2023 International Conference on Wireless Communications and Signal Processing (WCSP) | 979-8-3503-2466-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/WCSP58612.2023.10404854

DNAT: Multi-scale Transformer with Dilated Neighborhood Attention for Image Classification

Chenfeng Jiang[1,2], Quan Zhou[1,2,*], Zhiyi Mo[2], Jing Wang[1], Yawen Fan[1], Xiaofu Wu[1], Suofei Zhang[3], Bin Kang[3],
[1]National Engineering Research Center of Communications and Networking,
Nanjing University of Posts & Telecommunications, P.R. China.
[2]Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software,
Wuzhou University, Wuzhou, China.
[3]Department of Internet of Things, Nanjing University of Posts & Telecommunications, P.R. China.
1221013840@njupt.edu.cn, quan.zhou@njupt.edu.cn*

***Abstract*—Recently, Transformer-based models have achieved remarkable progress for image classification. In order to reduce the computational costs of multi-head self-attention (MHSA) that is widely adopted in recently-proposed visual Transformers, some window-based Transformers employ local self-attention via shrinking number of input tokens. Nevertheless, these approaches to some extent are short to model long-range dependencies. To address this problem, this paper presents an effective and powerful Transformer backbone, called Dilated Neighborhood Attention Transformer (DNAT) that has a mutli-path architecture, for image classification. DNAT is able to explore long-range interactions through dilated neighborhood attention (DNA). When dilation rate is small, DNAT degenerates to local self-attention that is computationally efficient to capture local features. As the dilation rate increases, more and more long-range token interactions are explored step-by-step. Finally, by integrating weighted features derived from various dilation rates, DNAT seamlessly harvests local-to-global context, yet with acceptable growth of model size and computational budgets. The extensive experiments on ImageNet-1K and CIFAR100 datasets demonstrate the effectiveness of DNAT for image classification.**

***Index Terms*—Image Classification, visual Transformer, Self-attention, Dilated Neighborhood Attention**

I. Introduction

Recent years have witnessed remarkable progress of image classification using convolutional neural networks (CNNs), such as VGG [1], GoogleNet [2], and ResNet [3]. Although CNNs are powerful to encode local features using convolutions, they are short to capture global dependencies. Instead, due to powerful capability to formulate long-range interactions using multi-head self-attention (MHSA), Vision Transformers (ViTs) [4], [5] have become dominant for image classification [5], [6] in recent years. As a pioneer work, ViT [5] keeps feature resolution unchanged to achieve excellent performance using large-scale pre-training. Since then, a huge amount of Transformer-based approaches and their variants [7], [8] are springing up like mushrooms in the community of computer vision. Nevertheless, the computation of global-based MHSA requires a lager number of computational resources, as the computational costs quadratically increase along with the input image resolutions. Therefore, recent works often adopt hierarchical Transformer-based backbones [9]–[11], where feature resolutions are gradually reduced. Most recently, window-based Transformers [9], [13]–[15] have also been proposed,

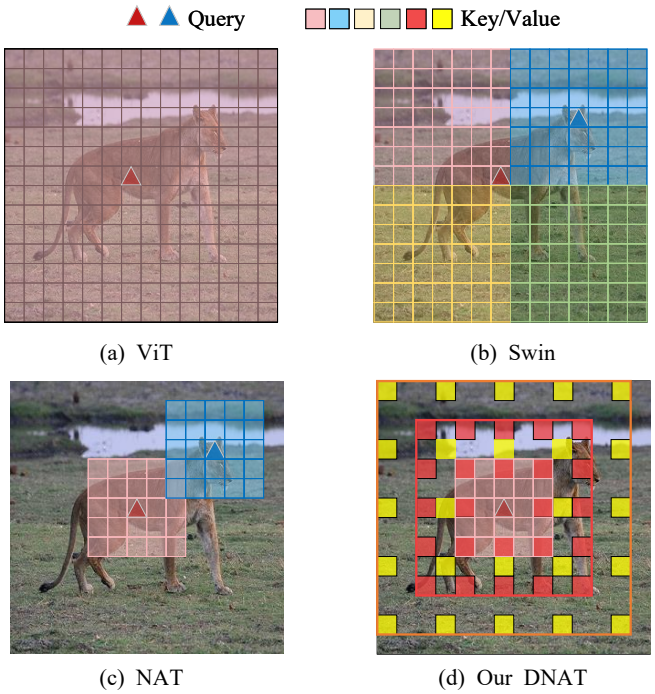

Fig. 1: Comparison of (a) ViT [5], (b) Swin [9], (c) NAT [12], and (d) DNAT. ViT captures global context using MHSA, yet Swin restricts MHSA in local divided windows. NAT encodes interactions between center token and its neighbourhoods in local range. DNAT employs multi-scale dependencies among surroundings with different dilation rates. The query tokens are marked by triangles, while key and value tokes are represented by squares with different colors. In (d), neighbouring tokens with different dilation rates are denoted by squares with pink, red, and yellow colors, respectively. (Best viewed in color)

