DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition

Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, Wei-Shi Zheng*

*Corresponding author: Wei-Shi Zheng.

Code: https://github.com/JIAOJIAYUASD/dilateformer

Project page: https://isee-ai.cn/ jiaojiayu/DilteFormer.html

Submission date: 22-Sep-2022 to IEEE Transaction on Multimedia

DilateFormer: Multi-Scale Dilated Transformer for Visual Recognition

Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, 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-1K 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-1K classification task, 53.5% box mAP/46.1% mask mAP on COCO object detection/instance segmentation task and 51.1% MS mIoU on ADE20K semantic segmentation task.**

***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 [35], [41], [55], [60], [68], [70], [107], object detection [34], [49], [52], [64], [65] and semantic segmentation [9], [56], [67], 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 [79]. Inspired by sequence modeling tasks [3], [62] in natural language processing (NLP) [2], [62], [76], pioneer works [14], [22], [58], [73], [74] introduce Transformers with long-range dependency modeling ability into computer vision, achieving exciting results in various vision tasks.

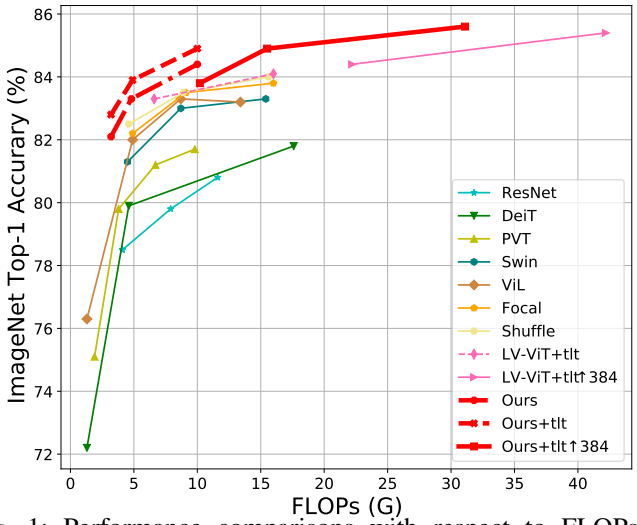

Fig. 1: Performance comparisons with respect to FLOPs on ImageNet-1K classification. Without extra training data, our DilateFormer variants achieve comparable or even better performance with fewer FLOPs.

With global attention, the vanilla Vision Transformers (ViTs) [22], [73] 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 [13], [32], [54], [91], [104] 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 [22], [73], we find that the attention matrix in shallow layers has two key properties, namely *locality* and *sparsity*. As shown in Figure 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

J. Jiao and Y.-M. Tang are equally-contributed authors. *Corresponding author: W.-S. Zheng.

J. Jiao, Y.-M. Tang, K.-Y. Lin, Y. Gao and J. Ma are with the School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China (e-mail: {jiaojy6, tangym9, linky5, gaoyp23}@mail2.sysu.edu.cn, majh8@mail.sysu.edu.cn).

W.-S. Zheng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, with Peng Cheng Laboratory, Shenzhen, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China. (e-mail: wszheng@ieee.org /zhwshi@mail.sysu.edu.cn).

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

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.

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 learning. Following PVT [77] and Swin [54], 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 [22], [73] 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 [22], [32], [54], [73], [91], [104] on various datasets across different model sizes. As depicted in Figure 1, we demonstrate the performance of our DilateFormers on ImageNet-1K classification task. Without extra training data, our Dilate-S (4.8 GFLOPs) achieves comparable performance with Swin-B (15.4 GFLOPs) [54] on ImageNet-1K using only 1/3 FLOPs. With the assistance of Token Labeling [39], our DilateFormers achieve better performance than LV-ViTs [39] 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 [39] (6.6 GFLOPs) and LV-ViT-M [39](16 GFLOPs). Besides, our Dilate-B achieves 85.6% top-1 accuracy on ImageNet-1K classification [19] task, 53.5% box mAP/46.1% mask mAP on COCO [50] object detection/instance segmentation task and 51.1% MS mIoU on ADE20K [110] 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.

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

A. Global Attention in Vision Transformers

Inspired by the success in NLP [3], [20], [76], the vanilla Vision Transformers (ViTs) [22], [73] directly apply self-attention mechanisms to patches split from images. By utilizing sufficient training data [22], [101] and strong data augmentation strategies [37], [71], [73], [100], [103], [109], Transformer-based methods [48], [58], [81], [86], [98], [99], [102], [105] achieve exciting performance improvements on various vision tasks, *i.e.*, image classification [21], [46], [47], [54], [61], [73], object detection [5], [17], [21], [24], [32], [44], [58], [104], semantic segmentation [13], [28], [42], [48], [66], [112], [113], and re-identification [12], [36], [108]. 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 Sec. 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 [13], [77], [94] 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 [21], [26], [54] apply the self-attention mechanism to patches in a fixed local region to reduce computational cost. For example, Swin [54] 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 [21] improves the window-fixed setting in Swin [54], performing self-attention to cross-shaped windows. Other works [38], [75], [78] use grouped sampling or spatial shuffling operation for information exchange between different local windows. Inspired by the convolution operation in CNNs [35], [41], [43], [55], [70], ViL [104] and NAT [32] propose sliding window attention, which models dependencies only with neighboring patches in the window centering each query patch. Moreover, some works [10], [27], [61], [69], [82], [85], [96] combine CNNs and Transformers

TABLE I: COMPARISON OF TECHNICAL DETAILS WITH OTHER MODELS. "-" 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.

