TransXNet: Learning Both Global and Local Dynamics with a Dual Dynamic Token Mixer for Visual Recognition

Meng Lou, Hong-Yu Zhou, *Member, IEEE*, Sibei Yang, Yizhou Yu, *Fellow, IEEE*

arXiv:2310.19380v2 [cs.CV] 30 Nov 2023

Abstract—Recent studies have integrated convolution into transformers to introduce inductive bias and improve generalization performance. However, the static nature of conventional convolution prevents it from dynamically adapting to input variations, resulting in a representation discrepancy between convolution and self-attention as self-attention calculates attention matrices dynamically. Furthermore, when stacking token mixers that consist of convolution and self-attention to form a deep network, the static nature of convolution hinders the fusion of features previously generated by self-attention into convolution kernels. These two limitations result in a sub-optimal representation capacity of the constructed networks. To find a solution, we propose a lightweight Dual Dynamic Token Mixer (D-Mixer) that aggregates global information and local details in an input-dependent way. D-Mixer works by applying an efficient global attention module and an input-dependent depthwise convolution separately on evenly split feature segments, endowing the network with strong inductive bias and an enlarged effective receptive field. We use D-Mixer as the basic building block to design TransXNet, a novel hybrid CNN-Transformer vision backbone network that delivers compelling performance. In the ImageNet-1K image classification task, TransXNet-T surpasses Swin-T by 0.3% in top-1 accuracy while requiring less than half of the computational cost. Furthermore, TransXNet-S and TransXNet-B exhibit excellent model scalability, achieving top-1 accuracy of 83.8% and 84.6% respectively, with reasonable computational costs. Additionally, our proposed network architecture demonstrates strong generalization capabilities in various dense prediction tasks, outperforming other state-of-the-art networks while having lower computational costs. Our code will be available at https://github.com/LMMMEng/TransXNet.

Index Terms—Visual Recognition, Vision Transformer, Dual Dynamic Token Mixer

I. INTRODUCTION

Vision Transformer (ViT) [1] has shown promising progress in computer vision by using multi-head self-attention (MHSA) to achieve long-range modeling. However, it does not inherently encode inductive bias as convolutional neural networks (CNNs), resulting in a relatively weak generalization ability [2], [3]. To address this limitation, Swin Transformer [4]

Manuscript received mm, dd, yyyy; revised mm, dd, yyyy. (Corresponding author: Yizhou Yu.)

Meng Lou is with the Artificial Intelligence Laboratory, Deepwise Healthcare, Beijing, China (E-mail: loumeng@deepwise.com).

Hong-Yu Zhou and Yizhou Yu are with the Department of Computer Science, The University of Hong Kong, Hong Kong (E-mail: whuzhouhongyu@gmail.com; yizhouy@acm.org).

Sibei Yang is with the Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University, Shanghai, China (E-mail: yangsb@shanghaitech.edu.cn).
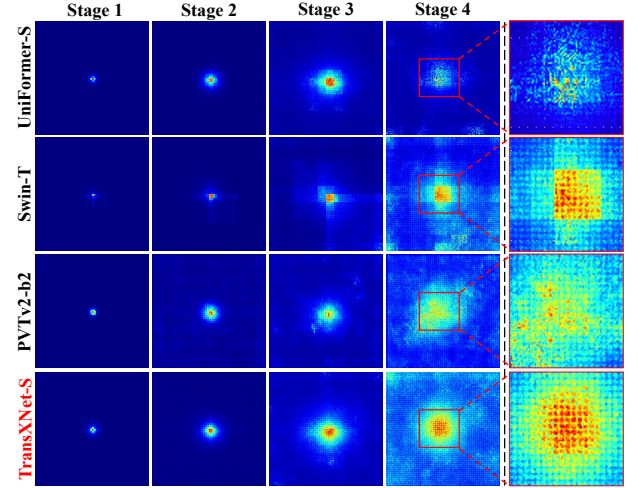

Fig. 1. Visualization of effective receptive fields (ERF). The results are obtained by averaging over 100 images from ImageNet-1K.

introduces shifted window self-attention, which incorporates inductive bias and reduces the computational cost of MHSA. However, Swin Transformer has a limited receptive field due to the local nature of its window-based attention.

To enable vision transformers to possess inductive bias, many recent works [5]–[8] have constructed hybrid networks that integrate self-attention and convolution within token mixers. However, the utilization of standard convolutions in these hybrid networks leads to limited performance improvements despite the presence of inductive bias. The reason is twofold. First, unlike self-attention which dynamically calculates attention matrices when given an input, standard convolution kernels are input-independent and unable to adapt to different inputs. This results in a discrepancy in representation capacity between convolution and self-attention. This discrepancy dilutes the modeling capability of self-attention as well as existing hybrid token mixers. Second, existing hybrid token mixers face challenges in deeply integrating convolution and self-attention. As a model goes deeper by stacking multiple hybrid token mixers, self-attention is capable of dynamically incorporating features generated by convolution in the preceding blocks while the static nature of convolution prevents it from effectively incorporating and utilizing features previously generated by self-attention. In this work, we aim to design an input-dependent dynamic convolution mechanism that is well
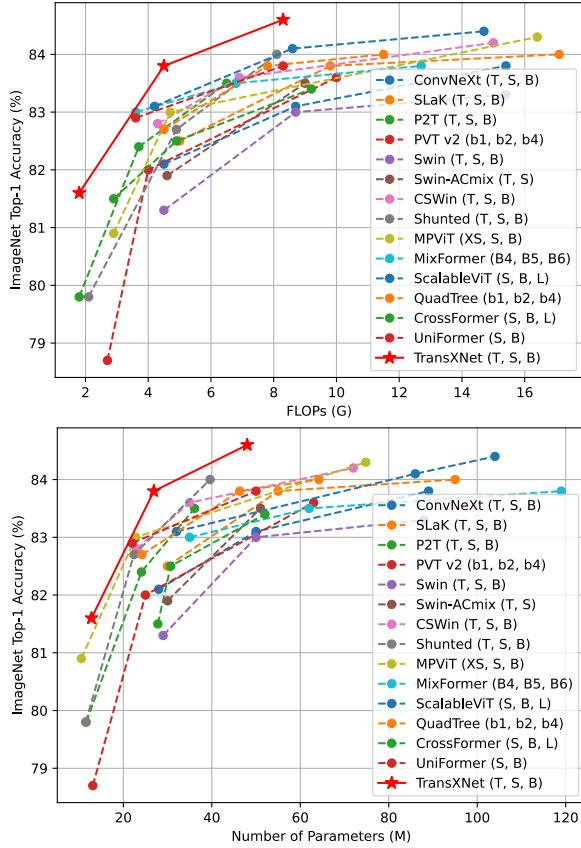
Fig. 2. Comparison of Top-1 accuracy on ImageNet-1K with recent SOTA methods. Our proposed TransXNet outperforms all other models.

suited for deep integration with self-attention within a hybrid token mixer so as to overcome the aforementioned challenges, resulting in a stronger representation capacity of the entire network.

On the other hand, a network should also have a large receptive field along with inductive bias to capture abundant contextual information. To this end, we obtain an interesting insight through effective receptive field (ERF) [9] analysis: leveraging global self-attention across all stages can effectively enlarge a model's ERF. Specifically, we visualize the ERF of three representative networks with similar computational cost, including UniFormer-S [10], Swin-T [4], and PVTv2-b2 [11]. Given a 224×224 input image, UniFormer-S and Swin-T exhibit locality at shallow stages and capture global information at the deepest stage, while PVTv2-b2 enjoys global information throughout the entire network. Results in Fig. 1 indicate that while all three networks employ global attention in the deepest layer, the ERF of PVTv2-b2 is clearly larger than that of UniFormer-S and Swin-T. According to this observation, to encourage a large receptive field, an efficient global self-attention mechanism should be encapsulated into all stages of a network. We also empirically find out that integrating dynamic convolutions with global self-attention can further enlarge the receptive field.