Corresponding author: Quan Zhou, quan.zhou@njupt.edu.cn. This work is partly supported by NSFC (No. 61876093), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX22_0962), and Guangxi Colleges and Universities Key Laboratory of Intelligent Industry Software (No. 2023B01)

restricting the calculation of MHSA in divided local windows. Moreover, the computational costs are reduced from quadratic to linear, alleviating a lot of computing burdens. However, they also accompany some limitations, particularly the one that most window-based attention independently focuses on local token interactions within divided windows, ignoring to inherit the merit of MHSA that is able to capture dense and global token-based dependencies.

To alleviate this problem, this paper designs a novel Transformer ackbone, called dilated neighborhood attention Transformer (DNAT), for image classification. As shown in Fig. 1 (d), DNAT employs a dilated neighborhood attention (DNA) to encode token-based interactions from different scales. It is worthy that DNAT integrates local-to-global context features step-by-step, rather than only employing one scale neighboring information [12] that is too weak to investigate global context. When dilation rate is small, DNAT degenerates to NAT [12], indicating NAT is a special case of our DNAT. The detail structure of DNA is also motivated from dilated convolution that investigates multi-scale information via convolution with various dilation rates. Nevertheless, DNAT achieves more powerful representation capabilities from multi-path perspective, where each path encodes various-scale contextual features from long-range surroundings instead of local neighborhoods. Most importantly, DNAT is also computationally efficient, as the learned model parameters are shared within each path. In nutshell, the major contributions of our paper are three-fold:

- Instead of using single-scale features [5], [12], a multi-path architecture is adopted in DNA to encode feature representation from multiple scales. In addition, the attention calculation requires no further parameters, making it computationally efficient for image classification.
- Our DNA inherits the merits of NAT [12] and dilated convolution together. On one hand, limiting interactions between center query and very few surroundings is beneficial for speeding up the computation of MHSA. The dilated operations, on the other hand, are helpful to enlarge receptive field from local to global step-by-step.
- To evaluate DANT, extensive experiments have been conducted on two widely-used image classification datasets: ImageNet-1K [16] and CIFAR100 [17]. The exhausted experimental results demonstrate the effectiveness of our method. Specifically, DNAT achieves 82.7/83.2% Top-1 accuracy on two datasets, with only 23M model size, 4.6/2.4GFLOPs, and 482/1,672 Throughput, respectively.

The remainder of this paper is organized as follows. Section II briefly reviews the related works. The detailed architecture of DNAT is introduced in Section III. Section IV shows the experimental settings and results on all datasets. Finally, the concluding remarks and future works are given in Section V.

II. Related Work

This section reviews related works from two perspectives: single-scale [5], [12], [14], [18], and multi-scale interactions [10], [11], [19] for image classification, respectively.

A. Transformers based on Single-scale Interactions

At present, many ViTs and their variants [5], [12], [18], [20] are proposed for image classification using single-scale token-based interactions. The most representative one is ViT [5], where images are represented by a series of non-overlapped 16×16 patches (also known as tokens), and token-to-token based interactions are investigated to capture global context using MHSA. Instead of computing MHSA based on image patches, BOAT [14] advocates to explore token interactions in feature space, where MHSA is defined on the tokens that have most similar features, without considering their spatial locations. SepViT [18] restricts attention computation in local windows, and establishes window connections using window tokens. NAT [12] computes MHSA between each query and its nearest neighborhoods, which is essentially equivalent to the single-scale window Transformer. An alternative method to explore single-scale interactions is aggregating neighboring tokens into a larger one [21], gradually reducing number of input tokens. Unlike these approaches that employ pure Transformer backbones, Slide Transformer [20] is a CNN-Transformer hybrid that introduces a deformed shifting module based on the re-parameterization technique. Although these advanced ViTs have achieved impressive classification results, using single-scale interactions is evidently not enough to capture long-range, even global dependencies. In contrast, DNAT encodes local-to-global context in terms of multi-path manner, where each path investigates various scale interactions.