Model	Overlapping Tokenizer	Overlapping Downsampler	Positional Embedding	Attention			Multi-scale	
				Local	Global	Sparse	Stage-level	Block-level
ViT [22]/DeiT [73]	×	-	APE	×	✓	×	×	×
PVT [77]	×	×	APE	×	✓	×	✓	×
Swin [54]	×	×	RPE	✓	×	×	✓	×
Twins [13]	×	×	CPE	✓	✓	×	✓	×
GG [94]	×	×	RPE/APE	×	×	✓	✓	×
Shuffle [38]	✓	×	RPE	✓	×	✓	✓	×
MaxViT [75]	✓	✓	CPE	✓	×	✓	✓	×
CrossFormer [78]	✓	✓	RPE	✓	×	✓	✓	✓
ViL [104]	×	×	RPE/APE	✓	✓	×	✓	×
NAT [32]	✓	✓	RPE	✓	×	×	✓	×
Mobile-Former [10]	-	-	CPE	×	✓	×	✓	×
Conformer [61]	✓	-	-	×	✓	×	✓	×
Shunted [66]	✓	✓	CPE	×	✓	×	✓	✓
MPViT [66]	✓	✓	CPE	×	✓	×	✓	✓
ViTAE [89]	✓	-	APE	×	✓	×	✓	✓
UniFormer [46]	✓	✓	CPE	×	✓	×	✓	×
Focal [91]	×	×	RPE	✓	✓	×	✓	✓
DilateFormer (ours)	✓	✓	CPE	✓	✓	✓	✓	✓

for introducing the locality prior, and they usually design handcrafted and complex modules for interaction between CNNs and Transformers features, leading to a lack of scalability to large-scale parameters [33], [88]. However, some works [21], [32], [54], [104] above only consider the locality of the self-attention mechanism and lack consideration of the sparsity. Although some works [38], [75], [78], [94] 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 performs self-attention in a dilated window centered on query patch.

C. Multi-scale Vision Transformer

The vanilla Vision Transformer [22], [73] is a "columnar" structure for visual tasks. Since multi-scale information [7], [35], [41], [51], [53], [70], [93], [114] is beneficial for dense prediction tasks such as object detection, instance and semantic segmentation, recent works [21], [23], [32], [38], [42], [54], [58], [66], [77], [78], [95], [104] introduce multi-scale modeling capability by using a pyramid structure to design their transformer backbones. Several works [6], [10], [42], [61], [66], [78], [89], [91] introduce multi-scale information in patch embedding layers [78] or self-attention blocks [42], [66], [91] or add extra branches [10], [61], [89] to perform convolution operation. CrossFormer [78] utilizes different convolution operations or different patch sizes for designing patch embedding. Shunted Transformer [66] uses multi-scale token aggregation for obtaining keys and values of various sizes. MPViT [42] consists of multi-scale patch embedding and multi-path transformer blocks. Conformer [61], Mobile-Former [10] and ViTAE [89] 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 [35], [41], [55], [70] usually use downsampling or convolution with a large stride to increase the receptive field and reduce computational cost. However, these approaches [35], [41], [55], [70] result in reduced resolution of feature maps, affecting model performance in many tasks such as object detection [34], [49], [52], [64], [65] and semantic segmentation [9], [56], [67]. Therefore, Cohen et al. [15], [92] propose dilated convolution [59], [90], 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 [11], 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