On the basis of the above discussions, we introduce a novel Dual Dynamic Token Mixer (D-Mixer) that aggregates features both globally and locally in an input-dependent way. Specifi-

cally, the input features are split into two half segments, which are respectively processed by an Overlapping Spatial Reduction Attention module and an Input-dependent Depthwise Convolution. The resulting two outputs are then concatenated together. Such a simple design can make a network see global contextual information while injecting effective inductive bias. As shown in Fig. 1, our method stands out among its competitors, yielding the largest ERF. In addition, zoom-in views (last column) reveal that our proposed mixer has remarkable local sensitivity in addition to non-local attention. We further introduce a Multi-scale Feed-forward Network (MS-FFN) that explores multi-scale information during token aggregation. By hierarchically stacking basic blocks composed of a D-Mixer and an MS-FFN, we construct a versatile backbone network called TransXNet for visual recognition. As illustrated in Fig. 2, our method showcases superior performance when compared to recent state-of-the-art (SOTA) methods in ImageNet-1K [12] image classification. In particular, our TransXNet-T achieves 81.6% top-1 accuracy with only 1.8G FLOPs and 12.8M Parameters (Params), outperforming Swin-T while incurring less than half of its computational cost. Also, our TransXNet-S/B models achieve 83.8%/84.6% top-1 accuracy, surpassing recently proposed InternImage [13] while incurring less computational cost.

In summary, our main contributions include: First, we propose a novel token mixer called D-Mixer, which aggregates sparse global information and local details in an input-dependent way, giving rise to both large ERF and strong inductive bias. Second, we design a novel and powerful vision backbone called TransXNet by employing D-Mixer as its token mixer. Finally, we conduct extensive experiments on image classification, object detection, and semantic and instance segmentation tasks. Results show that our method outperforms previous methods while having lower computational cost, achieving SOTA performance.

II. RELATED WORK

A. Convolutional Neural Networks

Throughout the field of computer vision, Convolutional Neural Networks (CNNs) have emerged as the standard deep model. Modern CNNs abandon the classical 3×3 convolution kernel and gradually adopt a model design centered on large kernels. For instance, ConvNeXt [14] employs 7×7 depthwise convolution as the network's building block. RepLKNet [15] investigates the potential of large kernel and further extends the convolution kernel to 31×31. Furthermore, SLaK [16] exploits the sparsity of convolution kernel and enlarges the kernel size beyond 51×51. More recently, InternImage [13] proposes a large-scale vision foundation model that surpasses state-of-the-art CNN and transformer models by using 3×3 deformable convolutions as the core operator.

B. Vision Transformer

Transformer was first proposed in the field of natural language processing [17], it can effectively perform dense relations among tokens in a sequence by adopting MHSA. To adapt computer vision tasks, ViT [1] split an image into

many image tokens through patch embedding operation, thus MHSA can be successfully utilized to model token-wise dependencies. However, vanilla MHSA is computationally expensive for processing high-resolution inputs, while dense prediction tasks such as object detection and segmentation generally require hierarchical feature representations to handle objects with different scales. To this end, many subsequent works adopted efficient attention mechanisms with pyramid architecture designs to achieve dense predictions, such as window attention [4], [18], [19], sparse attention [11], [20]–[22], and cross-layer attention [23], [24].

C. CNN-Transformer Hybrid Networks

Since relatively weak generalization is caused by lacking inductive biases in pure transformers [2], [3], CNN-Transformer hybrid models have emerged as a promising alternative that can leverage the advantages of both CNNs and transformers in vision tasks. For instance, some models used CNNs in the shallow layers and transformers in the deep layers [3], [10], [25], [26], while others integrated CNNs and transformers in a building block [5]–[8]. These models demonstrate that CNN-Transformer hybrid models can effectively combine the strengths of both paradigms and achieve notable results in various vision tasks.

D. Dynamic Weight

Dynamic weight is a powerful factor for the superiority of self-attention, enabling it to extract features dynamically according to the input, in addition to its long-range modeling capability. Similarly, dynamic convolution has been shown to be effective in improving the performance of CNN models [27]–[30] by extracting more discriminative local features with input-dependent filters. Among these methods, Han et al. [30] demonstrated that replacing the shifted window attention modules in Swin Transformer with dynamic depthwise convolutions achieves better results with lower computational cost.

Different from the aforementioned works, our goal is to construct an efficient yet powerful token mixer that processes both global and local information in an input-dependent manner, equipping the entire model with a large ERF and strong inductive bias.

III. METHOD

As illustrated in Fig. 3, our proposed TransXNet adopts a hierarchical architecture with four stages, which is similar to many previous works [18], [22], [31]. Each stage consists of a patch embedding layer and several sequentially stacked blocks. We implement the first patch embedding layer using a 7×7 convolutional layer (stride=4) followed by Batch Normalization (BN), while the patch embedding layers of the remaining stages use 3×3 convolutional layers (stride=2) with BN. Each block consists of a Dynamic Position Encoding (DPE) [10] layer, a Dual Dynamic Token Mixer (D-Mixer), and a Multi-scale Feed-forward Network (MS-FFN).

A. Dual Dynamic Token Mixer (D-Mixer)

To enhance the generalization ability of the Transformer model by incorporating inductive biases, many previous methods have combined convolution and self-attention to build a hybrid model [3], [5]–[8], [10], [25], [26]. However, their static convolutions dilute the input dependency of Transformers, i.e., although convolutions naturally introduce inductive bias, they have limited ability to improve the model's representation learning capability. In this work, we propose a lightweight token mixer termed Dual Dynamic Token Mixer (D-Mixer), which dynamically leverages global and local information, injecting the potential of large ERF and strong inductive bias without compromising input dependency. The overall workflow of the proposed D-Mixer is illustrated in Fig. 4 (a). Specifically, for a feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, we first divide it uniformly along the channel dimension into two sub-feature maps, denoted as $\{\mathbf{X_1}, \mathbf{X_2}\} \in \mathbb{R}^{\frac{C}{2} \times H \times W}$. Subsequently, $\mathbf{X_1}$ and $\mathbf{X_2}$ are respectively fed to a global self-attention module called OSRA and a dynamic depthwise convolution called IDConv, yielding corresponding feature maps $\{\mathbf{X_1'}, \mathbf{X_2'}\} \in \mathbb{R}^{\frac{C}{2} \times H \times W}$, which are then concatenated along the channel dimension to generate output feature map $\mathbf{X}' \in \mathbb{R}^{C \times H \times W}$. Finally, we employ a Squeezed Token Enhancer (STE) for efficient local token aggregation. Overall, the proposed D-Mixer is expressed as:

$$\begin{aligned} \mathbf{X_1}, \mathbf{X_2} &= \mathrm{Split}(\mathbf{X}) \\ \mathbf{X}' &= \mathrm{Concat}(\mathrm{OSRA}(\mathbf{X_1}), \mathrm{IDConv}(\mathbf{X_2})) \\ \mathbf{Y} &= \mathrm{STE}(\mathbf{X}') \end{aligned} \quad (1)$$

From the above equation, we can find out that by stacking D-Mixers, the dynamic feature aggregation weights generated in OSRA and IDConv take into account both global and local information, thus encapsulating powerful representation learning capabilities into the model.

