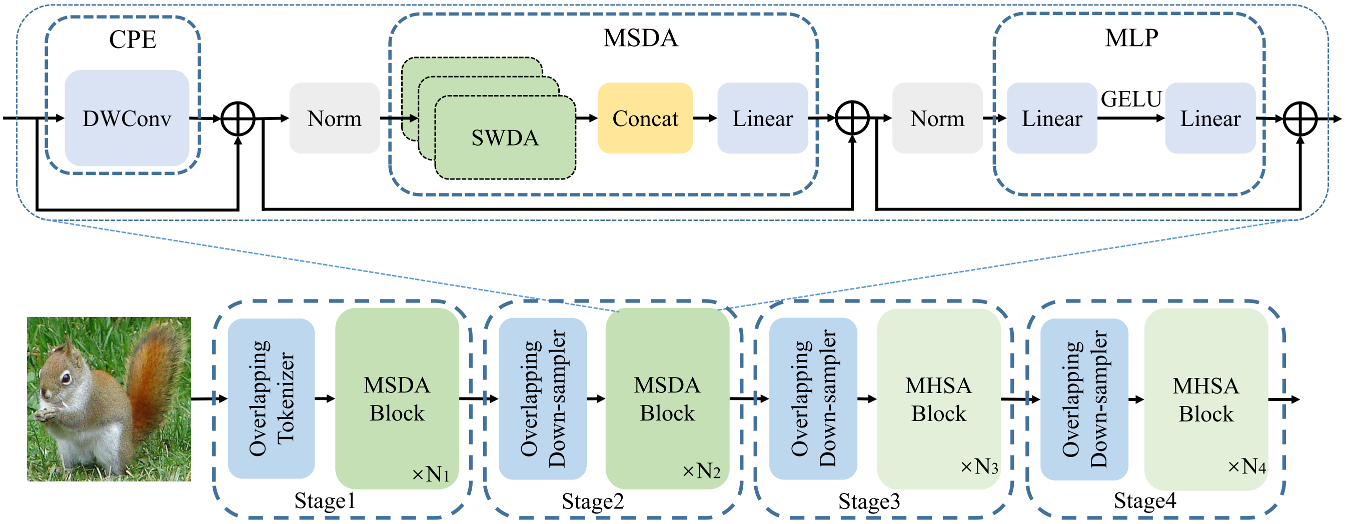


Fig. 3. The overall architecture of our DilateFormer. The top part shows the proposed Multi-Scale Dilated Attention (MSDA) block, consisting of DwConv, Multi-Scale Sliding Window Dilated Attention operation (SWDA) and MLP. The bottom part shows DilateFormer, consisting of Overlapping Tokenizer, Overlapping Downsampler, Multi-Scale Dilated Attention (MSDA) block and Multi-Head Self-Attention (MHSA) block.

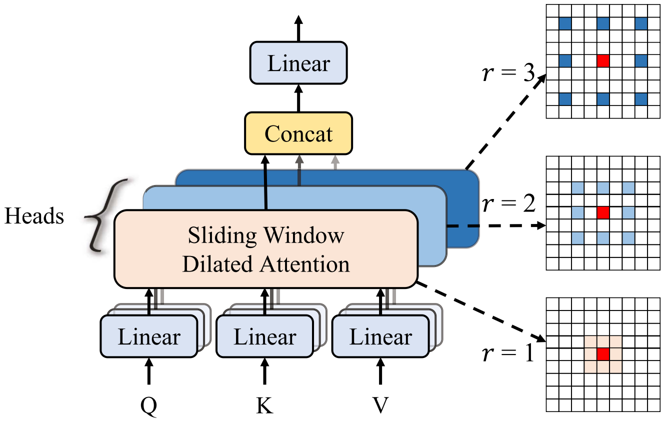


Fig. 4. **Illustration of Multi-Scale Dilated Attention (MSDA).** First, the channels of the feature map are split into different heads. Then, the self-attention operation is performed among the colored patches in the window surrounding the red query patch, using different dilation rates in different heads. Besides, features in different heads are concatenated together and then fed into a linear layer. By default, we use a 3×3 kernel size with dilation rates $r = 1$, 2 and 3, and the sizes of attended receptive fields in different heads are 3×3, 5×5 and 7×7.

embedding, which uses multiple overlapping 3×3 convolution modules with zero-padding. The resolution of the output feature map can be adjusted by controlling the stride size of convolution kernels to be 1 or 2 alternately. To merge patches in the previous stage, we utilize the overlapping downsampler [30], a convolution module with an overlapping kernel size of 3 and a stride of 2. To make the position encoding adaptive to inputs of different resolutions, we use Conditional Position Embedding (CPE) proposed in CPVT [24] whenever inputs are fed into MSDA or MHSA blocks. Specifically, our overall architecture is described as follows:

$$X = \mathrm{CPE}(\hat{X}) + \hat{X} = \mathrm{DwConv}(\hat{X}) + \hat{X}, \tag{6}$$

$$Y = \begin{cases} \mathrm{MSDA}(\mathrm{Norm}(X)) + X, & \text{at low-level stages,} \\ \mathrm{MHSA}(\mathrm{Norm}(X)) + X, & \text{at high-level stages,} \end{cases} \tag{7}$$

$$Z = \mathrm{MLP}(\mathrm{Norm}(Y)) + Y, \tag{8}$$

where $\hat{X}$ is the input of the current block, i.e., the image patches or the output from the last block. In practice, we implement CPE as a depth-wise convolution (DwConv) module with zero-padding and 3×3 kernel size. We add MLP following prior works [22], [26], which consists of two linear layers with the channel expansion ratio of 4 and one GELU activation.

Based on the above network structure, we introduce three variants of the proposed DilateFormer (i.e., Tiny, Small, and Base), and the specific model settings are given in Table II.

IV. Experiments

To evaluate the performance of our Multi-Scale Dilated Transformer (DilateFormer), we take our model as a vision backbone for ImageNet-1K [33] classification, COCO [34] object detection and instance segmentation, and ADE20K [35] semantic segmentation. Furthermore, we evaluate the effectiveness of our key modules via ablation studies. All experiments are conducted on a single server node with 8 A100 GPUs.

A. Image Classification on ImageNet-1 K

1) Dataset and Implementation Details: ImageNet-1k [33] is a large-scale 1000-classes dataset that contains 1.28 million training images and 50,000 validation images. We conduct classification experiments on ImageNet-1 K dataset to evaluate our variants, following the same training strategies of baseline Transformers as DeiT [22] and PVT [31] for a fair comparison. We use the AdamW optimizer [105] with 300 epochs including the first 10 warm-up epochs and the last 10 cool-down epochs and adopt a cosine decay learning rate scheduler decayed by a factor

TABLE II
MODEL VARIANTS OF OUR DILATEFORMER

Resolution	Block	Tiny	Small	Base
Stage 1 (56×56)	MSDA	$\begin{bmatrix}\text{72-d, 3-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 2$	$\begin{bmatrix}\text{72-d, 3-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 3$	$\begin{bmatrix}\text{96-d, 3-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 4$
Stage 2 (28×28)	MSDA	$\begin{bmatrix}\text{144-d, 6-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 2$	$\begin{bmatrix}\text{144-d, 6-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 5$	$\begin{bmatrix}\text{192-d, 6-h}\\ \text{ks. } 3\times 3\\ \text{dr. } [1,2,3]\end{bmatrix} \times 8$
Stage 3 (14×14)	MHSA	$\begin{bmatrix}\text{288-d, 12-h}\end{bmatrix} \times 6$	$\begin{bmatrix}\text{288-d, 12-h}\end{bmatrix} \times 8$	$\begin{bmatrix}\text{384-d, 12-h}\end{bmatrix} \times 10$
Stage 4 (7×7)	MHSA	$\begin{bmatrix}\text{576-d, 24-h}\end{bmatrix} \times 2$	$\begin{bmatrix}\text{576-d, 24-h}\end{bmatrix} \times 3$	$\begin{bmatrix}\text{768-d, 24-h}\end{bmatrix} \times 3$

MSDA and MHSA represent multi-scale dilated attention and multi-head self-attention, respectively. “D,” “H,” “KS.” and “DR.”indicate feature dimension, the number of head, kernel size and dilation rate, respectively.