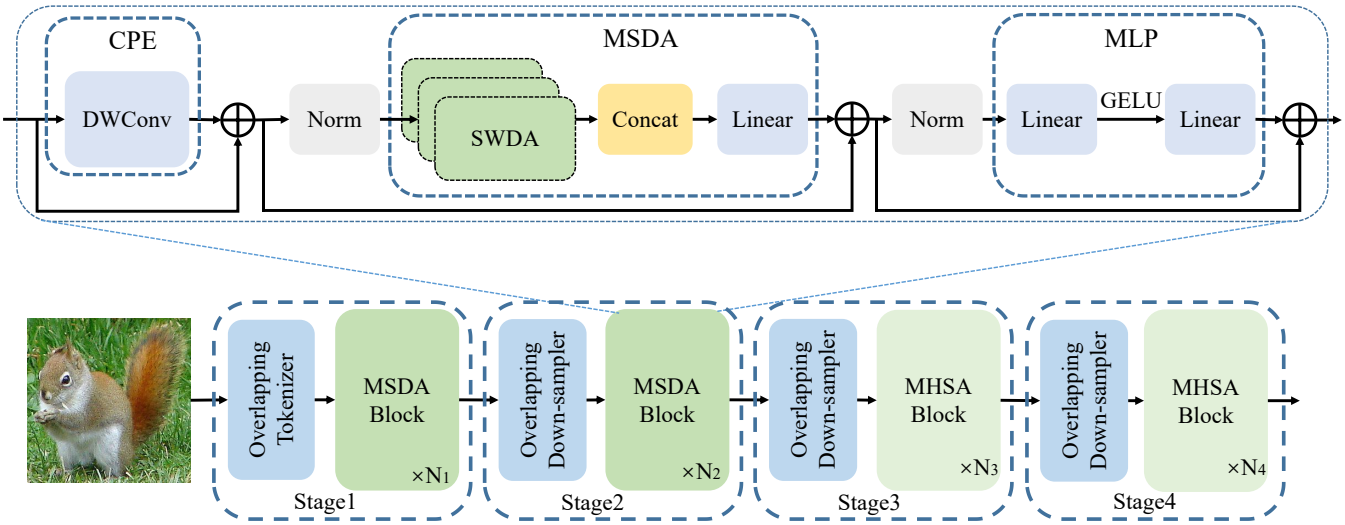

Fig. 3: The overall architecture of our DilateFormer. The top part shows the proposed Multi-Scale Dilated Attention (MSDA) block, consisting of DwConv, Multi-Scale Sliding Window Dilated Attention operation (SWDA) and MLP. The bottom part shows DilateFormer, consisting of Overlapping Tokenizer, Overlapping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

data-specific at the token level. In addition, we also notice a concurrent work, DiNAT [31], 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 Sec.III-A, we introduce our Sliding Window Dilated Attention (SWDA) operation, towards effective long-range dependency modeling in feature maps. In Sec.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 Sec.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:

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

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:

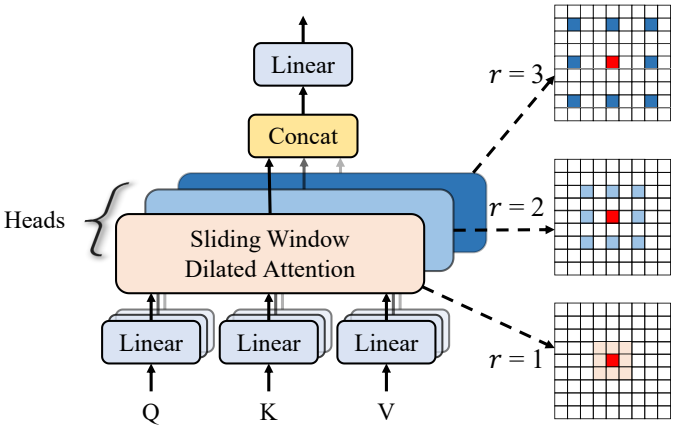

Fig. 4: **Illustration of Multi-Scale Dilated Attention (MSDA).** First, the channels of the feature map are split into different heads. Then, the self-attention operation is performed among the colored patches in the window surrounding the red query patch, using different dilation rates in different 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.

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

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 $(i',\ j')$

TABLE II: MODEL VARIANTS OF OUR DILATEFORMER. 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.

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	$\begin{bmatrix} \text{576-d, 24-h} \end{bmatrix} \times 2$	$\begin{bmatrix} \text{576-d, 24-h} \end{bmatrix} \times 3$	$\begin{bmatrix} \text{768-d, 24-h} \end{bmatrix} \times 3$

will be selected for conducting self-attention:

$$\begin{aligned} \Big\{(i', j')\Big| i' = i + p \times r, j' = j + q \times r\Big\}, \\ -\frac{w}{2} \leq p,\ q \leq \frac{w}{2}. \end{aligned} \tag{3}$$

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 Figure 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:

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

$$X = \mathrm{Linear}(\mathrm{Concat}[h_1, ..., h_n]), \tag{5}$$

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 Figure 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 Sec. III-B while the latter two stages utilize ordinary Multi-Head Self-Attention (MHSA). What's more, we use the overlapping tokenizer [85] for patch 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 [32], 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 [14] 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 [54], [73], 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 [19] classification, COCO [50]

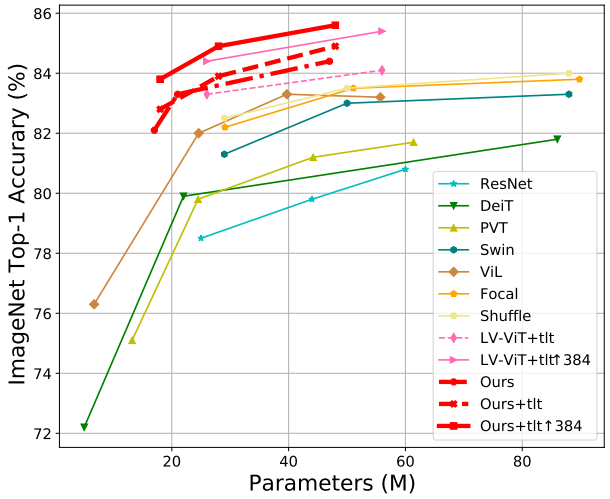

Fig. 5: Performance comparisons with respect to model parameters on ImageNet-1K classification. Without extra training data, our DilateFormer variants achieve comparable or even better performance with fewer model parameters.

object detection and instance segmentation, and ADE20K [110] 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-1K

- Dataset and implementation details. ImageNet-1k [19] 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-1K dataset to evaluate our variants, following the same training strategies of baseline Transformers as DeiT [73] and PVT [77] for a fair comparison. We use the AdamW optimizer [57] 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 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 [39] 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 [39] where CutMix [103] and Mixup [100] are replaced by MixToken [39]. 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.