1) Input-dependent Depthwise Convolution: To inject inductive bias and perform local feature aggregation in a dynamic input-dependent way, we propose a new type of dynamic depthwise convolution, termed Input-dependent Depthwise Convolution (IDConv). As shown in Fig. 4 (b), taking an input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, adaptive average pooling is used to aggregate spatial contexts, compressing spatial dimension to K^2, which is then forwarded into two sequential 1×1 convolutions, yielding attention maps $\mathbf{A}' \in \mathbb{R}^{(G \times C) \times K^2}$, where G denotes the number of attention groups. Then, $\mathbf{A}'$ is reshaped into $\mathbb{R}^{G \times C \times K^2}$ and a softmax function is employed over the G dimension, thus generating attention weights $\mathbf{A} \in \mathbb{R}^{G \times C \times K^2}$. Finally, $\mathbf{A}$ is element-wise multiplied with a set of learnable parameters $\mathbf{P} \in \mathbb{R}^{G \times C \times K^2}$, and the output is summed over the G dimension, resulting in input-dependent depthwise convolution kernels $\mathbf{W} \in \mathbb{R}^{C \times K^2}$, which can be expressed as:

$$\begin{aligned} \mathbf{A}' &= \mathrm{Conv}_{1 \times 1}^{\frac{C}{r} \to (G \times C)}(\mathrm{Conv}_{1 \times 1}^{C \to \frac{C}{r}}(\mathrm{AdaptivePool}(\mathbf{X}))) \\ \mathbf{A} &= \mathrm{Softmax}(\mathrm{Reshape}(\mathbf{A}')) \\ \mathbf{W} &= \sum_{i=0}^{G} \mathbf{P}_i \mathbf{A}_i \end{aligned} \quad (2)$$
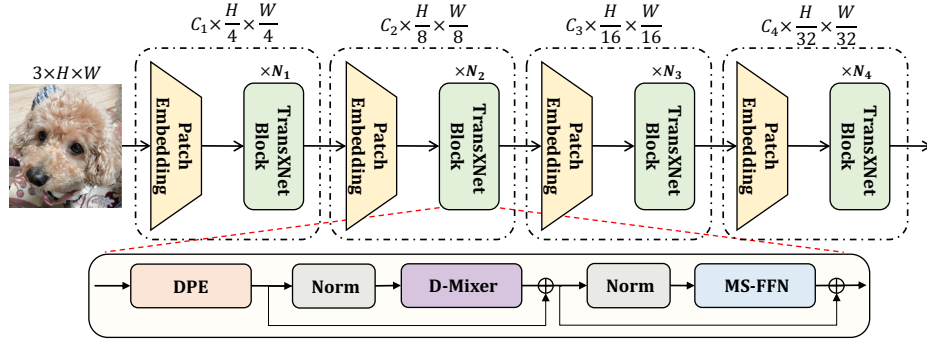
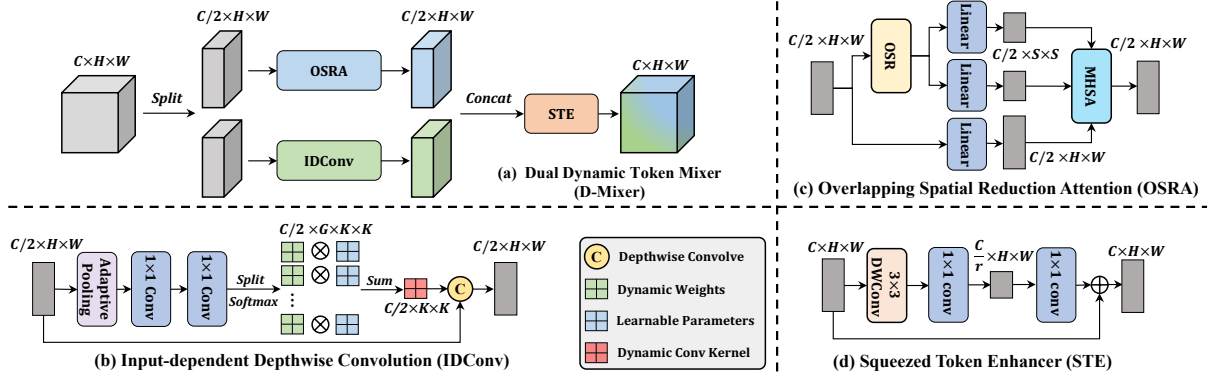Fig. 3. The overall architecture of the proposed TransXNet.

Fig. 4. Workflow of the proposed D-Mixer.

Since different inputs generate different attention maps $\mathbf{A}$, convolution kernels $\mathbf{W}$ vary with inputs. There are existing dynamic convolution schemes [29], [30]. In comparison to Dynamic Convolution (DyConv) [29], IDConv generates a spatially varying attention map for every attention group and the spatial dimensions ($K \times K$) of such attention maps exactly match those of convolution kernels while DyConv only generates a scalar attention weight for each attention group. Hence, our IDConv enables more dynamic local feature encoding. In comparison to recently proposed Dynamic Depthwise Convolution (D-DWConv) [30], IDConv combines dynamic attention maps with static learnable parameters to significantly reduce computational overhead. It is noted that D-DWConv applies global average pooling followed by channel squeeze-and-expansion pointwise convolutions on input features, resulting in an output with dimension $(C \times K^2) \times 1 \times 1$, which is then reshaped to match the depthwise convolutional kernel. The number of Params incurred in this procedure is $\frac{C^2}{r}(K^2 + 1)$, while our IDConv results in $\frac{C^2}{r}(G + 1) + GCK^2$ Params. In practice, when the maximum value of G is set to 4, and r and K are set to 4 and 7, respectively, the number of Params of IDConv ($1.25C^2 + 196C$) is much smaller than that of D-DWConv ($12.5C^2$).

2) Overlapping Spatial Reduction Attention (OSRA): Spatial Reduction Attention (SRA) [21] has been widely used in previous works [7], [11], [20], [32] to efficiently extract global information by exploiting sparse token-region relations. However, non-overlapping spatial reduction for reducing the token count breaks spatial structures near patch boundaries

and degrades the quality of tokens. To address this issue, we introduce Overlapping Spatial Reduction (OSR) for SRA to better represent spatial structures near patch boundaries by using larger and overlapping patches. In practice, we instantiate OSR as depthwise separable convolution [33], where the stride follows PVT [21] and the kernel size equals the stride plus 3. Fig. 4 (c) depicts the OSRA pipeline, which can also be formulated as:

$$
\begin{aligned}
\mathbf{Y} &= \mathrm{OSR}(\mathbf{X}) \\
\mathbf{Q} &= \mathrm{Linear}(\mathbf{X}) \\
\mathbf{K}, \mathbf{V} &= \mathrm{Split}(\mathrm{Linear}(\mathbf{Y} + \mathrm{LR}(\mathbf{Y}))) \\
\mathbf{Z} &= \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}} + \mathbf{B})\mathbf{V}
\end{aligned}
\tag{3}
$$

where $\mathrm{LR}(\cdot)$ denotes a local refinement module that is instantiated by a 3×3 depthwise convolution, $\mathbf{B}$ is a relative position bias matrix that encodes the spatial relations in attention maps [7], [26], and d is the number of channels in each attention head.

3) Squeezed Token Enhancer (STE): After performing token mixing, most previous methods use a 1×1 convolution to achieve cross-channel communications, which incurs considerable computational overhead. To reduce the computational cost without compromising performance, we propose a lightweight Squeezed Token Enhancer (STE), as shown in Fig. 4 (d). STE comprises a 3×3 depthwise convolution for enhancing local relations, channel squeeze-and-expansion 1×1 convolutions for reducing the computational cost, and a residual connection
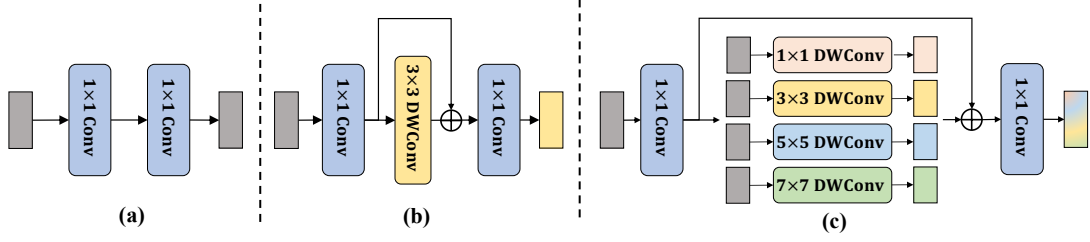
Fig. 5. (a) Vanilla FFN only handles cross-channel information. (b) Inverted Residual FFN further aggregates tokens in a small region. (c) Our MS-FFN performs multi-scale token aggregations.

for preserving the representation capacity. The STE can be expressed as follows:

$$\text{STE}(\mathbf{X}) = \text{Conv}_{1\times1}^{\frac{C}{r}\to C}(\text{Conv}_{1\times1}^{C\to\frac{C}{r}}(\text{DWConv}_{3\times3}(\mathbf{X}))) + \mathbf{X}. \tag{4}$$

According to the above equation, the FLOPs of STE is $HWC(2C/r + 9)$. In practice, we set the channel reduction ratio r to 8, but ensure that the number of compressed channels is not less than 16, resulting in FLOPs significantly less than that of a 1×1 convolution, i.e., HWC^2.