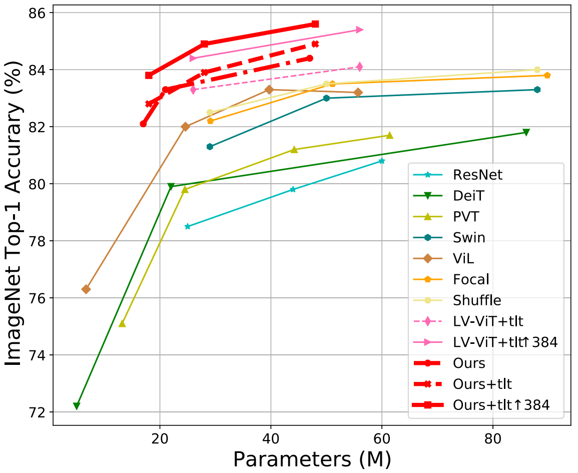


Fig. 5. Performance comparisons with respect to model parameters on ImageNet-1 K classification. Without extra training data, our DilateFormer variants achieve comparable or even better performance with fewer model parameters.

of 10 every 30 epochs with a base learning rate of 0.001, a batch size of 1024, and a weight decay of 0.05. To further demonstrate the performance of DilateFormer, Token Labeling [32] is used to auxiliarily train DilateFormer. We add an extra fully connected layer and an auxiliary loss to DilateFormer and follow the training strategy of LV-ViT [32] where CutMix [47] and Mixup [48] are replaced by MixToken [32]. For fine-tuning our models on a larger resolution, i.e., 384×384, the special hyperparameters are set as follows: weight decay, learning rate, batch size, warm-up epoch and total epoch are set to 1e-8, 5e-6, 512, 5 and 30.

2) Results and Analysis: As shown in Table III, Figs. 1 and 5, our proposed DilateFormer outperforms previous state-of-the-art models at different model sizes. Specifically, Dilate-S achieves 83.3% top-1 accuracy on ImageNet-1 K with a resolution of 224, surpasses Swin-T [26], ViL-S [28] by 2.0% and 1.3% respectively and has fewer parameters and FLOPs than these models. With the assistance of Token Labeling [32] (denoted by ‘$\star$’), our models achieve better performance than LV-ViTs [32] at different model sizes, i.e., Dilate-S* (4.9 GFLOPs) and Dilate-B* (10.0 GFLOPs) achieve 83.9% and 84.9% respectively, surpassing LV-ViT-S [32] (6.6 GFLOPs) and LV-ViT-M [32] (16 GFLOPs). The results in Table III also show the efficiency and effectiveness of the proposed model. Without extra assistance or high-resolution finetuning, Dilate-T consumes only 3.2 GFLOPs and achieves 82.1% accuracy, which is comparable to the performance of ViL-S [28] (4.9 G, 82.0%), Focal-T [27] (4.9 G, 82.2%) and PVT-L [31] (9.8 G, 81.7%). Similar conclusions can be found in larger models: our Dilate-S (83.3%) with 4.8 GFLOPs outperforms ViL-B [28] (13.4 G, 83.2%), Swin-B [26] (15.4 G, 83.4%), and DeiT-B [22] (17.5 G, 81.8%), indicating that our MSDA can effectively capture long-range dependencies as previous methods but save up to 70% FLOPs. To demonstrate the strong learning capability of DilateFormer, our Dilate-B fine-tuned on 384 × 384 images obtains 85.6% top-1 accuracy and outperforms LV-ViT-M [32] (85.4%) which needs 1.37 times more FLOPs.

B. Object Detection and Instance Segmentation on COCO

1) Dataset and Implementation Details: We evaluate our variants on object detection and instance segmentation on COCO2017 dataset [34]. COCO2017 dataset contains 118 K images for training, 5 K images for validation and 20 K images for testing. We utilize two representative frameworks: Mask R-CNN [12] and Cascade Mask R-CNN [108] implemented in mmdetection [109] and adopt the ImageNet-1 K pre-trained variants as backbones. For Mask R-CNN and Cascade Mask R-CNN frameworks, we use the AdamW optimizer with a base learning rate of 0.0001, a weight decay of 0.05, and a batch size of 16. For a fair comparison, we train our variants Dilate-S and Dilate-B via two strategies: (1) 1 × schedule with 12 epochs where the shorter side of the image is resized to 800 and the longer side is less than 1333; (2) 3 × schedule with 36