- Results and analysis. As shown in Table III, Figure 1 and Figure 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-1K with a resolution of 224, surpasses Swin-T [54], ViL-S [104] by 2.0% and 1.3% respectively and has fewer parameters and FLOPs than these models. With the assistance of Token Labeling [39] (denoted by '⋆') , our models achieve better performance than LV-ViTs [39] at different model sizes, *i.e.*, Dilate-S⋆ (4.9 GFLOPs) and Dilate-B⋆ (10.0 GFLOPs) achieve 83.9% and

TABLE III: COMPARISON WITH THE STATE-OF-THE-ART ON IMAGENET-1K. '⋆' INDICATES TOKEN LABELING PROPOSED IN LV-VIT [39], AND '↑' INDICATES THAT THE MODEL IS FINE-TUNED AT A LARGER RESOLUTIONS.

Method	Params (M)	FLOPs (G)	Train	Test	Top1 (%)
RegNetY-4G [63]	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 [77]	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 [80]	20	4.5	224	224	81.6
GG-T [94]	28	4.5	224	224	82.0
DeiT-S [73]	22	4.6	224	224	79.9
Distilled DeiT-S [73]	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 [69]	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 ↑ 384 [80]	20	16.3	224	384	83.0
Dilate-T⋆ ↑ 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 [63]	84	16.0	224	224	82.9
Focal-T [91]	29	4.9	224	224	82.2
CrossFormer-S 78	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 [80]	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 [77]	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↑ 384 [18]	20	13.4	224	384	83.9
T2T-14 ↑ 384 [97]	22	17.1	224	384	83.3
LV-ViT-S⋆ ↑ 384 [39]	26	22.2	224	384	84.4
CvT-21 ↑ 384 [80]	32	24.9	224	384	83.3
Dilate-S⋆ ↑ 384 (ours)	22	15.5	224	384	**84.9**
ResNet-152 [35]	60	11.6	224	224	80.8
EffNet-B7 [72]	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 [91]	90	16.0	224	224	83.8
LV-ViT-M⋆ [39]	56	16.0	224	224	84.1
CrossFormer-L [78]	92	16.1	224	224	84.0
MPViT-B [42]	75	16.4	224	224	84.3
DeiT-B [73]	86	17.5	224	224	83.4
Distilled DeiT-B [73]	86	17.5	224	224	81.8
NesT-B [106]	68	17.9	224	224	83.8
BoTNet-T7 [69]	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 ↑ 384 [18]	42	27.4	224	384	85.1
LV-ViT-M⋆ ↑ 384 [39]	56	42.2	224	384	85.4
BoTNet-S1-128↑ 384 [69]	79	45.8	256	384	84.7
Dilate-B⋆ ↑ 384 (ours)	48	31.1	224	384	**85.6**

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

Method	Params (M)	FLOPs (G)	Mask R-CNN 1× schedule						Mask R-CNN 3× + MS schedule					
			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 [91]	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 [83]	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 [77]	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 [87]	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 [83]	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 [91]	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 [77]	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 [91]	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

84.9% respectively, surpassing LV-ViT-S [39](6.6 GFLOPs) and LV-ViT-M [39] (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 [104] (4.9G, 82.0%), Focal-T [91] (4.9G, 82.2%) and PVT-L [77] (9.8G, 81.7%). Similar conclusions can be found in larger models: our Dilate-S (83.3%) with 4.8 GFLOPs outperforms ViL-B [104] (13.4G, 83.2%), Swin-B [54] (15.4G, 83.4%), and DeiT-B [73] (17.5G, 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 [39] (85.4%) which needs 1.37 times more FLOPs.

B. Object Detection and Instance Segmentation on COCO

- Dataset and implementation details. We evaluate our variants on object detection and instance segmentation on COCO2017 dataset [50]. COCO2017 dataset contains 118K images for training, 5K images for validation and 20K images for testing. We utilize two representative frameworks: Mask R-CNN [34] and Cascade Mask R-CNN [4] implemented in mmdetection [8] and adopt the ImageNet-1K 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 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 [54] and global attention in stage3 of DilateFormer to reduce computational cost.

- Results and analysis. Table IV and Table V report box mAP (AP^b) and mask mAP (AP^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 [54] 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 [77] by 2.2% box mAP, 2.7% mask mAP at 1× schedule with 13.2% fewer FLOPs.