B. Multi-scale Feed-forward Network (MS-FFN)

Compared to vanilla FFN [1], Inverted Residual FFN [7] achieves local token aggregation by introducing a 3×3 depthwise convolution into the hidden layer. However, due to the larger number of channels in the hidden layer, i.e., typically four times the number of input channels, single-scale token aggregation cannot fully exploit such rich channel representations. To this end, we introduce a simple yet effective MS-FFN. As shown in Fig. 5, instead of using a single 3×3 depthwise convolution, we use four parallel depthwise convolutions with different scales, each of which handles a quarter of the channels. The depthwise convolution kernels with kernel size=$\{3, 5, 7\}$ can effectively capture multi-scale information, while a 1×1 depthwise convolution kernel is in fact a learnable channel-wise scaling factor.

C. Architecture Variants

The proposed TransXNet has three different variants: TransXNet-T (Tiny), TransXNet-S (Small), and TransXNet-B (Base). To control the computational cost of different variants, there are two other adjustable hyperparameters in addition to the number of channels and blocks. First, since the computational cost of IDConv is directly related to the number of attention groups, we use a different number of attention groups in IDConv for different variants. In the tiny version, the number of attention groups is fixed at 2 to ensure a reasonable computational cost, while in the deeper small and base models, an increasing number of attention groups is used to improve the flexibility of IDConv, which is similar to the increase in the number of heads of the MHSA module as the model goes deeper. Second, many previous works [11], [20]–[22] set the expansion ratio of the FFNs in stages 1 and 2 to 8. However, since feature maps in stages 1 and 2 usually have larger resolutions, this leads to high FLOPs. Hence, we

gradually increase the expansion ratio in different architecture variants. Details of different architecture variants are listed in Table I.

IV. EXPERIMENTS

To assess the efficacy of our TransXNet, we evaluate it on various tasks, including image classification on the ImageNet-1K dataset [12], object detection and instance segmentation on the COCO dataset [34], and semantic segmentation on the ADE20K dataset [35]. Additionally, we conduct extensive ablation studies to analyze the impact of different components of our model.

A. Image classification

Setup. Image classification is performed on the ImageNet-1K dataset, following the experimental settings of DeiT [2] for a fair comparison with SOTA methods, i.e., all models are trained for 300 epochs with the AdamW optimizer [36]. The stochastic depth rate [37] is set to 0.1/0.2/0.4 for tiny, small, and base models, respectively. All the experiments are conducted on 8 NVIDIA V100 GPUs. Furthermore, to further demonstrate the generalizability of our method, we perform additional assessments on the ImageNet-1K-V2 dataset [38] using our ImageNet-pretrained weights, adhering settings outlined in [39].

Results. The proposed method outperforms other competitors in ImageNet-1K image classification with 224×224 images, as summarized in Table II. First, TransXNet-T achieves an impressive top-1 accuracy of 81.6% with only 1.8 GFLOPs and 12.8M Params, surpassing other methods by a large margin. Despite having less than half of the computational cost, TransXNet-T achieves 0.3% higher top-1 accuracy than Swin-T [4]. Second, TransXNet-S achieves a remarkable top-1 accuracy of 83.8%, which is higher than InternImage-T [13] by 0.2% without requiring specialized CUDA implementations[1]. Moreover, our method outperforms well-known hybrid models, including MixFormer [6] and MaxViT [8], while having a lower computational cost. Notably, our small model performs better than MixFormer-B5 whose number of Params actually exceeds our base model. Note that the performance improvement of TransXNet-S over CMT-S [7] in image classification appears limited because CMT has a more

---

[1]As the core operator of InternImage, DCNv3 relies on specialized CUDA implementations for accelerating on GPU, while our method can be more easily generalized to various devices without CUDA support.

TABLE I
DETAILED CONFIGURATIONS OF TRANSXNET VARIANTS, INCLUDING STRIDE OF OSRA (S), NUMBER OF ATTENTION HEADS OF OSRA (H), KERNEL SIZE OF IDCONV (K), NUMBER OF ATTENTION GROUPS IN IDCONV (G), AND EXPANSION RATIO OF MS-FFN (E).

Input Size	Operator	TransXNet-T	TransXNet-S	TransXNet-B
224×224	Patch Embed	7×7, 48, stride=4	7×7, 64, stride=4	7×7, 76, stride=4
56×56	DPE D − Mixer MS − FFN	$\begin{bmatrix} S=8, H=1 \\ K=7, G=2 \\ E=4 \end{bmatrix} \times 3$	$\begin{bmatrix} S=8, H=1 \\ K=7, G=2 \\ E=6 \end{bmatrix} \times 4$	$\begin{bmatrix} S=8, H=2 \\ K=7, G=2 \\ E=8 \end{bmatrix} \times 4$
	Patch Embed	3×3, 96, stride=2	3×3, 128, stride=2	3×3, 152, stride=2
28×28	DPE D − Mixer MS − FFN	$\begin{bmatrix} S=4, H=2 \\ K=7, G=2 \\ E=4 \end{bmatrix} \times 3$	$\begin{bmatrix} S=4, H=2 \\ K=7, G=2 \\ E=6 \end{bmatrix} \times 4$	$\begin{bmatrix} S=4, H=4 \\ K=7, G=2 \\ E=8 \end{bmatrix} \times 4$
	Patch Embed	3×3, 224, stride=2	3×3, 320, stride=2	3×3, 336, stride=2
14×14	DPE D − Mixer MS − FFN	$\begin{bmatrix} S=2, H=4 \\ K=7, G=2 \\ E=4 \end{bmatrix} \times 9$	$\begin{bmatrix} S=2, H=5 \\ K=7, G=3 \\ E=4 \end{bmatrix} \times 12$	$\begin{bmatrix} S=2, H=8 \\ K=7, G=4 \\ E=4 \end{bmatrix} \times 21$
	Patch Embed	3×3, 448, stride=2	3×3, 512, stride=2	3×3, 672, stride=2
7×7	DPE D − Mixer MS − FFN	$\begin{bmatrix} S=1, H=8 \\ K=7, G=2 \\ E=4 \end{bmatrix} \times 3$	$\begin{bmatrix} S=1, H=8 \\ K=7, G=4 \\ E=4 \end{bmatrix} \times 4$	$\begin{bmatrix} S=1, H=16 \\ K=7, G=4 \\ E=4 \end{bmatrix} \times 4$
		Global Average Pooling		
1×1		Fully Connected Layer, 1000		
# FLOPs		1.8 G	4.5 G	8.3 G
# Params		12.8 M	26.9 M	48.0 M

complex classification head to boost performance. In contrast, benefiting from the stronger representation capacity of the backbone network, our method exhibits very clear advantages in downstream tasks including object detection and instance segmentation (refer to section IV-B). Finally, TransXNet-B leads other methods by achieving an excellent balance between performance and computational cost, boasting a top-1 accuracy of 84.6%. On the other hand, it is worth highlighting that our method exhibits a more pronounced performance advantage on the ImageNet-1K-V2 dataset. Specifically, TransXNet-tiny, -small, and -base achieve top-1 of 70.7%, 73.8%, and 75.0%, respectively. This demonstrates the superior generalization and transferability of our method compared to its counterparts.

B. Object Detection and Instance Segmentation

Setup. To evaluate our method on object detection and instance segmentation tasks, we conduct experiments on COCO 2017 [34] using the MMDetection [46] codebase. Specifically, for object detection, we use the RetinaNet framework [47], while instance segmentation is performed using the Mask R-CNN framework [48]. For fair comparisons, we initialize all backbone networks with weights pre-trained on ImageNet-1K, while training settings follow the 1× schedule provided by PVT [21].