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART ON IMAGENET-1 K

Method	Params (M)	FLOPs (G)	Train	Test	Top1 (%)
RegNetY-4G [62]	21	4.0	224	224	80.0
ResNet-50 [35]	25	4.1	224	224	78.5
ConvNeXt-T [55]	28	4.5	224	224	82.1
Mobile-Former-508M [10]	14	1.0	224	224	79.3
PVT-S [76]	25	3.8	224	224	79.8
DW-Conv.-T [30]	24	3.8	224	224	81.3
CoAtNet-0 [18]	25	4.2	224	224	81.6
Swin-T [54]	29	4.5	224	224	81.3
CvT-13 [79]	20	4.5	224	224	81.6
GG-T [94]	28	4.5	224	224	82.0
DeiT-S [72]	22	4.6	224	224	79.9
Distilled DeiT-S [72]	22	4.6	224	224	81.2
ViL-S [104]	25	4.9	224	224	82.0
TNT-S [29]	24	5.2	224	224	81.3
NesT-T [106]	17	5.8	224	224	81.3
BoTNet-S1-59 [68]	34	7.3	224	224	81.7
Dilate-T (ours)	17	3.2	224	224	82.1
Dilate-T* (ours)	18	3.2	224	224	**82.8**
CvT-13 $\uparrow$ 384 [79]	20	16.3	224	384	83.0
Dilate-T* $\uparrow$ 384 (ours)	18	10.2	224	384	**83.8**
ResNet-101 [35]	44	7.9	224	224	79.8
ConvNeXt-S [55]	50	8.7	224	224	83.1
RegNetY-16G [62]	84	16.0	224	224	82.9
Focal-T [90]	29	4.9	224	224	82.2
CrossFormer-S 77	31	4.9	224	224	82.5
T2T-14 [97]	22	5.2	224	224	80.7
DiNAT-T [31]	28	4.3	224	224	82.7
LV-ViT-S* [39]	26	6.6	224	224	83.3
CvT-21 [79]	32	7.1	224	224	82.5
Twins-SVT-B [13]	56	8.3	224	224	83.1
Swin-S [54]	50	8.7	224	224	83.0
PoolFormer-M36 [95]	56	8.8	224	224	82.1
PVT-L [76]	61	9.8	224	224	81.7
NesT-S [106]	38	10.4	224	224	83.3
DeepVit-L	55	12.5	224	224	82.2
CoaT-S	22	12.6	224	224	82.1
TNT-B [29]	66	14.1	224	224	82.8
Dilate-S (ours)	21	4.8	224	224	83.3
Dilate-S* (ours)	22	4.9	224	224	**83.9**
CoAtNet-0$\uparrow$ 384 [18]	20	13.4	224	384	83.9
T2T-14 $\uparrow$ 384 [97]	22	17.1	224	384	83.3
LV-ViT-S* $\uparrow$ 384 [39]	26	22.2	224	384	84.4
CvT-21 $\uparrow$ 384 [79]	32	24.9	224	384	83.3
Dilate-S* $\uparrow$ 384 (ours)	22	15.5	224	384	**84.9**
ResNet-152 [35]	60	11.6	224	224	80.8
EffNet-B7 [71]	54	39.2	600	600	84.3
Next-ViT-L [45]	58	10.8	224	224	83.6
PoolFormer-M48 [95]	73	11.6	224	224	82.5
DeepViT-L [111]	55	12.5	224	224	83.1
DW-Conv.-B [30]	74	12.9	224	224	83.2
DiNAT-S [31]	51	7.8	224	224	83.8
T2T-24 [97]	64	13.2	224	224	82.2
ViL-B [104]	56	13.4	224	224	83.2
Twins-SVT-L [13]	99	14.8	224	224	83.3
Swin-B [54]	88	15.4	224	224	83.4
Shuffle-B [38]	88	15.6	224	224	84.0
CoAtNet-2 [18]	75	15.7	224	224	84.1
Focal-B [90]	90	16.0	224	224	83.8
LV-ViT-M* [39]	56	16.0	224	224	84.1
CrossFormer-L [77]	92	16.1	224	224	84.0
MPViT-B [42]	75	16.4	224	224	84.3
DeiT-B [72]	86	17.5	224	224	83.4
Distilled DeiT-B [72]	86	17.5	224	224	81.8
NesT-B [106]	68	17.9	224	224	83.8
BoTNet-T7 [68]	79	19.3	256	256	84.2
Dilate-B (ours)	47	10.0	224	224	84.4
Dilate-B* (ours)	48	10.0	224	224	**84.9**
CoAtNet-1 $\uparrow$ 384 [18]	42	27.4	224	384	85.1
LV-ViT-M* $\uparrow$ 384 [39]	56	42.2	224	384	85.4
BoTNet-S1-128$\uparrow$ 384 [68]	79	45.8	256	384	84.7
Dilate-B* $\uparrow$ 384 (ours)	48	31.1	224	384	**85.6**

'$\star$' Indicates token labeling proposed in LV-ViT [32], and '$\uparrow$' indicates that the model is fine-tuned at a larger resolutions.

epochs where the multi-scale training strategy is adopted and the shorter side of the image is resized in $[480, 800]$. Because image resolution in object detection and instance segmentation is generally larger than that in image classification, we use a combination of local window attention, local window attention with shifted operation [26] and global attention in stage3 of DilateFormer to reduce computational cost.

2) Results and Analysis: Table IV and Table V report box mAP (AP$^{\text{b}}$) and mask mAP (AP$^{\text{m}}$) of Mask R-CNN framework and Cascade Mask R-CNN framework, respectively. Our DilateFormer variants outperform recent Transformers on both object detection and instance segmentation in two frameworks. For Mask R-CNN 1× schedule, DilateFormer surpasses Swin Transformer [26] by 2.8-3.6% of box mAP and 2.5-2.6% mask mAP at comparable settings, respectively. For 3× + MS schedule, Dilate-B achieves 49.9% box mAP and 43.7% mask mAP in Mask R-CNN framework, 53.3% box mAP and 46.1% mask mAP in Cascade Mask R-CNN framework. Furthermore, our Dilate-S outperforms PVT-M [31] by 2.2% box mAP, 2.7% mask mAP at 1× schedule with 13.2% fewer FLOPs.

C. Semantic Segmentation on ADE20 K

1) Dataset and Implementation Details: ADE20 K dataset [35] contains 150 semantic categories, and there are 20,000 images for training, 2000 images for validation and 3000 images for testing. We evaluate the proposed variants for DilateFormer on semantic segmentation on ADE20 K and utilize two representative frameworks: Upernet [110] and Semantic FPN [111] implemented in mmsegmentation [112] with our ImageNet-1 K pre-trained variants as backbones. For training Upernet, we follow the configuration of Swin Transformer and train our variants for 160 K iterations. We employ the AdamW [105] optimizer with a base learning rate of 0.00006, a weight decay of 0.01, a batch size of 16, and a linear scheduler with a linear warmup of 1,500 iterations. As for Semantic FPN 80 K iterations, we follow the same configuration of PVT with a cosine learning rate schedule with an initial learning rate of 0.0002 and a weight decay of 0.0001.

2) Results and Analysis: Table VI shows the results of DilateFormer equipped with UperNet and Semantic FPN frameworks. Our variants DilateFormer-Small/Base equipped with UperNet framework achieve 47.1/50.4% mIoU and 47.6/50.5% MS mIoU, outperforming Swin [26] by at least 2.6% of mIoU and 1.0% of MS mIoU respectively. For Semantic FPN framework, our variants achieve 47.1/48.8% mIoU, and exceed Swin [26] by 3.6-5.6%.

D. Ablation Studies

We conduct ablation studies from the perspectives of sparse and local patterns, dilation scale, block setting, stage setting and overlapping tokenizer/downsampler. More ablation studies about the kernel size are given in the supplementary material.

1) SWDA vs. Other Sparse and Local Patterns: We replace Sliding Window Dilated Attention (SWDA) in the first two