C. Semantic Segmentation on ADE20K

- Dataset and implementation details. ADE20K dataset [110] 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 ADE20K and utilize two representative frameworks: Upernet [84] and Semantic FPN [40] implemented in mmsegmentation [16] with our ImageNet-1K pre-trained variants as backbones. For training Upernet, we follow the configuration of Swin Transformer and train our variants for 160K iterations. We employ the AdamW [57] 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 80K 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.

- 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 [54] by at least 2.6% of mIoU and

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

Method	Params (M)	FLOPs (G)	3× + MS schedule					
			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 [73]	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 [87]	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**

TABLE VI: SEMANTIC SEGMENTATION EXPERIMENTAL RESULTS ON ADE20K VALIDATION SET. LEFT: WITH UPERNET; RIGHT: WITH SEMANTIC FPN.

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 [83]	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 [91]	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 [83]	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 [91]	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 [77]	28	161	39.8
TwinsP-S [13]	28	162	44.3
XCiT-S12/8 [1]	30	-	44.2
TRT-ViT-B [83]	46	176	45.4
Swin-T [54]	32	182	41.5
Next-ViT-S [45]	36	208	46.5
CrossFormer-S [78]	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 [77]	65	283	42.1
TwinsP-L [13]	65	283	46.4
TRT-ViT-D [83]	106	296	46.7
CrossFormer-B [78]	56	320	48.0
Swin-B [54]	91	422	46.0
Dilate-B (ours)	51	288	**48.8**

1.0% of MS mIoU respectively. For Semantic FPN framework, our variants achieve 47.1/48.8% mIoU, and exceed Swin [54] 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.

- SWDA vs. other sparse and local patterns. We replace Sliding Window Dilated Attention (SWDA) in the first two stages with other sparse and local patterns, *i.e.*, Dilated Convolution (DC) [92], Dynamic Dilated Convolution (DDC) [11] , Window Attention with Spatial Shuffle (WASS[2]) [38] and Sliding Window Attention (SWA) [43].

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-1K, 44.9% box mAP/40.9% mask mAP on COCO and 45.84% mIoU on ADE20K. 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-

[2]The WASS is an approximate sparse sampling operation which divides patches into local Windows like Swin [54] and then shuffles keys and values between different windows.

TABLE VII: EXPERIMENT RESULTS WITH LOCAL AND SPARSE PATTERNS IN THE FIRST TWO STAGES. THE Top-1 IS ON IMAGENET-1K, AP^b AND AP^m ARE ON COCO VAL2017 WITH MASK R-CNN 1× SCHEDULE, mIoU IS ON ADE20K VALIDATION SET WITH SEMANTIC FPN.

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**

TABLE VIII: TOP-1 ACCURACY ON IMAGENET-1K 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

TABLE IX: EXPERIMENT RESULTS WITH DIFFERENT BLOCKS IN STAGE1/2. "sr." INDICATES SPATIAL REDUCTION OPERATION. "D", "G" AND "L" INDICATE DILATION, GLOBAL AND LOCAL OPERATIONS, RESPECTIVELY. THE Top-1 IS ON IMAGENET-1K, AP^b AND AP^m ARE ON COCO VAL2017 WITH MASK R-CNN 1 × SCHEDULE, mIoU IS ON ADE20K VALIDATION SET WITH SEMANTIC FPN.

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**

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

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

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.

- 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 the effects of dilation scales according to the performance on ImageNet-1K 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.

- 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) [54] and global attention (G-G-G-G) [22] 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.) [77], 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) [54] (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) [22], 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 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.

- Stage setting. To demonstrate the modeling capability of the MSDA block at shallow stages, we conduct a set of experiments

TABLE XI: Top-1 ACCURACY ON IMAGENET-1K 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**

TABLE XII: COMPARISON OF MODEL INFERENCE. "MEM" DENOTES THE PEAK MEMORY FOR EVALUATION. "FPS" IS THE NUMBER OF IMAGES PROCESSED FOR ONE SECOND.

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**

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 (+ 0.35G). Therefore, we use MSDA blocks in both stage1 and stage2 by default.

- 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.

- 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.