Results. We present results in Table III. For object detection with RetinaNet, our method attains the best performance in comparison to other competitors. It is noted that previous methods often fail to simultaneously perform well on both small and large objects. However, our method, supported by global and local dynamics and multi-scale token aggregation, not only achieves excellent results on small targets but also significantly outperforms previous methods on medium and large targets. For example, the recently proposed Slide-PVTv2-b1 [19], which focuses on local information modeling, achieves comparable AP_S to our tiny model, while our

method improves AP_M/AP_L by 1.9%/2.5% while having less computational cost, underscoring its effectiveness in modeling both global and local information. This phenomenon is more prominent in the comparison groups of small and base models, demonstrating the superior performance of our method across different object sizes. Regarding instance segmentation with Mask-RCNN, our method also has a clear advantage over previous methods with a comparable computational cost. It is worth mentioning that even though TransXNet-S shows limited performance improvement over CMT-S [7] in ImageNet-1K classification, it achieves obvious performance improvements in object detection and instance segmentation, which indicates that our backbone has stronger representation capacity and better transferability.

C. Semantic Segmentation

Setup. We conduct semantic segmentation on the ADE20K dataset [35] using the MMSegmentation [51] codebase. The commonly used Semantic FPN [52] is employed as the segmentation framework. For fair comparisons, all backbone networks are initialized with ImageNet-1K pre-trained weights, and training settings follow PVT [11].

Results. Table IV demonstrates the performance achieved by our method in comparison to other competitors. Note that since some methods (e.g., CMT [7] and MaxViT [8]) do not report semantic segmentation results in their papers, we do not compare with them. Specifically, our TransXNet-T achieves a remarkable 45.5% mIoU, surpassing the second-best method by 2.1% in mIoU while maintaining a similar computational cost. Additionally, TransXNet-S improves the mIoU by modest 0.3% over CSWin-T [18] but with fewer GFLOPs. Finally, TransXNet-B achieves the highest mIoU of 49.9%, surpassing other competitors but with less computational cost.

Method	#F (G)	#P (M)	Top-1	V2 Top-1
RSB-ResNet-18 [40]	1.8	11.7	70.6	-
RegNetY-1.6G [41]	1.6	11.2	78.0	66.2
PVT-ACmix-T [5]	2.0	13.0	78.0	-
PVTv2-b1 [11]	2.1	13.1	78.7	66.9
Shunted-T [20]	2.1	11.5	79.8	66.9
P2T-T [22]	1.8	11.6	79.8	70.0
QuadTree-B-b1 [42]	2.3	13.6	80.0	67.2
MPViT-XS [43]	2.9	10.5	80.9	70.0
TransXNet-T	1.8	12.8	**81.6**	**70.7**
ConvNeXt-T [14]	4.5	29.0	82.1	71.0
SLaK-T [16]	5.0	30.0	82.5	-
InternImage-T [13]	5.0	30.0	83.5	73.0
Swin-T [4]	4.5	29.0	81.3	69.7
CSWin-T [18]	4.5	23.0	82.7	72.5
PVTv2-b2 [11]	4.0	25.4	82.0	71.8
Shunted-S [20]	4.9	22.4	82.9	72.4
UniFormer-S [10]	3.6	22.0	82.9	71.9
ScalableViT-S [44]	4.2	32.0	83.1	71.5
MOAT-0 [39]	5.7	27.8	83.3	72.8
BSwin-T [24]	4.5	29.0	82.0	-
Slide-Swin-T [19]	4.6	29.0	82.3	-
Slide-CSWin-T [19]	4.3	23.0	83.2	-
QuadTree-B-b2 [42]	4.5	24.2	82.7	70.4
MPViT-S [43]	4.7	22.8	83.0	72.3
CrossFormer-S [45]	4.9	30.7	82.5	71.5
CMT-S [7]	4.0	26.3	83.5	73.4
MaxViT-T [8]	5.6	31.0	83.6	73.1
Swin-ACmix-T [5]	4.6	30.0	81.6	70.7
MixFormer-B4 [6]	3.6	35.0	83.0	-
TransXNet-S	4.5	26.9	**83.8**	**73.8**
ConvNeXt-S [14]	8.7	50.0	83.1	72.4
SLaK-S [16]	9.8	55.0	83.8	-
InternImage-S [13]	8.0	50.0	84.2	73.9
Swin-B [4]	15.4	88.0	83.5	72.3
CSWin-B [18]	15.0	78.0	84.2	74.1
PVTv2-b4 [11]	10.1	62.6	83.6	73.7
P2T-L [22]	9.8	54.5	83.9	72.7
Shunted-B [20]	8.1	39.6	84.0	73.9
UniFormer-B [10]	8.3	50.0	83.9	73.0
ScalableViT-B [44]	8.6	81.0	84.1	72.9
MOAT-1 [39]	9.1	41.6	84.2	74.2
BSwin-S [24]	8.7	50.0	84.2	-
Slide-Swin-B [19]	15.5	89.0	84.2	-
QuadTree-B-b4 [42]	11.5	64.2	84.1	72.7
MPViT-B [43]	16.4	74.8	84.3	73.7
CrossFormer-L [45]	16.1	92.0	84.0	73.5
Swin-ACmix-S [5]	9.0	51.0	83.5	73.0
MixFormer-B5 [6]	6.8	62.0	83.5	-
TransXNet-B	8.3	48.0	**84.6**	**75.0**

D. Ablation Study

Setup. To evaluate the impact of each component in TransXNet, we conduct extensive ablation experiments on ImageNet-1K. Due to limited resources, we adjust the number of training epochs to 200 for all models, while keeping the rest of the experimental settings consistent with section IV-A. Subsequently, we proceeded to fine-tune the ImageNet-pretrained model on the ADE20K dataset, applying the identical training configurations as described in section IV-C.

Comparison of token mixers. To perform a fair comparison of token mixers, we adjust the tiny model to a similar style as Swin-T [4], i.e., setting the numbers of blocks and channels in the four stages to [2,2,6,2] and [64,128,256,512], respectively, and using non-overlapping patch embedding and vanilla FFN. The performance of different token mixers is shown in Table V.

It can be found that our D-Mixer has clear advantages in terms of performance and computational cost. In ImageNet-1K classification, D-Mixer ties with ACmix block [5] which is also a hybrid module, but our D-Mixer has a significantly lower computational cost. Furthermore, D-Mixer demonstrates a pronounced superiority in semantic segmentation, as demonstrated on the ADE20K dataset.

Comparison of depthwise convolutions. To evaluate the effectiveness of IDConv, we replace it in the tiny model with a series of alternatives including the standard depthwise convolution (DWConv), window attention [32], DyConv [29], and D-DWConv [30]. The kernel/window sizes of the above methods are set to 7×7 for fair comparisons. As listed in Table VI, IDConv exceeds DyConv by 0.2% top-1 accuracy and 0.7% mIoU with only a slight increase in Params. Then, window attention performs worse with higher computational cost, possibly due to its non-overlapping locality. Compared with the recently proposed D-DWConv, IDConv has fewer parameters while achieving comparable performance in image classification and superior result in semantic segmentation.

Impact of MS-FFN. Based on the tiny model, we investigate the impact of multi-scale token aggregation in MS-FFN by conducting comparisons between MS-FFN and vanilla FFN [1], while adjusting the kernel size in the middle layer of MS-FFN. The results presented in Table VII reveal that MS-FFN surpasses vanilla FFN and Inverted Residual FFN [7] (i.e., scale=3) by 0.6% in top-1 accuracy and 1.0% in mIoU, and 0.3% in top-1 accuracy and 0.4% in mIoU, respectively. Importantly, this performance boost comes with only a minor increase in computational cost. Our investigation identifies the optimal set of scales for MS-FFN as $\{1, 3, 5, 7\}$, striking a favorable balance between performance and computational efficiency. Although we observe further performance gains by extending the scale set to $\{1, 3, 5, 7, 9\}$, we opt to discard this configuration to avoid the accompanying increase in the number of parameters.

Impact of channel ratio between attention and convolution. Channel ratio represents the proportion of channels allocated to OSRA in a given feature map. We investigate the impact of channel ratio by setting it to different values in a tiny model. As shown in Table VIII, top-1 accuracy and mIoU are greatly improved when channel ratio is increased from 0.25 to 0.5. However, when the channel ratio becomes greater than 0.5, the improvement in top-1 accuracy and mIoU becomes unconspicuous even though the number of Params increases. Hence, we conclude that a channel ratio of 0.5 has the best trade-off between performance and model complexity.