B. Transformers based on Multi-scale Interactions

The multi-scale interactions have gained great attraction in recent ViTs [9], [10], [22], [23], which can be roughly divided into three categories: window-based ViTs [9], [13], token aggregation [22] and partition [6], and feature pyramid [10], [11], [24]. In first category, the most representative works are Swin Transformer families [9], [15]. For instance, Swin [9], as one of the earliest window-based ViTs, evenly divides input image into non-overlapped windows. Thereafter, it computes MHSA in these independent windows. Finally, a shifted window mechanism is utilized to recover window connections using mask operations. VSA [13] employs a window regression module to predict the size and location of target window. The second category explores multi-scale interactions by aggregating or dividing tokens. Shunted Transformer [22] selectively merges tokens to produce larger ones. Conversely, TNT [6] considers 16×16 tokens are too large to extract detail features. It thus partitions image patches into smaller 4×4 ones. The final category produces hierarchy feature pyramid or uses image pyramid to investigate multi-scale clues. For example, MVT [24] embeds feature pyramid into ViTs through channel expansion. PVT [10] directly introduces feature pyramid for multi-scale interactions. P2T [11] adopts an adaptive pyramid pooling to MHSA, capturing powerful contextual features. Instead, CorssFormer [23] projects image pyramid into feature embeddings and concatenates them to encode multi-scale cues. The most similar work to DNAT is DiNAT [19], where each Transformer block adopts different

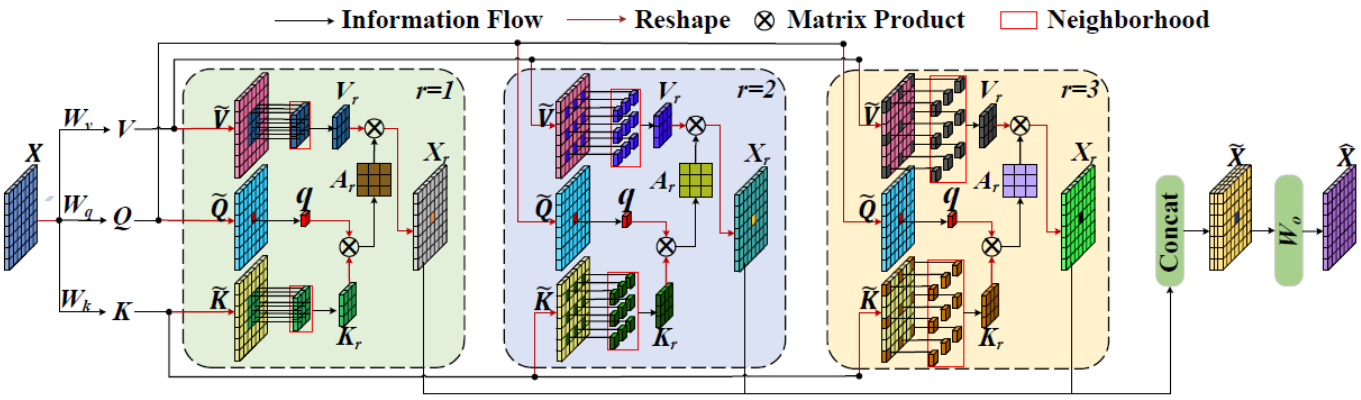

Fig. 2: An illustration of DNA. DNA has a multi-path architecture, where each path employs MHSA with a specified dilation rate, corresponding to various ranges of neighbor tokens that are ready to interact with queries. (Best viewed in color)

dilation rate to explore single-scale interactions. Nevertheless, DNAT achieves more powerful representation capabilities, where each DNAT block utilizes a multi-path architecture with different dilation rates to investigate multi-scale interactions.

III. Our Method

This section first introduces DNA, used to extract various scale interactions via multi-path dilated attention. Thereafter, we elaborate on the detailed architecture of DNAT.

A. Dilated Neighborhood Attention (DNA)

As illustrated in Fig. 2, to effectively capture multi-scale information, a parallel multi-path architecture is utilized in our DNA, where each path shares the same paradigm of MHSA computation, yet with different dilation rates. Immediately below, we introduce the details of how to compute DNA.

Let $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ be input features, where H, W and C stand for the height, width, and channel number of $\boldsymbol{X}$, respectively. Following traditional ViTs [5], [9], [20], three linear projections $\boldsymbol{W}_q \in \mathbb{R}^{C\times C}$, $\boldsymbol{W}_k \in \mathbb{R}^{C\times C}$, and $\boldsymbol{W}_v \in \mathbb{R}^{C\times C}$ are firstly used to map $\boldsymbol{X}$ into query $\boldsymbol{Q} \in \mathbb{R}^{HW\times C}$, key $\boldsymbol{K} \in \mathbb{R}^{HW\times C}$, and value $\boldsymbol{V} \in \mathbb{R}^{HW\times C}$, respectively:

$$\boldsymbol{Q} = \mathrm{Re}(\boldsymbol{X})\boldsymbol{W}_q, \quad \boldsymbol{K} = \mathrm{Re}(\boldsymbol{X})\boldsymbol{W}_k, \quad \boldsymbol{V} = \mathrm{Re}(\boldsymbol{X})\boldsymbol{W}v \tag{1}$$

where $\mathrm{Re}(\cdot)$ denotes reshape operation. Note $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are shared for each path to save model size, and the following computation of DNA does not require any further parameters.