TABLE IV
OBJECT DETECTION AND INSTANCE SEGMENTATION WITH MASK R-CNN ON COCO VAL2017

Method	Params	FLOPs	Mask R-CNN 1× schedule						Mask R-CNN 3× + MS schedule					
	(M)	(G)	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}
Res50 [35]	44	260	38.0	58.6	41.4	34.4	55.1	36.7	41.0	61.7	44.9	37.1	58.4	40.1
NAT-T [32]	48	258	-	-	-	-	-	-	47.7	69.0	52.6	42.6	66.1	45.9
Swin-T [54]	48	264	42.2	64.6	46.2	39.1	61.6	42.0	46.0	68.2	50.2	41.6	65.1	44.8
MPViT-S [42]	43	268	-	-	-	-	-	-	48.4	70.5	52.6	**43.9**	67.6	**47.5**
UniFormer-S_{h14} [46]	41	269	45.6	68.1	49.7	41.6	64.8	**45.0**	48.2	70.4	52.5	43.4	67.1	47.0
Focal-T [90]	49	291	44.8	67.7	49.2	41.0	64.7	44.2	47.2	69.4	51.9	42.7	66.5	45.9
TRT-ViT-C [82]	86	294	44.7	66.9	48.8	40.8	63.9	44.0	47.3	68.8	51.9	42.7	65.9	46.0
PVT-M [76]	64	302	42.0	64.4	45.6	39.0	61.6	42.1	44.2	66.0	48.2	40.5	63.1	43.5
Dilate-S (ours)	44	262	**45.8**	**68.2**	**50.1**	**41.7**	**65.3**	44.7	**49.0**	**70.9**	**53.8**	43.7	**67.7**	46.9
X101-32 [86]	63	340	41.9	62.5	45.9	37.5	59.4	40.2	44.0	64.4	48.0	39.2	61.4	41.9
NAT-S [32]	70	330	-	-	-	-	-	-	48.4	69.8	53.2	43.2	66.9	46.5
TRT-ViT-D [82]	121	375	45.3	67.9	49.6	41.6	64.7	44.8	48.1	69.3	52.7	43.4	66.7	46.8
Focal-S [90]	71	401	47.4	69.8	51.9	42.8	66.6	46.1	48.8	70.5	53.6	43.8	67.7	47.2
PVT-L [76]	81	494	42.9	65.0	46.6	39.5	61.9	42.5	44.5	66.0	48.3	40.7	63.4	43.7
Swin-B [54]	107	496	46.9	-	-	42.3	-	-	48.5	69.8	53.2	43.4	66.8	46.9
MPViT-B [42]	95	503	-	-	-	-	-	-	49.5	70.9	54.0	**44.5**	68.3	**48.3**
Focal-B [90]	110	533	**47.8**	-	-	43.2	-	-	49.0	70.1	53.6	43.7	67.6	47.0
Dilate-B (ours)	67	370	47.6	**70.2**	**55.2**	**43.4**	**67.2**	**46.8**	**49.9**	**71.9**	**55.1**	**44.5**	**68.9**	47.7

TABLE V
OBJECT DETECTION AND INSTANCE SEGMENTATION WITH CASCADE MASK R-CNN ON COCO VAL2017

Method	Params	FLOPs	3× + MS schedule					
	(M)	(G)	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Res50 [35]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
NAT-T [32]	85	737	51.4	70.0	55.9	44.5	67.6	47.9
ConvNeXt-T [55]	86	741	50.4	69.1	54.8	43.7	66.5	47.3
Swin-T [54]	86	745	50.5	69.3	54.9	43.7	66.6	47.1
Shuffle-T [38]	86	746	50.8	69.6	55.1	44.1	66.9	48.0
UniFormer-S_{h14}	79	747	52.1	71.1	56.6	45.2	68.3	48.9
DeiT-S [72]	80	889	48.0	67.2	51.7	41.4	64.2	44.3
Dilate-S (ours)	82	740	**52.4**	**71.6**	**56.9**	**45.2**	**68.6**	**49.0**
X101-32 [86]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
NAT-S [32]	108	809	52.0	70.4	56.3	44.9	68.1	48.6
ConvNeXt-S [55]	108	827	51.9	70.8	56.5	45.0	68.4	49.1
Swin-S [54]	107	838	51.8	70.4	56.3	44.7	67.9	48.5
NAT-B [32]	147	931	52.3	70.9	56.9	45.1	68.3	49.1
ConvNeXt-B [55]	146	964	52.7	71.3	57.2	45.6	68.9	49.5
Swin-B [54]	145	982	51.9	70.5	56.4	45.0	68.1	48.9
Dilate-B (ours)	105	849	**53.5**	**72.4**	**58.0**	**46.1**	**69.9**	**50.3**

stages with other sparse and local patterns, i.e., Dilated Convolution (DC) [89], Dynamic Dilated Convolution (DDC) [93], Window Attention with Spatial Shuffle (WASS[2]) [37] and Sliding Window Attention (SWA) [73].

As shown in Table VII, our SWDA outperforms other sparse and local patterns in various vision tasks. SWDA achieves 82.1% Top-1 accuracy on ImageNet-1 K, 44.9% box mAP/40.9% mask mAP on COCO and 45.84% mIoU on ADE20 K. SWDA outperforms DC (+0.4%, +1.4%/+0.6%, +1.69%) because attention is data-specific compared to conventional convolution. Although DDC is local, sparse and data-specific like SWDA, SWDA still outperforms DDC (+0.3%, +0.6%/+0.3%, +0.94%). DDC uses the entire feature map to generate the kernel parameter of convolution, which is data-specific at the feature-map level; and in comparison, SWDA performs self-attention on keys and values sparsely selected in a sliding window centered on the query patch, which is data-specific at the token level. Therefore, SWDA has a stronger modeling capability than DDC. SWDA also outperforms WASS (+0.3%, +0.8%/+0.5%, +1.18%) and SWA (+0.3%, +0.5%/+0.1%, +2.21%), which demonstrates the importance of considering locality and sparsity in self-attention of the shallow layers.

2) Dilation Scale: Since the number of heads must be multiple of the number of dilation scales, we change the number of heads and feature dimensions in each head, keeping the same total length according to the number of dilation scales. We analyze

[2] The WASS is an approximate sparse sampling operation which divides patches into local Windows like Swin [26] and then shuffles keys and values between different windows.

TABLE VI
SEMANTIC SEGMENTATION EXPERIMENTAL RESULTS ON ADE20 K VALIDATION SET

Method	Upernet 160K Params (M)	 FLOPs (G)	 mIoU (%)	 MS mIoU (%)
Res101 [35]	86	1029	-	44.9
Twins-S [13]	54	901	46.2	47.1
TwinsP-S [13]	55	919	46.2	47.5
ConvNeXt-T [55]	60	939	46.0	46.7
TRT-ViT-B [82]	81	941	46.5	47.5
GG-T [94]	60	942	46.4	47.2
Swin-T [54]	60	945	44.5	45.8
Shuffle-T [38]	60	949	46.6	**47.8**
Focal-T [90]	62	998	45.8	47.0
Dilate-S (ours)	54	935	**47.1**	47.6
NAT-S [32]	82	1010	48.0	49.5
Twins-B [13]	89	1020	47.7	48.9
ConvNeXt-S [55]	82	1027	48.7	49.6
Swin-S [54]	81	1038	47.6	49.5
GG-S [94]	81	1035	48.4	49.6
TRT-ViT-D [82]	144	1065	48.8	49.8
Shuffle-B [38]	121	1196	49.0	50.5
Next-ViT-L [45]	92	1072	50.1	50.8
Focal-S [90]	85	1130	48.0	50.0
Dilate-B (ours)	79	1046	**50.8**	**51.1**

Method	Semantic FPN 80K Params (M)	 FLOPs (G)	 mIoU (%)
Res50 [35]	29	183	36.7
Twins-S [13]	28	144	43.2
PVT-S [76]	28	161	39.8
TwinsP-S [13]	28	162	44.3
XCiT-S12/8 [1]	30	-	44.2
TRT-ViT-B [82]	46	176	45.4
Swin-T [54]	32	182	41.5
Next-ViT-S [45]	36	208	46.5
CrossFormer-S [77]	34	209	46.4
Dilate-S (ours)	28	178	**47.1**
Res101 [35]	48	260	38.8
Next-ViT-B [45]	49	260	48.6
XCiT-S24/8 [1]	52	-	47.1
Swin-S [54]	53	274	45.2
PVT-L [76]	65	283	42.1
TwinsP-L [13]	65	283	46.4
TRT-ViT-D [82]	106	296	46.7
CrossFormer-B [77]	56	320	48.0
Swin-B [54]	91	422	46.0
Dilate-B (ours)	51	288	**48.8**

Left: With upernet; Right: With semantic FPN.