Other model design choices. We verify the impact of DPE, OSR, and STE on a tiny model by removing or replacing these components. As shown in Table IX, DPE brings clear performance improvement, which is consistent with previous works [10], [53]. Regarding the design choice of the self-attention module, OSR demonstrates a slight yet noteworthy improvement of 0.1% in top-1 accuracy and 0.3% in mIoU, all

TABLE III
PERFORMANCE COMPARISON OF OBJECT DETECTION AND INSTANCE SEGMENTATION ON THE COCO DATASET. FLOPS ARE CALCULATED WITH RESOLUTION 800×1280.

Backbone	RetinaNet 1× schedule								Mask R-CNN 1× schedule							
	#F (G)	#P (M)	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#F (G)	#P (M)	AP^b	AP_{50}^b	AP_{75}^b	AP^m	AP_{50}^m	AP_{75}^m
ResNet-18 [49]	190	21.3	31.8	49.6	33.6	16.3	34.3	43.2	209	31.2	34.0	54.0	36.7	31.2	51.0	32.7
PoolFormer-S12 [31]	188	21.7	36.2	56.2	38.2	20.8	39.1	48.0	207	31.6	37.3	59.0	40.1	34.6	55.8	36.9
PVTv2-b1 [11]	209	23.8	40.2	60.7	42.4	22.8	43.3	54.0	227	33.7	41.8	64.3	45.9	38.8	61.2	41.6
PVT-ACmix-T [5]	232	-	40.5	61.2	42.7	-	-	-	-	-	-	-	-	-	-	-
ViL-T [50]	204	16.6	40.8	61.3	43.6	26.7	44.9	53.6	223	26.9	41.4	63.5	45.0	38.1	60.3	40.8
P2T-T [22]	206	21.1	41.3	62.0	44.1	24.6	44.8	56.0	225	31.3	43.3	65.7	47.3	39.6	62.5	42.3
MixFormer-B3 [6]	-	-	-	-	-	-	-	-	207	35.0	42.8	64.5	46.7	39.3	61.8	42.2
Slide-PVTv2-b1 [19]	204	-	41.5	62.3	44.0	26.0	44.8	54.9	222	33.0	42.6	65.3	46.8	39.7	62.6	42.6
TransXNet-T	187	22.4	**43.1**	**64.1**	**46.0**	**26.0**	**46.7**	**57.4**	205	32.5	**44.5**	**66.5**	**48.6**	**40.6**	**63.7**	**43.8**
Swin-T [4]	248	38.5	41.5	62.1	44.2	25.1	44.9	55.5	264	47.8	42.2	64.6	46.2	39.1	61.6	42.0
CSWin-T [18]	266	35.1	43.8	64.8	46.8	26.0	47.6	59.2	285	45.0	45.3	67.1	49.6	41.2	64.2	44.4
PVTv2-b2 [11]	-	-	-	-	-	-	-	-	279	42.0	46.7	68.6	51.3	42.2	65.6	45.4
PVT-ACmix-S [5]	232	-	40.5	61.2	42.7	-	-	-	-	-	-	-	-	-	-	-
Swin-ACmix-T [5]*	-	-	-	-	-	-	-	-	275	-	47.0	69.0	51.8	-	-	-
P2T-S [22]	260	33.8	44.4	65.3	47.6	27.0	48.3	59.4	279	43.7	45.5	67.7	49.8	41.4	64.6	44.5
MixFormer-B4 [6]	-	-	-	-	-	-	-	-	243	53.0	45.1	67.1	49.2	41.2	64.3	44.1
CrossFormer-S [45]	282	40.8	44.4	65.8	47.4	28.2	48.4	59.4	301	50.2	45.4	68.0	49.7	41.4	64.8	44.6
CMT-S [7]	231	44.3	44.3	65.5	47.5	27.1	48.3	59.1	249	44.5	44.6	66.8	48.9	40.7	63.9	43.4
InternImage-T [13]	-	-	-	-	-	-	-	-	270	49.0	47.2	69.0	52.1	42.5	66.1	45.8
Slide-PVTv2-b2 [19]	255	-	45.0	66.2	48.4	28.8	48.8	59.7	274	43.0	46.0	68.2	50.3	41.9	65.1	45.4
TransXNet-S	242	36.6	**46.4**	**67.7**	**50.0**	**28.9**	**50.3**	**61.1**	261	46.5	**47.7**	**69.9**	**52.3**	**43.1**	**66.9**	**46.5**
Swin-S [4]	336	59.8	44.5	65.7	47.5	27.4	48.0	59.9	354	69.1	44.8	66.6	48.9	40.9	63.4	44.2
CSWin-S [18]	-	-	-	-	-	-	-	-	342	54.0	47.9	70.1	52.6	43.2	67.1	46.2
PVTv2-b3 [11]	354	55.0	45.9	66.8	49.3	28.6	49.8	61.4	372	64.9	47.0	68.1	51.7	42.5	65.7	45.7
P2T-B [22]	344	45.8	46.1	67.5	49.6	30.2	50.6	60.9	363	55.7	47.2	69.3	51.6	42.7	66.1	45.9
CrossFormer-B [45]	389	62.1	46.2	67.8	49.5	30.1	49.9	61.8	408	71.5	47.2	69.9	51.8	42.7	66.6	46.2
InternImage-S [13]	-	-	-	-	-	-	-	-	340	69.0	47.8	69.8	52.8	43.3	67.1	46.7
Slide-PVTv2-b3 [19]	343	-	46.8	67.7	50.3	30.5	51.1	61.6	362	63.0	47.8	69.5	52.6	43.2	66.5	46.6
TransXNet-B	317	58.0	**47.6**	**69.0**	**51.1**	**31.3**	**51.7**	**62.2**	336	67.6	**48.8**	**70.8**	**53.5**	**43.8**	**67.9**	**47.2**

* ACmix uses the 3× schedule to train Mask R-CNN, while our method has better results despite using 1× schedule.

TABLE IV
PERFORMANCE COMPARISON OF SEMANTIC SEGMENTATION ON THE ADE20K DATASET. FLOPS ARE CALCULATED WITH RESOLUTION 512×2048.

Backbone	#F (G)	#P (M)	mIoU
ResNet-18 [49]	129	15.5	32.9
PoolFormer-S12 [31]	124	15.7	37.2
PVTv2-b1 [11]	129	17.8	42.5
PVT-ACmix-T [5]	160	17.0	42.7
P2T-T [22]	121	15.8	43.4
Slide-PVT-T [19]	136	16.0	38.4
TransXNet-T	121	16.6	**45.5**
Swin-T [4]	182	31.9	41.5
PVTv2-b2 [11]	167	29.1	45.2
PVT-ACmix-S [5]	228	29.0	46.4
CSWin-T [18]	202	26.1	48.2
ScalableViT-S [44]	174	30.0	44.9
CrossFormer-S [45]	221	34.3	46.0
UniFormer-S [10]	247	25.0	46.6
P2T-S [22]	162	28.4	46.7
Slide-PVT-S [19]	188	26.0	42.5
TransXNet-S	179	30.6	**48.5**
Swin-S [4]	274	53.2	45.2
PVTv2-b4 [11]	291	66.3	47.9
ScalableViT-B [44]	270	79.0	48.4
CrossFormer-B [45]	270	55.6	47.7
UniFormer-B [10]	471	54.0	48.0
P2T-L [22]	281	58.8	49.4
Slide-PVT-M [19]	278	46.0	44.0
TransXNet-B	256	51.7	**49.9**

TABLE V
COMPARISON OF TOKEN MIXERS.

Token Mixer	ImageNet-1K			ADE20K		
	#F (G)	#P (M)	Top-1	#F (G)	#P (M)	mIoU
Sep Conv [33]	1.6	11.4	76.9	119.7	14.6	39.9
SRA [21]	2.0	16.2	77.4	126.6	20.0	41.9
Swin [4]	2.0	13.6	78.2	130.5	17.4	40.3
Mixing Block [6]	2.0	13.6	78.9	130.0	17.4	42.3
ACmix Block [5]	2.1	13.9	**79.0**	131.4	17.6	41.5
D-Mixer (Ours)	1.6	11.4	**79.0**	118.0	15.2	**42.7**

TABLE VI
COMPARISON OF DEPTHWISE CONVOLUTIONS.