For better understanding how DNA works, as shown in Fig. 2, we once again reshape $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ into three-dimensional tensors $\widetilde{\boldsymbol{Q}}$, $\widetilde{\boldsymbol{K}}$, and $\widetilde{\boldsymbol{V}}$ that have the same shape of input $\boldsymbol{X}$:

$$\widetilde{\boldsymbol{Q}} = \mathrm{Re}(\boldsymbol{Q}), \quad \widetilde{\boldsymbol{K}} = \mathrm{Re}(\boldsymbol{K}), \quad \widetilde{\boldsymbol{V}} = \mathrm{Re}(\boldsymbol{V}) \tag{2}$$

Given one specific query $\boldsymbol{q} \in \mathbb{R}^{1\times 1\times C}$ at position (x, y), we define $\mathcal{N}_r(x, y)$ as the neighborhood of $\boldsymbol{q}$ in r^{th} path, which has a $n \times n$ fixed-length set of indices of tokens nearest to (x, y), yet have different dilation rate r. Note expanding the size of $\mathcal{N}_r(x, y)$ that covers all possible positions of $\boldsymbol{X}$ will result in the fact that neighborhood attention equals to self-attention, as all tokens are treated as neighbors and involved in self-attention computation. Collecting and reshaping all neighborhood tokens in $\widetilde{\boldsymbol{K}}$ and $\widetilde{\boldsymbol{V}}$ by $\mathcal{N}_r(x, y)$ produces new key $\boldsymbol{K}_r \in \mathbb{R}^{n^2\times C}$ and value $\boldsymbol{V}_r \in \mathbb{R}^{n^2\times C}$, where $n << \min\{W, H\}$. Thereafter, an attention map $\boldsymbol{A}_r \in \mathbb{R}^{1\times n^2}$ that represents the dependencies between query $\boldsymbol{q}$ and its neighborhood $\boldsymbol{K}_r$ is produced using a SoftMax$(\cdot)$ function:

$$\boldsymbol{A}_r = \mathrm{SoftMax}(\frac{\mathrm{Re}(\boldsymbol{q})\boldsymbol{K}_r^\top}{\sqrt{d}}) \tag{3}$$

where $\sqrt{d}$ is a scaling parameter. Finally, the output intermediate feature $\boldsymbol{x}_r \in \mathbb{R}^{1\times 1\times C}$ in position (x, y) is produced by reweighting $\boldsymbol{V}_r$ with attention map $\boldsymbol{A}_r$:

$$\boldsymbol{x}_r = \mathrm{Re}(\boldsymbol{A}_r\boldsymbol{V}_r) \tag{4}$$

This operation is repeated for every position, producing feature maps $\boldsymbol{X}_r \in \mathbb{R}^{W\times H\times C}$ that has the same shape of input $\boldsymbol{X}$. In our method, the sizes of neighborhood are default set as 3×3, 5×5, and 7×7, respectively, for three paths, which has been demonstrated by our ablation studies in Section IV-E2. Note when there is only one path, DNA degenerates to NAT [12], indicating NAT is a special and simplified case of our method.

So far, we have introduced how to independently compute DNA with different dilation rates. To recover relationship of all paths, we first produce an integrated feature $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{H\times W\times 3C}$ by concatenating all $\boldsymbol{X}_r$ together, and then fed $\widetilde{\boldsymbol{X}}$ into a linear projection $\boldsymbol{W}_o \in \mathbb{R}^{3C\times C}$, resulting in the final outputs $\widehat{\boldsymbol{X}} \in \mathbb{R}^{H\times W\times C}$ that have the same shape of input features $\boldsymbol{X}$.

B. Dilated Neighborhood Attention Transformer (DNAT)

As shown in Fig. 3, DNAT inherits the hierarchical architecture [9], [10] that has four stages, where feature resolutions are gradually reduced, and channel numbers are expanded by factor 2. Therefore, it produces a series of feature maps that have the shape of $\frac{H}{4}\times\frac{W}{4}\times C$, $\frac{H}{8}\times\frac{W}{8}\times 2C$, $\frac{H}{16}\times\frac{W}{16}\times 4C$, $\frac{H}{32}\times\frac{W}{32}\times 8C$, where H and W represent the width and height of input image, and C denotes channel numbers. Except first stage that contains patch embedding module [9], the rest stages are composed by patch merging module [10] and repeated