TABLE VII
EXPERIMENT RESULTS WITH LOCAL AND SPARSE PATTERNS IN THE FIRST TWO STAGES

Pattern	Locality	Sparity	Data Specific	Top-1 (%)	AP^b (%)	AP^m (%)	mIoU (%)
DC	✓	✓		81.7	43.5	40.3	44.15
DDC	✓	✓	✓	81.8	44.3	40.6	44.90
WASS		✓	✓	81.8	44.1	40.4	44.66
SWA	✓		✓	81.8	44.4	40.8	43.63
SWDA	✓	✓	✓	**82.1**	**44.9**	**40.9**	**45.84**

The Top-1 is on Imagenet-1 k, AP^b and AP^m are on COCO VAL 2017 with mask R-CNN 1× schedule, MIOU is on ADE20 K validation set with semantic FPN.

TABLE VIII
TOP-1 ACCURACY ON IMAGENET-1 K OF DIFFERENT DILATION SCALES

Scale	Head Num in Stage1/2	Dilation Rate	Top-1 (%)
Multi-	[2, 4]	[1, 2]	81.7
	[3, 6]	[1, 2, 3]	**82.1**
		[2, 3, 4]	81.8
		[3, 4, 5]	81.7
	[4, 8]	[1, 2, 3, 4]	81.9
Single	[3, 6]	[1]	81.7
		[2]	81.9
		[3]	81.8

the effects of dilation scales according to the performance on ImageNet-1 K classification task. The number of heads in stage 1 or 2, dilation scales and top-1 accuracy are shown in Table VIII. With the same number of heads in the block, the top-1 accuracy (82.1%) of multi-scale dilated attention, i.e., [1, 2, 3], is better than that of single-scale, i.e., [1], [2], and [3], because multi-scale can provide richer information than single-scale. What is more, the dilation rates in the block need to be moderate so that it can simultaneously model both locality and sparsity of attention, without introducing redundant information modeling due to the large receptive field such as global attention. Therefore, we set the dilation scale of the model to 3, i.e., [1, 2, 3] by default.

3) MSDA vs. Other Blocks Setting: In our DilateFormer, we stack Multi-Scale Dilated Attention (MSDA) blocks in the first two stages. To demonstrate the effectiveness of our proposed MSDA, we replace MSDA in the first two stages of the default setting (D-D-G-G) with local attention in a shifted window (L-L-G-G) [26] and global attention (G-G-G-G) [21] for comparison. We also compare with the global attention cooperated with a naïve downsampling technique, namely global attention with spatial reduction (G-G-G-G + sr.) [31], which reduces the redundant interaction between patches by decreasing the number of patches. The maximum size of attended receptive field in MSDA is 7 × 7 with dilation, the size of attended receptive field in local attention is 7 × 7, and the size of attended receptive field in global attention is the size of the entire feature map.

Table IX summarizes the comparison results. By using the same size of maximum attended receptive field, our MSDA (82.1%) outperforms local attention with shifted window (L-L-G-G) [26] (81.7%) with fewer FLOPs, which demonstrates the effectiveness of sparse and local attention mechanisms in shallow layers. Compared with the global attention (G-G-G-G) [21], our MSDA achieves an improvement of 0.3% with half of FLOPs, which further demonstrates the effectiveness and efficiency of the proposed local and sparse attention mechanism. Also, the superiority of MSDA against the global attention shows the redundancy of modeling dependencies among all image patches. To reduce the redundant interaction, the global attention with spatial reduction utilizes downsampling by convolution but

TABLE IX
EXPERIMENT RESULTS WITH DIFFERENT BLOCKS IN STAGE1/2

Block Type	Params (M)	FLOPs (G)	Top-1 (%)	AP^b (%)	AP^m (%)	mIoU (%)
G-G-G-G + sr.	20.6	3.02	81.6	40.9	37.9	44.4
G-G-G-G	17.2	6.36	81.8	42.0	38.7	44.5
L-L-G-G	17.2	3.24	81.7	40.9	37.6	44.3
D-D-G-G	17.2	3.18	**82.1**	**44.2**	**40.9**	**45.8**

“SR.” Indicates spatial reduction operation. “D,” “G” and “L” indicate dilation, global and local operations, respectively. The top-1 is on imagenet-1 k, AP^b and AP^m are on COCO VAL2017 with mask R-CNN 1 × schedule, miou is on ADE20 K validation set with semantic FPN.

TABLE X
ANALYSIS OF MULTI-SCALE DILATED ATTENTION BLOCKS IN DIFFERENT STAGES ON IMAGENET-1 K

Stage Setting	FLOPs (G)	Top-1 (%)
G-G-G-G	6.36	81.8
D-G-G-G	3.53	82.2
D-D-G-G	3.18	**82.1**
D-D-D-G	3.05	81.3
D-D-D-D	3.04	80.5

TABLE XI
TOP-1 ACCURACY ON IMAGENET-1 K OF USING OVERLAPPING TOKENIZER AND DOWNSAMPLER

Overlapping Tokenizer	Overlapping Downsampler	Params (M)	FLOPs (G)	Top-1 (%)
		16.1	2.62	81.7
	✓	17.2	2.74	81.8
✓		16.2	3.12	81.9
✓	✓	17.2	3.18	**82.1**

introduces extra parameters. By contrast, our MSDA exploits the locality and sparsity property without extra parameters. The results show that our MSDA surpasses the global attention with spatial reduction by 0.5%, which indicates the effectiveness of redundancy reduction of the proposed MSDA. In downstream tasks, our MSDA block also outperforms other types of blocks, indicating that MSDA has a stronger modeling capability.

4) Stage Setting: To demonstrate the modeling capability of the MSDA block at shallow stages, we conduct a set of experiments to explore the performance of using MSDA in different stages. In the four stages of the model, we progressively replace the global MHSA block in each stage with the MSDA block. Table X shows FLOPs and top-1 accuracy of models with different structures. The model performance shows a decreasing trend, from 82.2% down to 80.5%, as the proportion of MSDA blocks in the model stage increases. The results show that it is more effective to consider the locality and sparsity property of the self-attention mechanism in shallow stages rather than in deeper stages. What's more, the model with MSDA block only in stage1 (82.2%) performs slightly better than the model with MSDA blocks in both stage1 and stage2 (82.1%), but the former has larger FLOPs (+ 0.35 G). Therefore, we use MSDA blocks in both stage1 and stage2 by default.