Local Operator	ImageNet-1K			ADE20K		
	#F (G)	#P (M)	Top-1	#F (G)	#P (M)	mIoU
DWConv [33]	1.8	12.5	80.3	122.0	16.3	44.1
DyConv [29]	1.8	12.3	80.7	121.4	16.5	44.3
Window Attention [32]	1.9	13.3	80.8	124.4	17.0	44.5
D-DWConv [30]	1.8	14.2	**80.9**	121.4	18.0	44.6
IDConv (Ours)	1.8	12.8	**80.9**	121.4	16.6	**45.0**

TABLE VII
ABLATION ON MS-FFN SCALES.

FFN Scale	ImageNet-1K			ADE20K		
	#F (G)	#P (M)	Top-1	#F (G)	#P (M)	mIoU
N/A	1.7	12.5	80.3	119.3	16.2	44.0
{1, 3}	1.7	12.6	80.6	119.9	16.4	44.5
3 (Single-scale)	1.7	12.7	80.6	120.3	16.4	44.6
{1, 3, 5}	1.7	12.7	80.7	120.5	16.5	44.7
{1, 3, 5, 7, }	1.8	12.8	80.9	121.4	16.6	**45.0**
{1, 3, 5, 7, 9}	1.8	13.0	**81.0**	122.5	16.8	44.7

TABLE VIII
ABLATION ON CHANNEL RATIO BETWEEN ATTENTION AND
CONVOLUTION.

Attention Ratio	ImageNet-1K			ADE20K		
	#F (G)	#P (M)	Top-1	#F (G)	#P (M)	mIoU
0.25	1.7	12.5	80.6	120.3	16.3	44.1
0.5	1.8	12.8	**80.9**	121.4	16.6	**45.0**
0.75	1.8	13.6	**80.9**	123.2	17.4	44.9

without incurring additional computational costs when compared to Non-overlapping Spatial Reduction (NSR). Notably, our proposed STE significantly reduces computational cost while maintaining consistent performance on ImageNet-1K and boosting mIoU by 0.3% on ADE20K, highlighting the effectiveness of STE as an efficient design choice for our model.

TABLE IX
ABLATION STUDY ON DPE, OSR, AND STE.

NSR	DPE	OSR	STE	ImageNet-1K			ADE20K		
				#F (G)	#P (M)	Top-1	#F (G)	#P (M)	mIoU
✓				1.8	13.4	80.5	122.2	17.1	44.1
✓	✓			1.9	13.6	80.8	123.4	17.3	44.5
	✓	✓		1.9	13.6	**80.9**	123.5	17.4	44.7
	✓	✓	✓	1.8	12.8	**80.9**	121.4	16.6	**45.0**

E. Network Visualization

1) Effective Receptive Field Analysis: To gain further insights into the superiority of our IDConv over the standard DWConv, we visualize the Effective Receptive Field (ERF) [9] of the deepest stage of all models considered in Table VI. As shown in Fig. 6 (a), DWConv has the smallest ERF in comparison to dynamic operators, including DyConv [28], Window Attention [32], D-DWConv [30], and IDConv. Furthermore, among the dynamic operators, it is evident that IDConv enables our model to achieve the largest ERF while preserving a strong locality. These observations substantiate the claim that incorporating suitable dynamic convolutions assists Transformers in better capturing global contexts while carrying potent inductive biases, thereby improving their representation capacity.

On the other hand, to demonstrate the powerful representation capacity of our TransXNet, we also compare the ERF of several SOTA methods with similar computational costs. As shown in Fig. 6 (b), our TransXNet-S has the largest ERF among these methods while maintaining strong local sensitivity, which is challenging to achieve.

2) Grad-CAM Analysis: To comprehensively assess the quality of the learned visual representations, we employ Grad-CAM technique [21] to generate activation maps for visual representations at various stages of TransXNet-S, Swin-T, UniFormer-S, and PVT v2-b2. These activation maps provide insight into the significance of individual pixels in depicting class discrimination for each input image. As depicted in Fig. 7, our approach stands out by revealing more intricate details in early layers and identifying more semantically meaningful regions in deeper layers. This compellingly demonstrates the
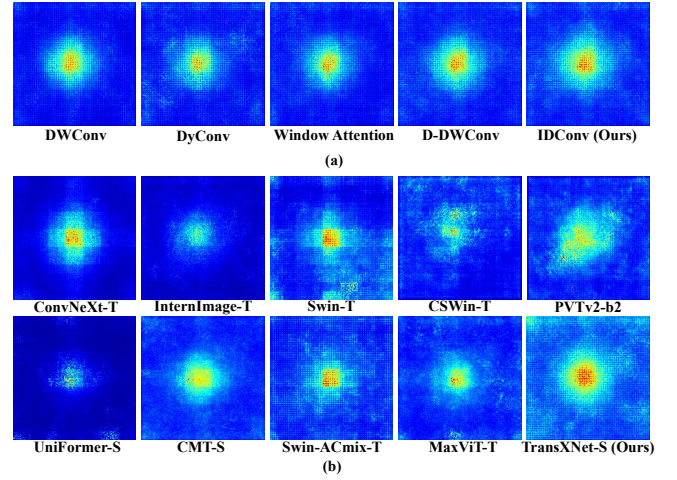

Fig. 6. ERF visualization of (a) models incorporating various local operators and (b) SOTA methods. The results are obtained by averaging over 100 images (resized to 224×224) from ImageNet.

robust visual representation capabilities of our method compared to other competitors.

F. Throughput Analysis

In this section, we present a comparison of inference throughput among different methods. This experiment is conducted on a single NVIDIA RTX TITAN GPU with a batch size of 32 and an image resolution of 224×224. First, we compare the throughput of TransXNet-S with some representative SOTA methods, including Swin [4], Swin-ACmix [5], QuadTree [42], P2T [22], Shunted [20], and MPViT [43]. As shown in Fig. 8 (a), it is apparent that TransXNet demonstrates reasonable throughput. However, the advantages are not pronounced when compared to other methods. We attribute this to the presence of operators that may not be as compatible with GPU-based parallel computing, such as multi-scale depthwise convolutions in MS-FFN. To confirm this point, we further evaluate the performance of all models considered in Table V. These models share a similar architectural design, and the only difference is the token mixer, i.e., Sep Conv [33], SRA [21], Shifted Window [4], Mixing block [6], ACmix block [5], and our D-Mixer. As shown in Fig. 8 (b), our D-Mixer performs comparably to the Shifted Window mixer of Swin while exhibiting advantages over other hybrid mixers, which demonstrates that our D-Mixer is both effective and GPU friendly.

It is worth mentioning that better GPU-based implementation of IDConv and MS-FFN can possibly improve the throughput of our network, and we leave this as our future work.