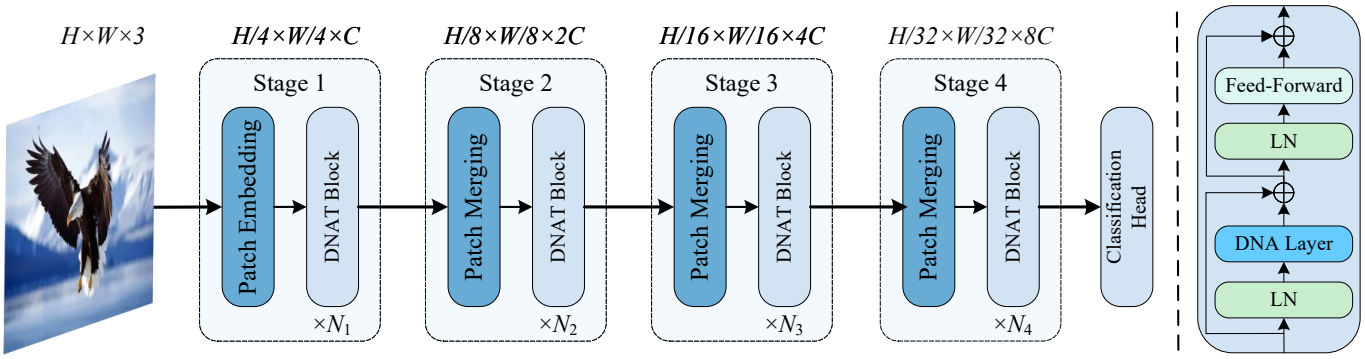

Fig. 3: Overview of the proposed Dilated Neighborhood Attention Transformer (DNAT). The architecture is divided into four stages, each of which consists of a patch embedding block or patch merging block and N_i repeated DNAT blocks. The right column is the detailed implementation of one DNAT block. (Best viewed in color)

TABLE I: The detailed architecture of backbone in DNAT. Conv$(k \times k, s)$ represents convolution using $k \times k$ filter kernel size with stride s, H is the number of heads in MHSA, r stands for dilation rate, and $n \times n$ denotes neighborhood $\mathcal{N}_r(x, y)$.

Stage	Input Size	Layer Name	Operator
1	$224 \times 224 \times 3$	Patch Embedding	Conv(3×3 , s=2) $\times$ 2
	$56 \times 56 \times 64$	DNAT Block $\times$ 3	$H = 2$ $3 \times 3, r = 1$ $5 \times 5, r = 2$ $7 \times 7, r = 3$
2	$56 \times 56 \times 64$	Patch Merging	Conv$(3 \times 3, s = 2)$
	$28 \times 28 \times 128$	DNAT Block $\times$ 4	$H = 4$ $3 \times 3, r = 1$ $5 \times 5, r = 2$ $7 \times 7, r = 3$
3	$28 \times 28 \times 128$	Patch Merging	Conv(3×3 , s=2)
	$14 \times 14 \times 256$	DNAT Block $\times$ 6	$H = 8$ $3 \times 3, r = 1$ $5 \times 5, r = 2$ $7 \times 7, r = 3$
4	$14 \times 14 \times 256$	Patch Merging	Conv$(3 \times 3, s = 2)$
	$7 \times 7 \times 512$	DNAT Block $\times$ 5	$H = 16$ $3 \times 3, r = 1$ $5 \times 5, r = 2$ $7 \times 7, r = 3$

DNAT blocks. As illustrated in the right panel of Fig. 3, the DNA layers are equipped with Feed-Forward Networks (FFN), residual connection, and Layer Normalization (LN), following traditional ViT-based design [5], [9].

The detailed structure of DNAT backbone is given in Table I. More specifically, the patch embedding module employs two 3×3 convolutions with stride 2, directly shrinking 4 times the input resolution, while patch merging module utilizes 3×3 convolution with stride 2, only downsampling 2 times of input size. Except input size and number of heads used to compute MHSA, each DNAT block shares the same architecture that has three paths to explore multi-scale information, where dilation rates are chosen as 1, 2, and 3, corresponding to the neighborhood $\mathcal{N}_r(x, y)$ of 3×3, 5×5, and 7×7, respectively.

IV. Experiments

To evaluate our DNAT, we have conducted exhausted experiments on ImageNet-1K [16] and CIFAR100 [17] datasets. In addition, a series of ablation studies have been carried on to reveal the potential impact of various components, and better understand the underlying behavior of DNAT.

A. Datasets and Evaluation Metrics

1) Datasets: **ImageNet-1K** [16] is the most popular dataset for image classification. It involves 1,000 categories with approximately 1,000 images per category. Specifically, this dataset contains 1.28 million training, and 50K testing images. In contrast, **CIFAR100** [17] is a smaller dataset that only includes 100 classes and 60K color images with resolution 32×32, which are divided into 50K/10K images for training and testing, respectively.