TABLE XII
COMPARISON OF MODEL INFERENCE

Method	Params (M)	FLOPs (G)	FPS (s)	Mem. (G)	Top-1 (%)
ConvNeXt-T [55]	28	4.5	2450	3.5	82.1
Swin-T [54]	28	4.5	1681	5.0	81.3
NAT-T [32]	28	4.5	1515	3.7	83.2
DiNAT-T [31]	28	4.5	1479	3.7	82.7
Dilate-S (ours)	21	4.8	1368	3.1	**83.3**
ConvNeXt-S [55]	50	8.7	1558	4.8	83.1
Swin-S [54]	50	8.7	1045	6.7	83.0
NAT-S [32]	51	7.8	1055	5.0	83.7
DiNAT-S [31]	51	7.8	1069	5.0	83.8
Dilate-B (ours)	47	10.0	1122	4.0	**84.4**

“MEM” denotes the peak memory for evaluation. “FPS” is the number of images processed for one second.

5) Overlapping Tokenizer/Downsampler: We further study how the overlapping tokenizer or downsampler affect the performance. While keeping the same settings, we replace our overlapping tokenizer or downsampler with a simple non-overlapping tokenizer or downsampler, i.e., convolution with kernel size 4 and stride 4 or convolution with kernel size 2 and stride 2. As shown in Table XI, our model achieves a slight improvement (+0.4%) with overlapping tokenizer/downsampler, indicating that the main improvement of our model does not rely on these two modules.

6) Comparisons of Real Running Times: We provide a comparison of model inference about FPS, peak memory about our DilateFormers and current SOTA models in Table XII. FPS and peak memory usage are measured from forward passes with a batch size of 256 on a single A100 GPU. With comparable model parameters and FLOPs, our DilateFormers have comparable FPS and better performance than current SOTA models.

7) Grad-CAM Visualization: To further illustrate the recognition ability of the proposed DilateFormer, we apply Grad-CAM [114] to visualize the areas of the greatest concern in the last layer of DeiT-Tiny [22], Swin-Tiny [26] and Dilate-Tiny. As shown in Fig. 6, our Dilate-Tiny model performs better in locating the target objects and attends to semantic areas more continuously and completely, suggesting the stronger recognition ability of our model. Such ability yields better classification performance compared with DeiT-Tiny and Swin-Tiny.

8) More Visualization Results on Global Attention: In Section I, we discuss two key properties i.e., *locality* and *sparsity* of global attention in shallow layers. To further analyze these two properties, we visualize more attention maps in the shallow layers of ViT-Small [21]. As shown in Fig. 7, the attention maps in the shallow layers of ViT-Small show that activated key patches are sparsely distributed in the neighborhood of the query patch. Specifically, the patches with high attention scores sparsely scatter around the query patch and other patches have low attention scores.

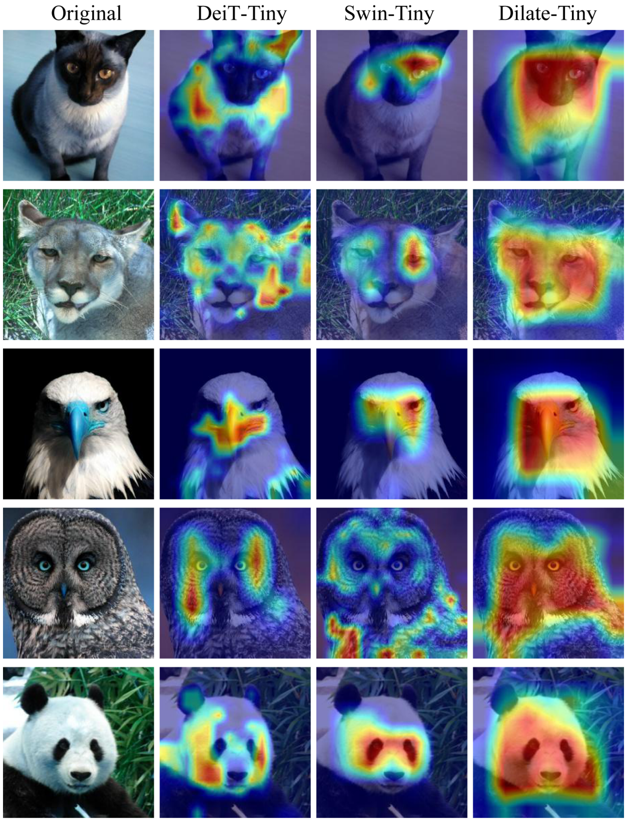


Fig. 6. Grad-CAM Visualization of the last layer of DeiT-Tiny, Swin-Tiny and Dilate-Tiny. Images are from the validation set of ImageNet-1 k.

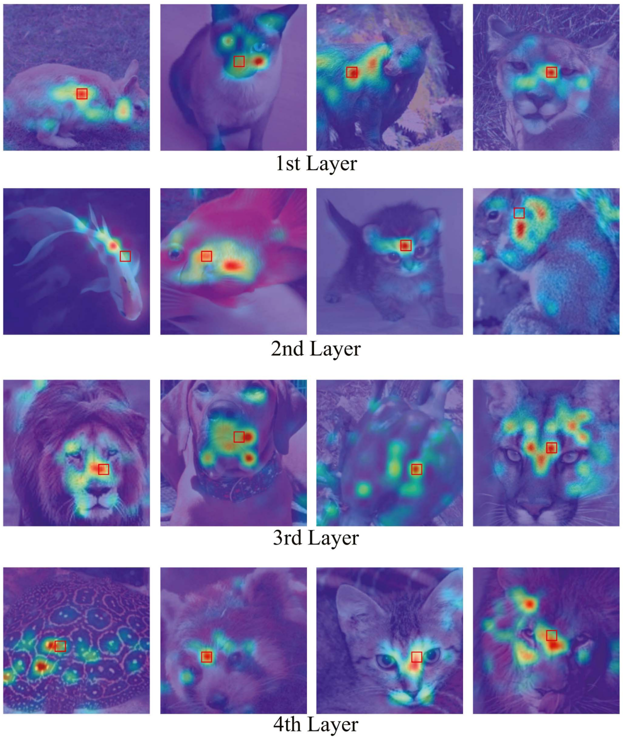


Fig. 7. More Visualization of attention maps of shallow layers of ViT-Small. We visualize the activations in attention maps of the query patches (in the red box). The attention maps show that patches with high attention scores sparsely scatter around the query patch, and other patches have low attention scores.

V. Conclusion

In this work, we propose a strong and effective Vision Transformer, called DilateFormer, which can provide powerful and general representations for various vision tasks. Our proposed Multi-Scale Dilated Attention (MSDA) takes both the locality and sparsity of the self-attention mechanism in the shallow layers into consideration, which can effectively aggregate semantic multi-scale information and efficiently reduce the redundancy of the self-attention mechanism without complex operations and extra computational cost. Extensive experiment results show that the proposed method achieves state-of-the-art performance in both ImageNet-1 k classification and down-stream vision tasks such as object detection and semantic segmentation.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, 2012, pp. 84–90. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[4] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594

[5] Z. Liu et al., "A ConvNet for the 2020 s," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.

[6] Y.-X. Peng, J. Jiao, X. Feng, and W.-S. Zheng, "Consistent discrepancy learning for intra-camera supervised person re-identification," *IEEE Trans. Multimedia*, early access, Jan. 27, 2022, doi: 10.1109/TMM.2022.3146775.