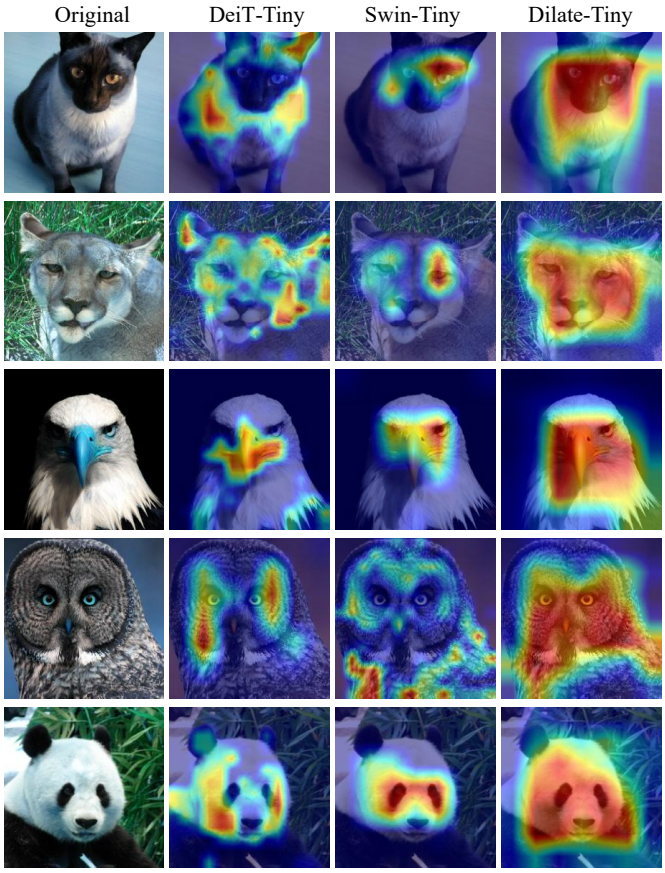

Fig. 6: Grad-CAM Visualization of the last layer of DeiT-Tiny, Swin-Tiny and Dilate-Tiny. Images are from the validation set of ImageNet-1k.

- Grad-CAM Visualization. To further illustrate the recognition ability of the proposed DilateFormer, we apply Grad-CAM [25] to visualize the areas of the greatest concern in the last layer of DeiT-Tiny [73], Swin-Tiny [54] and Dilate-Tiny. As shown in Figure 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.

- More Visualization Results on Global Attention. In Sec.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 [22]. As shown in Figure 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.

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

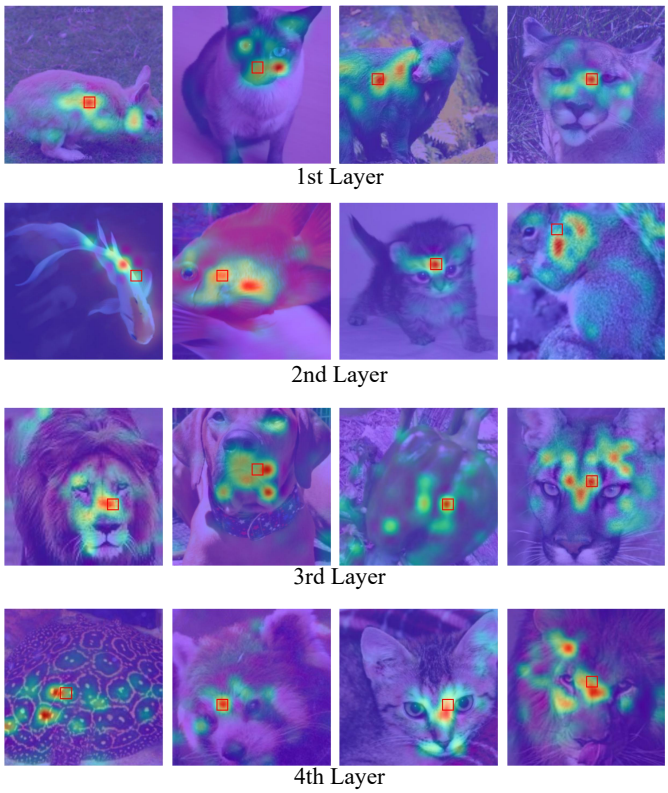

Fig. 7: More Visualization of attention maps of shallow layers of ViT-Small. We visualize the activations in attention maps of the query patches (in the red box). The attention maps show that patches with high attention scores sparsely scatter around the query patch, and other patches have low attention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

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-1k classification and down-stream vision tasks such as object detection and semantic segmentation.

ACKNOWLEDGMENTS

This work was supported partially by the NSFC (U21A20471,U1911401,U1811461), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), and the Key-Area Research and Development Program of Guangzhou (202007030004).

REFERENCES

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. XCiT: Cross-Covariance Image Transformers. In *Annual Conference on Neural Information Processing Systems*, 2021.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150, 2020.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Annual Conference on Neural Information Processing Systems*, 2020.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Computer Vision 16th European Conference, Glasgow*, Lecture Notes in Computer Science. Springer, 2020.

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

[7] Jun Chen, Xuejiao Li, Linbo Luo, and Jiayi Ma. Multi-focus image fusion based on multi-scale gradients and image matting. *IEEE Trans. Multim.*, 24:655–667, 2022.

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab Detection Toolbox and Benchmark. *CoRR*, abs/1906.07155, 2019.

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR*, abs/1706.05587, 2017.

[10] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-Former: Bridging MobileNet and Transformer. *CoRR*, abs/2108.05895, 2021.

[11] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[12] Ying Chen, Shixiong Xia, Jiaqi Zhao, Yong Zhou, Qiang Niu, Rui Yao, Dongjun Zhu, and Dongjingdian Liu. ResT-ReID: Transformer block-based residual learning for person re-identification. *Pattern Recognition Letters*, 2022.

[13] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. In *Annual Conference on Neural Information Processing Systems*, 2021.

[14] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do We Really Need Explicit Position Encodings for Vision Transformers? *CoRR*, abs/2102.10882, 2021.

[15] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *Proceedings of the 33nd International Conference on Machine Learning*, 2016.