2) Evaluation metrics: For fair comparison with other state-of-the-art methods, we employ the standard evaluation metric of Top-1 recognition accuracy. On the other hand, the popular floating-point operations per second (FLOPs) and Throughput that evaluates how many images are recognized per second are both used to measure implementation efficiency.

B. Implementation Details

1) Training settings: DNAT is implemented in the hardware server platform with 8 RTX 3090 GPU cards. The software code is based on the MMPretrain toolbox that is an open-source repository for image classification. The widely-used AdamW is employed to optimize DNAT, where the weight decay and initial learning rate are set to 0.05 and 0.5×10^{-3} respectively. Otherwise, the cosine learning policy is adopted with the minimum learning rate 0.5×10^{-5}. We trained DNAT for 300 epochs on two datasets, where the first 20 epochs are used to warm up. To increase data diversity, we employed various data augmentation techniques, including

TABLE II: Comparison with state-of-the-art methods on ImageNet-1K dataset [16]. For fair comparison, the input resolutions of all methods are fixed to 224×224. '-' denotes that the results are not reported.

Method	Year	Param (M)	FLOPs (G)	Thru.	Top-1 (%)
CNN					
ResNet34 [3]	CVPR2016	21.8	3.6	**860**	74.8
ResNet50 [3]	CVPR2016	25.0	4.1	585	76.2
Transformer					
DeiT-S [25]	ICML2021	22.1	4.6	489	79.9
PVT-S [10]	ICCV2021	24.5	3.8	336	79.8
Swin-T [9]	ICCV2021	28.3	4.5	431	81.2
T2T-T [21]	ICCV2021	22.0	5.2	305	81.5
DPT-S [26]	ACMMM2021	26.0	4.0	-	81.0
Twins-S [27]	NIPS2021	24.1	3.8	439	81.2
DAT-T [28]	CVPR2022	29.0	4.6	-	82.0
CrossFormer [23]	ICLR2022	27.8	2.9	686	81.5
Slide-PVT-S [20]	CVPR2023	22.7	4.0	790	81.7
DiNAT-M [19]	ARXIV2022	**20.0**	**2.7**	520	81.8
NAT-M [12]	CVPR2023	**20.0**	**2.7**	533	81.8
DNAT	-	23.0	4.6	482	**82.7**

random cropping, random flipping, Mixup, CutMix, and random erasing [9]. The ablation studies are only conducted on the CIFAR100 [17] dataset.

2) Loss settings: Label Smoothing Loss (LSL) [9] is employed to supervise the entire DNAT. It combines cross entropy [3] and a label smoothness term that produces target labels by averaging the probability distribution of the true label with other classes:

$$\mathcal{L} = (1-\epsilon) \cdot \mathcal{L}_{CE} - \epsilon/N \sum \log y_{pred} \tag{5}$$

where $\mathcal{L}_{CE}$ is the cross entropy Loss, ϵ denotes non-negative smoothing parameter that leverage the trade-off between $\mathcal{L}_{CE}$ and smoothness term, N indicates the number of classes, and y_{pred} stands for the predicted probabilities.

C. Results on ImageNet-1K

Table II reports some quantitative results with recent state-of-the-art networks. In order to increase diversity of models, CNN-based [3] and ViT-based [9], [12], [20], [23], [25], [28] approaches are both adopted. Our DNAT achieves the best 82.7% Top-1 accuracy, yet with only 23M model size and 4.6 GFLOPs. Compared with CNN-based ResNet [3], DNAT has comparable model size (23M vs 21.8/25.0M) and a slight increment of GFLOPs (4.6G vs 3.6/4.1G), yet it delivers 7.9% and 6.5% improvements. When compared with ViT-based methods, on the other hand, DNAT surpasses PVT [10], Swin [9], and T2T [21] by large margins in terms of Top-1 accuracy (e.g., 2.9%, 1.5%, and 1.2%), and Throughput (e.g., 482 vs 336/431/305), respectively. Some ViT-based networks have smaller model size, GFLOPs, and Throughput (e.g., DeiT-S [25] and Twins-S [27]), but their performance have 2.8% and 1.5% drops. Particularly, compared with NAT [12] that is a simplified version of our method, DNAT achieves a significant improvement (82.7% vs 81.8%), yet with only a slight growth in the number of parameters (23M vs 20M).

TABLE III: Comparison with state-of-the-art methods on CIFAR100 [17]. '-' denotes that the results are not reported.