[7] B. Zhao et al., "COMO: Efficient deep neural networks expansion with convolutional maxout," *IEEE Trans. Multimedia*, vol. 23, pp. 1722–1730, 2021. [Online]. Available: https://doi.org/10.1109/TMM.2020.3002614

[8] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1137–1149. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[10] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Comput. Vis. 14th Eur. Conf., Amsterdam*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[11] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988. [Online]. Available: https://doi.org/10.1109/ICCV.2017.324

[12] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969, doi: 10.1109/ICCV.2017.322.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.- Assist. Intervention 18th Int. Conf. Munich*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[14] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[16] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Non-Local_Neural_Networks_CVPR_2018_paper.html

[17] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/languageunsupervised/languageunderstandingpaper.pdf

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[20] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[21] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[22] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357. [Online]. Available: http://proceedings.mlr.press/v139/touvron21a.html

[23] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 32–42, doi: 10.1109/ICCV48922.2021.00010.

[24] X. Chu, B. Zhang, Z. Tian, X. Wei, and H. Xia, "Conditional positional encodings for vision transformers," 2021, *arXiv:2102.10882*.

[25] X. Ma et al., "Spatial pyramid attention for deep convolutional neural networks," *IEEE Trans. Multimedia*, vol. 23, pp. 3048–3058, 2021, doi: 10.1109/TMM.2021.3068576.

[26] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.

[27] J. Yang et al., "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*. https://arxiv.org/abs/2107.00641

[28] P. Zhang et al., "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2998–3008. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00299

[29] X. Chu et al., "Twins: Revisiting the design of spatial attention in vision transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9355–9366. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/4e0928de075538c593fbdabb0c5ef2c3-Abstract.html

[30] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," 2022, *arXiv:2204.07143*.

[31] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578, doi: 10.1109/ICCV48922.2021.00061.

[32] Z. Jiang et al., "All tokens matter: Token labeling for training better vision transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 18590–18602. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html

[33] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[34] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Comput. Vis. 13th Eur. Conf.*, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[35] B. Zhou et al., "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 633–641. doi: 10.1109/CVPR.2017.544.

[36] Q. Yu et al., "Glance-and-gaze vision transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12992–13003.

[37] Z. Huang et al., "Shuffle transformer: Rethinking spatial shuffle for vision transformer," 2021, *arXiv:2106.03650*.

[38] Z. Tu et al., "Maxvit: Multi-axis vision transformer," in *Proc. Comput. Vis. 17th Eur. Conf.*, 2022, pp. 459–479, doi: 10.1007/978-3-031-20053-3_27.

[39] W. Wang et al., "CrossFormer: A versatile vision transformer based on cross-scale attention," 2021, *arXiv:2108.00154*.

[40] Y. Chen et al., "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5260–5269, doi: 10.1109/CVPR52688.2022.00520.

[41] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 367–376, doi: 10.1109/ICCV48922.2021.00042.

[42] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10843–10852, doi: 10.1109/CVPR52688.2022.01058.

[43] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 28522–28535.

[44] K. Li et al., "UniFormer: Unifying convolution and self-attention for visual recognition," 2022, *arXiv:2201.09450*.

[45] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.

[46] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1204–1213, doi: 10.1109/CVPR52688.2022.01179.

[47] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[48] S. Yun et al., "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032, doi: 10.1109/ICCV.2019.00612.

[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[50] E. Hoffer et al., "Augment your batch: Improving generalization through instance repetition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8129–8138. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Hoffer_Augment_Your_Batch_Improving_Generalization_Through_Instance_Repetition_CVPR_2020_paper.html

[51] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7000

[52] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10033–10041, doi: 10.1109/ICCV48922.2021.00988.

[53] X. Yue et al., "Vision transformer with progressive sampling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 387–396, doi: 10.1109/ICCV48922.2021.00044.

[54] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, doi: 10.1109/TPAMI.2022.3206108.

[55] Z. Zhang et al., "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3417–3425.

[56] P. Xie, M. Zhao, and X. Hu, "PiSLTRc: Position-informed sign language transformer with content-aware convolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3908–3919, 2022, doi: 10.1109/TMM.2021.3109665.

[57] C. Zhang et al., "Self-attention-based multiscale feature learning optical flow with occlusion feature map prediction," *IEEE Trans. Multimedia*, vol. 24, pp. 3340–3354, 2022, doi: 10.1109/TMM.2021.3096083.

[58] L. Lin et al., "Structured attention network for referring image segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 1922–1932, 2022, doi: 10.1109/TMM.2021.3074008.

[59] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, Feb. 2023, doi: 10.1109/TPAMI.2022.3164083 2021,.

[60] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12114–12124, doi: 10.1109/CVPR52688.2022.01181.

[61] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Comput. Vis. 16th Eur. Conf., Glasgow,*, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[62] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3621–3630, doi: 10.1109/ICCV48922.2021.00360.

[63] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7373–7382. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Dai_Dynamic_Head_Unifying_Object_Detection_Heads_With_Attentions_CVPR_2021_paper.html

[64] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked U-shape network with channel-wise attention for salient object detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1397–1409, 2021, doi: 10.1109/TMM.2020.2997192.

[65] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7277–7286, doi: 10.1109/CVPR52688.2022.00714.

[66] R. Guo, D. Niu, L. Qu, and Z. Li, "SOTR: Segmenting objects with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7157–7166, doi: 10.1109/ICCV48922.2021.00707.

[67] F. Zhu et al., "A unified efficient pyramid transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2667–2677, doi: 10.1109/ICCVW54120.2021.00301.

[68] L. Zhou, C. Gong, Z. Liu, and K. Fu, "SAL: Selection and attention losses for weakly supervised semantic segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 1035–1048, 2021, doi: 10.1109/TMM.2020.2991592.

[69] Y. Chen et al., "ResT-ReID: Transformer block-based residual learning for person re-identification," *Pattern Recognit. Lett.*, vol. 157, pp. 90–96, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786552200085X

[70] S. He et al., "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15013–15022. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01474

[71] Y. Zheng, Z. Zhao, X. Yu, and D. Yu, "Template-aware transformer for person reidentification," *Comput. Intell. Neurosci.*, pp. 1–12, 2022.

[72] X. Gong et al., "LAG-net: Multi-granularity network for person re-identification via local attention system," *IEEE Trans. Multimedia*, vol. 24, pp. 217–229, 2022, doi: 10.1109/TMM.2021.3050082.

[73] G. Li, D. Xu, X. Cheng, L. Si, and C. Zheng, "Simvit: Exploring a simple vision transformer with sliding windows," in *Proc. IEEE Int. Conf. Multimedia Expo*2021, pp. 1–6.

[74] T. Xiao et al., "Early convolutions help transformers see better," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 30392–30400. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html

[75] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12165–12175, doi: 10.1109/CVPR52688.2022.01186.

[76] K. Yuan et al., "Incorporating convolution designs into visual transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 579–588, doi: 10.1109/ICCV48922.2021.00062.

[77] A. Srinivas et al., "Bottleneck transformers for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Srinivas_Bottleneck_Transformers_for_Visual_Recognition_CVPR_2021_paper.html

[78] R. Xia, Y. Li, and W. Luo, "Laga-Net: Local-and-global attention network for skeleton based action recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 2648–2661, 2022, doi: 10.1109/TMM.2021.3086758.