[16] MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[17] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic Head: Unifying Object Detection Heads With Attentions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[18] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying Convolution and Attention for All Data Sizes. *Annual Conference on Neural Information Processing Systems*, 2021.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[21] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *CoRR*, abs/2107.00652, 2021.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani,

Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*, 2021.

[23] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[24] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast Convergence of DETR with Spatially Modulated Co-Attention. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.

[25] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[26] Xun Gong, Zu Yao, Xin Li, Yueqiao Fan, Bin Luo, Jianfeng Fan, and Boji Lao. Lag-net: Multi-granularity network for person re-identification via local attention system. *IEEE Trans. Multim.*, 24:217–229, 2022.

[27] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. CMT: Convolutional Neural Networks Meet Vision Transformers. *CoRR*, abs/2107.06263, 2021.

[28] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. SOTR: Segmenting Objects with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.

[29] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021.

[30] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the Connection between Local Attention and Dynamic Depth-wise Convolution. In *International Conference on Learning Representations*, 2022.

[31] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. 2022.

[32] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood Attention Transformer. *CoRR*, abs/2204.07143, 2022.

[33] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked Autoencoders Are Scalable Vision Learners. *CoRR*, abs/2111.06377, 2021.

[34] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, 2017.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.

[37] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment Your Batch: Improving Generalization Through Instance Repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[38] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. *CoRR*, abs/2106.03650, 2021.

[39] Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All Tokens Matter: Token Labeling for Training Better Vision Transformers. In *Annual Conference on Neural Information Processing Systems*, 2021.

[40] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[41] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *26th Annual Conference on Neural Information Processing Systems 2012.*, 2012.

[42] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. MPViT: Multi-Path Vision Transformer for Dense Prediction. *CoRR*, abs/2112.11010, 2021.

[43] Gang Li, Di Xu, Xing Cheng, Lingyu Si, and Changwen Zheng. Simvit: Exploring a simple vision transformer with sliding windows, 2021.

[44] Junxia Li, Zefeng Pan, Qingshan Liu, and Ziyang Wang. Stacked u-shape network with channel-wise attention for salient object detection. *IEEE Trans. Multim.*, 23:1397–1409, 2021.

[45] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.

[46] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. UniFormer: Unifying Convolution and Self-attention for Visual Recognition. *CoRR*, abs/2201.09450, 2022.

[47] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual Transformer Networks for Visual Recognition. *CoRR*, abs/2107.12292, 2021.

[48] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Trans. Multim.*, 24:1922–1932, 2022.

[49] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision*, 2017.

[50] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision 13th European Conference*, 2014.

[51] Huabin Liu, Jianguo Li, Dian Li, John See, and Weiyao Lin. Learning scale-consistent attention part network for fine-grained image recognition. *IEEE Trans. Multim.*, 24:2902–2913, 2022.

[52] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *Computer Vision 14th European Conference, Amsterdam*, 2016.

[53] Yang Liu, Faming Fang, Tingting Wang, Juncheng Li, Yun Sheng, and Guixu Zhang. Multi-scale grid network for image deblurring with high-frequency guidance. *IEEE Trans. Multim.*, 24:2890–2901, 2022.

[54] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.

[55] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *CoRR*, abs/2201.03545, 2022.

[56] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[57] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations*, 2019.

[58] Xu Ma, Jingda Guo, Andrew Sansom, Mara McGuire, Andrew Kalaani, Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Spatial pyramid attention for deep convolutional neural networks. *IEEE Trans. Multim.*, 23:3048–3058, 2021.

[59] Yu-Jen Ma, Hong-Han Shuai, and Wen-Huang Cheng. Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation. *IEEE Trans. Multim.*, 24:261–273, 2022.

[60] Yi-Xing Peng, Jile Jiao, Xuetao Feng, and Wei-Shi Zheng. Consistent discrepancy learning for intra-camera supervised person re-identification. *IEEE Transactions on Multimedia*, 2022.

[61] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local Features Coupling Global Representations for Visual Recognition. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.

[62] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper.pdf*, 2018.

[63] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[64] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767, 2018.

[65] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Annual Conference on Neural Information Processing Systems*, 2015.

[66] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted Self-Attention via Multi-Scale Token Aggregation. *CoRR*, abs/2111.15193, 2021.

[67] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention 18th International Conference Munich*, 2015.

[68] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations*, 2015.