Model	Year	Param (M)	FLOPs (G)	Thru.	Top-1 (%)
CNN					
ResNet34 [3]	CVPR2016	21.3	1.8	-	76.7
SENet34 [29]	CVPR2018	21.6	-	-	77.9
SKNet-50 [30]	CVPR2019	27.7	4.2	778	82.6
Transformer					
DeiT-S [25]	ICML2021	21.4	1.4	1,800	66.5
PVT-S [10]	ICCV2021	27.0	**1.2**	1,270	69.8
Swin-T [9]	ICCV2021	27.5	1.5	2,399	78.1
NesT-S [31]	AAAI2022	23.4	6.6	628	81.7
NAT-M [12]	CVPR2023	20.0	1.8	**2,457**	82.7
CNN-Transformer Hybrid					
CvT-13 [32]	ICCV2021	**19.6**	4.5	-	81.8
DHVT-S [33]	NIPS2022	23.4	1.5	-	82.9
DNAT	-	23.0	2.4	1,672	**83.2**

TABLE IV: Ablation studies on path number r used in DNA.

Path number r	size of $\mathcal{N}_r(x,y)$				Params (M)	FLOPs (G)	Top-1(%)
	3×3	5×5	7×7	9×9			
$\{1\}$	✓				20.0	2.6	81.5
$\{1,2\}$	✓	✓			21.0	3.0	82.6
$\{1,2,3\}$	✓	✓	✓		23.0	4.6	83.2
$\{1,2,3,4\}$	✓	✓	✓	✓	25.0	9.6	**83.5**

D. Results on CIFAR100

Similar with ImageNet-1K dataset [16], besides CNN-based [3], [29], [30] and ViT-based [9], [10], [25], [31] networks, we also increase model diversity and compare with Hybrid-based [32], [33] methods. The results are reported in Table III. Regardless of which networks are utilized, our DNAT is able to consistently boost classification performance, yet with comparable even less model size, GFLOPs, and faster Throughput. Particularly, compared with SKNet-50 [30] and NesT-S [31], DNAT obtains 0.6% and 1.5% Top-1 accuracy improvement, with less model size (23.0M vs 27.7/23.4M), less FLOPs (2.4G vs 4.2/6.6G), and faster Throughput (1,672 vs 778/628). Some ViT-based networks, such as Dei-S [25] and NAT-M [12], require less model size and GFLOPs, and run faster than our DNAT, yet delivering 16.7% and 0.5% accuracy drops.

E. Ablation Study

To understand the underlying behavior of DNAT, this section reports some results of a series of ablation studies.

1) Ablation studies for Path number: The number of paths, also known as dilation rates, determines how many scales of information are explored, significantly influence the trade-off between classification accuracy and computational efficiency in DNA. We thus evaluate the performance variance along with

TABLE V: Ablation studies on the collection of neighborhood $\mathcal{N}_r(x, y)$, where $n \times n$ represents the size of $\mathcal{N}_r(x, y)$.

Path number r			Params (M)	FLOPs (G)	Top-1(%)
1	2	3			
3×3	5×5	3×3	23.0	3.2	82.6
3×3	5×5	5×5	23.0	3.7	82.9
3×3	5×5	7×7	23.0	4.6	**83.2**
3×3	7×7	3×3	23.0	3.6	82.8
3×3	7×7	5×5	23.0	3.8	83.0
3×3	7×7	7×7	23.0	4.8	82.7

the changes of path number r. Along with the increase of path numbers, the size of neighborhood $\mathcal{N}_r(x, y)$ also gradually becomes larger from 3×3 to 9×9. The results are reported in Table IV. DNAT obtains better results when more and more scales of information are introduced. Adopting four paths, however, requires more than $2 \times$ computational costs (9.6 vs 4.6 GFLOPs) yet with only 0.3% accuracy improvement. Thus the sizes of $\mathcal{N}_r(x, y)$ are chosen as 3×3, 5×5 and 7×7, respectively, for three paths in DNA.

2) Ablation studies for the size of neighborhood: The size of $\mathcal{N}_r(x, y)$ determines how many neighbors are collected to interact with center query. Thus this section evaluates the effect on performance by changing the size of $\mathcal{N}_r(x, y)$. However, it is impractical to enumerate all possible combinations of $\mathcal{N}_r(x, y)$ with $n \times n$ size using different dilation rates. We thus fix the number of path to three in DNA design, and change the size of $\mathcal{N}_r(x, y)$ to form different combinations as many as possible. The results are reported in Table V. As the projection parameters are shared for each path, there are no changes of model size using different $n \times n$ size combinations. Nevertheless, more computational costs always accompany by larger number of neighbor tokens that are associated with attention calculation. The best performance peaks when collecting 3×3, 5×5, and 7×7 for each path, which are also opt to default settings in our DNA.