[79] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15979–15988, doi: 10.1109/CVPR52688.2022.01553.

[80] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9643–9653, doi: 10.1109/CVPR52688.2022.00943.

[81] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE Trans. Multimedia*, vol. 24, pp. 1775–1786, 2022, doi: 10.1109/TMM.2021.3072479.

[82] H. Liu, J. Li, D. Li, J. See, and W. Lin, "Learning scale-consistent attention part network for fine-grained image recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 2902–2913, 2022, doi: 10.1109/TMM.2021.3090274.

[83] J. Chen, X. Li, L. Luo, and J. Ma, "Multi-focus image fusion based on multi-scale gradients and image matting," *IEEE Trans. Multimedia*, vol. 24, pp. 655–667, 2022, doi: 10.1109/TMM.2021.3057493.

[84] Y. Liu et al., "Multi-scale grid network for image deblurring with high-frequency guidance," *IEEE Trans. Multimedia*, vol. 24, pp. 2890–2901, 2022, doi: 10.1109/TMM.2021.3090206.

[85] Y. Zuo et al., "Mig-Net: Multi-scale network alternatively guided by intensity and gradient features for depth map super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3506–3519, 2022, doi: 10.1109/TMM.2021.3100766.

[86] W. Yu et al., "Metaformer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10819–10829.

[87] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.

[88] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 357–366.

[89] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. 4th Int. Conf. Learn. Representations*, 2016, pp. 1–13. [Online]. Available: http://arxiv.org/abs/1511.07122

[90] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. 33nd Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999. [Online]. Available: http://proceedings.mlr.press/v48/cohenc16.html

[91] Y. Ma, H. Shuai, and W. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Trans. Multimedia*, vol. 24, pp. 261–273, 2022, doi: 10.1109/TMM.2021.3050059.

[92] Z. Yan et al., "Crowd counting via perspective-guided fractional-dilation convolution," *IEEE Trans. Multimedia*, vol. 24, pp. 2633–2647, 2022, doi: 10.1109/TMM.2021.3086709.

[93] Y. Chen et al., "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11030–11039.

[94] A. Hassani and H. Shi, "Dilated neighborhood attention transformer," 2022, *arXiv:2209.15001*.

[95] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10428–10436.

[96] Q. Han et al., "On the connection between local attention and dynamic depth-wise convolution," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–25. [Online]. Available: https://openreview.net/forum?id=L3_SsSNMmy

[97] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 3965–3977.

[98] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[99] K. Han et al., "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*2021,vol. 34, pp. 15908–15919.

[100] Z. Zhang et al., "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," *AAAI*, vol. 36, no. 3, pp. 3417–3425, 2021, doi: 10.1609/aaai.v36i3.20252.

[101] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on imagenet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.

[102] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[103] J. Li et al., "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022, *arXiv:2207.05501*.

[104] D. Zhou et al., "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[105] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations*, 2019, pp. 1–19. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[106] X. Xia et al., "Trt-vit: TensorRT-oriented vision transformer," 2022, *arXiv:2205.09579*.

[107] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500. [Online]. Available: https://doi.org/10.1109/CVPR.2017.634

[108] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021. [Online]. Available: https://doi.org/10.1109/TPAMI.2019.2956516

[109] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.

[110] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Comput. Vis. 15th Eur. Conf.*, 2018, pp. 418–434, doi: 10.1007/978-3-030-01228-1_26.

[111] A. Kirillov, R. B. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6399–6408. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Feature_Pyramid_Networks_CVPR_2019_paper.html

[112] M. Contributors, "MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark," 2020. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

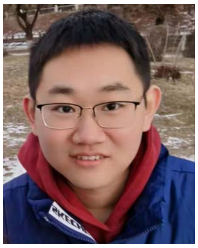
[113] A. Ali et al., "XCiT: Cross-covariance image transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 20014–20027. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/a655fbe4b8d7439994aa37ddad80de56-Abstract.html

[114] J. Gildenblat and contributors, "Pytorch library for cam methods," 2021. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

Jiayu Jiao received the bachelor's degree in information and computing science from the South China University of Technology, Guangzhou, China, in 2021. He is working toward the master's degree with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou. His research interests include computer vision and machine learning.

Yu-Ming Tang received the bachelor's degree in electronic information science and technology in 2021 from Sun Yat-sen University, Guangzhou, China, where he is currently working toward the Ph.D. degree with the School of Computer Science and Engineering. His research interests include computer vision and machine learning.

Kun-Yu Lin received the B.S. and M.S. degrees from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He is currently working toward the Ph.D. degree with the School of Computer Science and Engineering. His research interests include computer vision and machine learning.



Yipeng Gao received the bachelor's degree from the South China University of Technology, Guangzhou, China, in 2021. He is currently working toward the M.S. degree with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include computer vision and machine learning.

Andy J. Ma received the B.Sc. and M.Sc. degrees in applied mathematics from Sun Yat-sen University, Guangzhou, China, and the Ph.D. degree in computer science, Hong Kong Baptist University, Hong Kong. He was a Postdoctoral Fellow at Rutgers University, New Brunswick, NJ, USA, and Johns Hopkins University, Baltimore, MD, USA. He is currently an Associate Professor with Sun Yat-sen University. He has authored or coauthored more than 60 papers in top-tier journals and conferences including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, Critical Care Medicine, Alimentary Pharmacology & Therapeutics, ICCV, CVPR, ECCV, AAAI, IJCAI, MICCAI. His research interests include developing machine learning algorithms for intelligent video surveillance and medical applications. He is an Associate Editor for the *SPIE Journal of Electronic Imaging*.

Yaowei Wang (Member, IEEE) received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2005. He was a Distinguished Professor at the National Engineering Laboratory for Video Technology Shenzhen (NELVT), Peking University Shenzhen Graduate School, Shenzhen, China, in 2019. He is currently a Professor with the Peng Cheng Laboratory, Shenzhen, China. He is the author or coauthor of more than 120 technical articles in international journals and conferences, including TOMM, ACM MM, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, ICCV, IJCAI, and AAAI. His research interests include multimedia content analysis and understanding, machine learning and computer vision. He is the Chair of the IEEE Digital Retina Systems Working Group and a Member of CIE, CCF, CSIG. He was the recipient of the second prize of the National Technology Invention in 2017 and the first Prize of the CIE Technology Invention in 2015.

Wei-Shi Zheng received the Ph.D. degree in applied mathematics from Sun Yat-sen University, Guangzhou, China, in 2008. He is currently a full Professor with Sun Yat-sen University. His research interests include person/object association and activity understanding in visual surveillance, and the related large-scale machine learning algorithm. Especially, Dr. Zheng has active research on person re-identification in the last five years. He has ever joined Microsoft Research Asia Young Faculty Visiting Programme. He was the Area Chairs of CVPR, ICCV, BMVC and IJCAI. He is an IEEE MSA TC Member. He is an Associate Editor for the *Pattern Recognition Journal*. He was the recipient of the Excellent Young Scientists Fund of the National Natural Science Foundation of China, and the recipient of the Royal Society-Newton Advanced Fellowship of the United Kingdom.