[69] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck Transformers for Visual Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[70] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *IEEE*

Conference on Computer Vision and Pattern Recognition, 2015.
[71] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[72] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International conference on machine learning*. PMLR.
[73] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.
[74] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with Image Transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[75] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022.
[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Annual Conference on Neural Information Processing Systems*, 2017.
[77] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[78] Wenxiao Wang, Lu Yao, Long Chen, Deng Cai, Xiaofei He, and Wei Liu. CrossFormer: A Versatile Vision Transformer Based on Cross-scale Attention. *CoRR*, abs/2108.00154, 2021.
[79] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-Local Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[80] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
[81] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and Improving Relative Position Encoding for Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[82] Rongjie Xia, Yanshan Li, and Wenhan Luo. Laga-net: Local-and-global attention network for skeleton based action recognition. *IEEE Trans. Multim.*, 24:2648–2661, 2022.
[83] Xin Xia, Jiashi Li, Jie Wu, Xing Wang, Mingkai Wang, Xuefeng Xiao, Min Zheng, and Rui Wang. Trt-vit: Tensorrt-oriented vision transformer. *arXiv preprint arXiv:2205.09579*, 2022.
[84] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *Computer Vision 15th European Conference*, 2018.
[85] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross B. Girshick. Early Convolutions Help Transformers See Better. In *Annual Conference on Neural Information Processing Systems*, 2021.
[86] Pan Xie, Mengyi Zhao, and Xiaohui Hu. Pisltrc: Position-informed sign language transformer with content-aware convolution. *IEEE Trans. Multim.*, 24:3908–3919, 2022.
[87] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[88] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. *CoRR*, abs/2111.09886, 2021.
[89] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 2021.
[90] Zhaoyi Yan, Ruimao Zhang, Hongzhi Zhang, Qingfu Zhang, and Wangmeng Zuo. Crowd counting via perspective-guided fractional-dilation convolution. *IEEE Trans. Multim.*, 24:2633–2647, 2022.
[91] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. *CoRR*, abs/2107.00641, 2021.
[92] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *4th International Conference on Learning Representations*, 2016.
[93] Litao Yu, Jian Zhang, and Qiang Wu. Dual attention on pyramid feature maps for image captioning. *IEEE Trans. Multim.*, 24:1775–1786, 2022.
[94] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 2021.
[95] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
[96] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating Convolution Designs into Visual Transformers. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[97] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training Vision Transformers from scratch on Imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
[98] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. VOLO: Vision Outlooker for Visual Recognition. *CoRR*, abs/2106.13112, 2021.
[99] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip H. S. Torr, Wayne Zhang, and Dahua Lin. Vision Transformer with Progressive Sampling. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[100] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision*, 2019.
[101] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers. *CoRR*, abs/2106.04560, 2021.
[102] Congxuan Zhang, Zhongkai Zhou, Zhen Chen, Weiming Hu, Ming Li, and Shaofeng Jiang. Self-attention-based multiscale feature learning optical flow with occlusion feature map prediction. *IEEE Trans. Multim.*, 24:3340–3354, 2022.
[103] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations*, 2018.
[104] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In *2021 IEEE/CVF International Conference on Computer Vision*, 2021.
[105] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, , Sercan Ö. Arık, and Tomas Pfister. Nested Hierarchical Transformer: Towards Accurate, Data-Efficient and Interpretable Visual Understanding. In *AAAI Conference on Artificial Intelligence*, 2022.
[106] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating Nested Transformers. *arXiv preprint arXiv:2105.12723*, 2021.
[107] Baoxin Zhao, Haoyi Xiong, Jiang Bian, Zhishan Guo, Cheng-Zhong Xu, and Dejing Dou. COMO: efficient deep neural networks expansion with convolutional maxout. *IEEE Trans. Multim.*, 23:1722–1730, 2021.
[108] Yanwei Zheng, Zengrui Zhao, Xiaowei Yu, and Dongxiao Yu. Template-Aware Transformer for Person Reidentification. *Computational Intelligence and Neuroscience*, 2022.
[109] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. In *AAAI Conference on Artificial Intelligence*, 2020.
[110] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[111] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. DeepViT: Towards Deeper Vision Transformer. *CoRR*, abs/2103.11886, 2021.
[112] Lei Zhou, Chen Gong, Zhi Liu, and Keren Fu. SAL: selection and attention losses for weakly supervised semantic segmentation. *IEEE Trans. Multim.*, 23:1035–1048, 2021.
[113] Fangrui Zhu, Yi Zhu, Li Zhang, Chongruo Wu, Yanwei Fu, and Mu Li. A Unified Efficient Pyramid Transformer for Semantic Segmentation. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2021.
[114] Yifan Zuo, Hao Wang, Yuming Fang, Xiaoshui Huang, Xiwu Shang, and Qiang Wu. Mig-net: Multi-scale network alternatively guided by intensity and gradient features for depth map super-resolution. *IEEE Trans. Multim.*, 24:3506–3519, 2022.

Supplementary Material

More ablation studies

-Minimum Kernel size. The minimum kernel size affects the size of the attended receptive field of our Sliding Window Dilated Attention operation and naturally affects the performance of our DilateFormer. We analyze the effect of different minimum kernel sizes on the ImageNet-1K classification task. The experimental results in terms of FLOPs and Top-1 accuracy are shown in Table XIII. Experiments show that using a larger kernel size in SWDA can slightly improve the performance due to a larger attended receptive field, but the FLOPs will also increase. By default, we set the kernel size in SWDA to 3×3.

TABLE XIII: Top-1 accuracy on ImageNet-1K of different minimum kernel sizes.

Minimum Kernel Size	FLOPs (G)	Top-1 (%)
3×3	3.18	82.1
5×5	3.21	82.2
7×7	3.24	82.3