V. Conclusion remarks and future work

This paper presents a DNAT for image classification [16]. Unlike previous multi-scale Transformers [9], [13], [15], DNAT adopts a multi-path design in each Transformer block, where each path employs DNA with different dilation rates to investigate various scale interactions. Extensive experiments show that, with similar model size, GFLOPs and Throughput of state-of-the-art Transformer networks, DNAT is able to achieve impressive classification accuracy.

In the future, we are interested in introducing DNAT to downstream visual tasks such as object detection [34], and semantic segmentation [35].

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[2] C. Szegedy, W. Liu, Y. Jia *et al.*, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[3] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[4] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[6] K. Han, A. Xiao, E. Wu *et al.*, "Transformer in transformer," in *NIPS*, 2021, pp. 15 908–15 919.

[7] A. Hatamizadeh, H. Yin, G. Heinrich *et al.*, "Global context vision transformers," in *ICML*, 2023, pp. 12 633–12 646.

[8] Y. Lee, J. Kim, J. Willette *et al.*, "Mpvit: Multi-path vision transformer for dense prediction," in *CVPR*, 2022, pp. 7287–7296.

[9] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[10] W. Wang, E. Xie, X. Li *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[11] Y.-H. Wu, Y. Liu, X. Zhan *et al.*, "P2t: Pyramid pooling transformer for scene understanding," *IEEE TPAMI*, pp. 1–12, 2022.

[12] A. Hassani, S. Walton, J. Li *et al.*, "Neighborhood attention transformer," in *CVPR*, 2023, pp. 6185–6194.

[13] Q. Zhang, Y. Xu, J. Zhang *et al.*, "Vsa: learning varied-size window attention in vision transformers," in *ECCV*, 2022, pp. 466–483.

[14] T. Yu, G. Zhao, P. Li *et al.*, "Boat: Bilateral local attention vision transformer," *arXiv preprint arXiv:2201.13027*, 2022.

[15] Z. Liu, H. Hu, Y. Lin *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *CVPR*, 2022, pp. 12 009–12 019.

[16] J. Deng, W. Dong, R. Socher *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[17] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," https://www.cs.toronto.edu/~kriz/cifar.html, 2009.

[18] W. Li, X. Wang, X. Xia *et al.*, "Sepvit: Separable vision transformer," *arXiv preprint arXiv:2203.15380*, 2022.

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

[20] X. Pan, T. Ye, Z. Xia *et al.*, "Slide-transformer: Hierarchical vision transformer with local self-attention," in *CVPR*, 2023, pp. 2082–2091.

[21] L. Yuan, Y. Chen, T. Wang *et al.*, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *ICCV*, 2021, pp. 558–567.

[22] S. Ren, D. Zhou, S. He *et al.*, "Shunted self-attention via multi-scale token aggregation," in *CVPR*, 2022, pp. 10 853–10 862.

[23] W. Wang, L. Yao, L. Chen *et al.*, "Crossformer: A versatile vision transformer hinging on cross-scale attention," in *ICLR*, 2022.

[24] H. Fan, B. Xiong, K. Mangalam *et al.*, "Multiscale vision transformers," in *ICCV*, 2021, pp. 6824–6835.

[25] H. Touvron, M. Cord, M. Douze *et al.*, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021, pp. 10 347–10 357.

[26] Z. Chen, Y. Zhu, C. Zhao *et al.*, "Dpt: Deformable patch-based transformer for visual recognition," in *ACM MM*, 2021, pp. 2899–2907.

[27] X. Chu, Z. Tian, Y. Wang *et al.*, "Twins: Revisiting the design of spatial attention in vision transformers," in *NIPS*, 2021, pp. 9355–9366.

[28] Z. Xia, X. Pan, S. Song *et al.*, "Vision transformer with deformable attention," in *CVPR*, 2022, pp. 4794–4803.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[30] X. Li, W. Wang, X. Hu *et al.*, "Selective kernel networks," in *CVPR*, 2019, pp. 510–519.

[31] Z. Zhang, H. Zhang, L. Zhao *et al.*, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *AAAI*, 2022, pp. 3417–3425.

[32] H. Wu, B. Xiao, N. Codella *et al.*, "Cvt: Introducing convolutions to vision transformers," in *ICCV*, 2021, pp. 22–31.

[33] Z. Lu, H. Xie, C. Liu *et al.*, "Bridging the gap between vision transformers and convolutional neural networks on small datasets," in *NIPS*, 2022, pp. 14 663–14 677.

[34] G. Zhang, Z. Luo, Z. Tian *et al.*, "Towards efficient use of multi-scale features in transformer-based object detectors," in *CVPR*, 2023, pp. 6206–6216.

[35] C. Pan, Y. He, J. Peng *et al.*, "Baeformer: Bi-directional and early interaction transformers for bird's eye view semantic segmentation," in *CVPR*, 2023, pp. 9590–9599.