V. LIMITATIONS

Our ablation study suggests that a fixed 1:1 ratio between the numbers of channels allocated to self-attention and dynamic convolutions in all stages yields a favorable trade-off. However, we speculate that employing different ratios at different stages may further improve performance
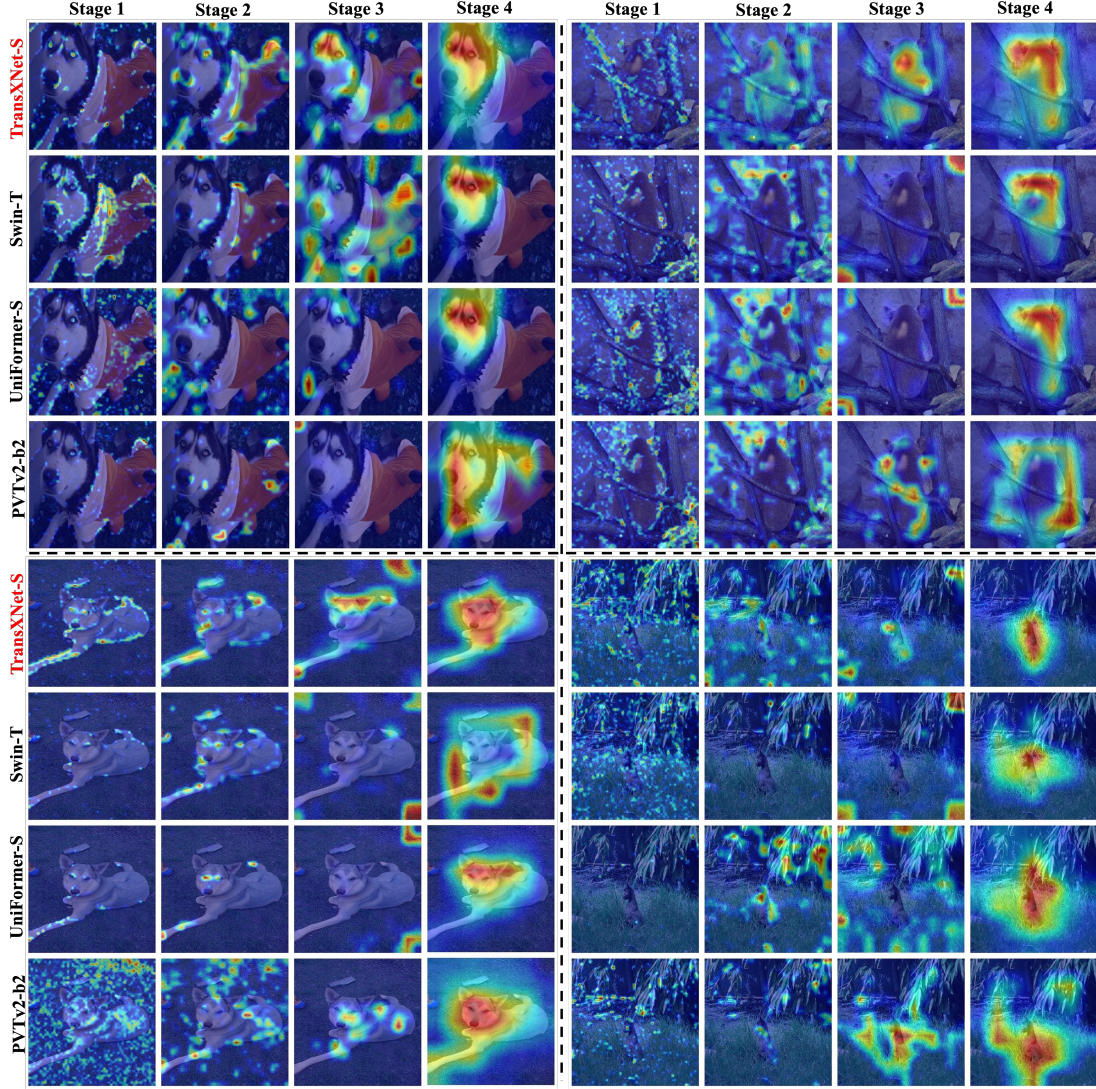
Fig. 7. Grad-CAM visualization of the models trained on ImageNet-1K. The visualized images are randomly selected from the validation set.
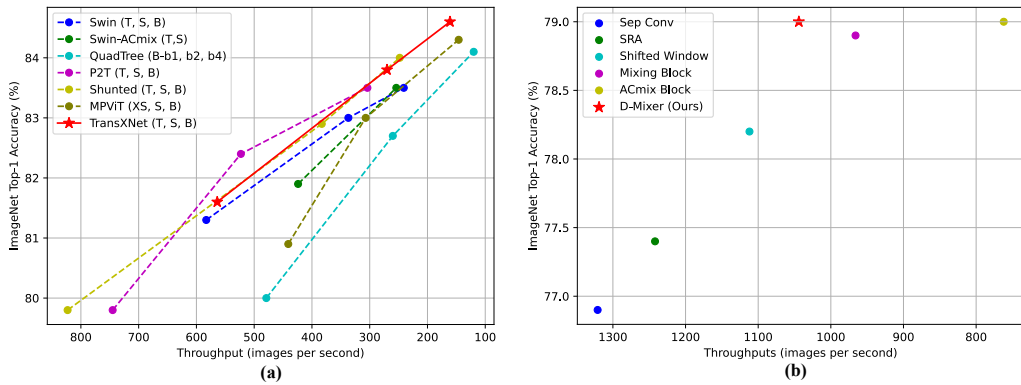

Fig. 8. Comparison of throughput of (a) representative backbone networks and (b) networks with different token mixers.

and reduce computational cost. Regarding the model design, our TransXNet series are manually stacked, and there exist potentially inefficient operators in the building blocks (e.g., multi-scale depthwise convolutions in MS-FFN). As a result, our model exhibits limited advantages in terms of speed when compared to other models with similar GFLOPs. Nonetheless, these inefficiencies can be mitigated through techniques such as Neural Architecture Search (NAS) [54] and specialized implementation engineering, which we plan to explore in our future work.

VI. Conclusion

In this work, we propose an efficient Dual Dynamic Token Mixer (D-Mixer), taking advantage of hybrid feature extraction provided by Overlapping Spatial Reduction Attention (OSRA) and Input-dependent Depthwise Convolution (IDConv). By stacking D-Mixer-based blocks to a deep network, the kernels in IDConv and attention matrices in OSRA are dynamically generated using both local and global information gathered in previous blocks, empowering the network with a stronger representation capacity by incorporating strong inductive bias and an expanded effective receptive field. Besides, we introduce an MS-FFN to explore multi-scale token aggregation in the feed-forward network. By alternating D-Mixer and MS-FFN, we construct a novel hybrid CNN-Transformer network termed TransXNet, which has shown SOTA performance on various vision tasks.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[3] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[4] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[5] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 815–825.

[6] Q. Chen, Q. Wu, J. Wang, Q. Hu, T. Hu, E. Ding, J. Cheng, and J. Wang, "Mixformer: Mixing features across windows and dimensions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5249–5259.

[7] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 175–12 185.

[8] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European Conference on Computer Vision*. Springer, 2022.

[9] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[10] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unifying convolution and self-attention for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[13] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[15] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 963–11 975.

[16] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, M. Pechenizkiy, D. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," in *International Conference on Learning Representations*, 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 124–12 134.

[19] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-transformer: Hierarchical vision transformer with local self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[20] S. Ren, D. Zhou, S. He, J. Feng, and X. Wang, "Shunted self-attention via multi-scale token aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 853–10 862.

[21] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[22] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2t: Pyramid pooling transformer for scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] H. Zhang, W. Hu, and X. Wang, "Fcaformer: Forward cross attention in hybrid vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6060–6069.

[24] N. Li, Y. Chen, W. Li, Z. Ding, D. Zhao, and S. Nie, "Bvit: Broad attention-based vision transformer," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[25] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 392–30 400, 2021.

[26] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 934–12 949, 2022.

[27] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3562–3572.

[29] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 030–11 039.

[30] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depth-wise convolution," in *International Conference on Learning Representations*, 2022.

[31] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[32] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision

transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.

[33] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[35] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[37] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European Conference on Computer Vision*. Springer, 2016, pp. 646–661.

[38] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[39] C. Yang, S. Qiao, Q. Yu, X. Yuan, Y. Zhu, A. Yuille, H. Adam, and L.-C. Chen, "Moat: Alternating mobile convolution and attention brings strong vision models," in *International Conference on Learning Representations*, 2023.

[40] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[41] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 428–10 436.

[42] S. Tang, J. Zhang, S. Zhu, and P. Tan, "Quadtree attention for vision transformers," in *International Conference on Learning Representations*, 2022.

[43] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "Mpvit: Multi-path vision transformer for dense prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7287–7296.

[44] R. Yang, H. Ma, J. Wu, Y. Tang, X. Xiao, M. Zheng, and X. Li, "Scalablevit: Rethinking the context-oriented generalization of vision transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 480–496.

[45] W. Wang, L. Yao, L. Chen, B. Lin, D. Cai, X. He, and W. Liu, "Crossformer: A versatile vision transformer hinging on cross-scale attention," in *International Conference on Learning Representations*, 2022.

[46] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

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

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

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[50] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2998–3008.

[51] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[52] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6399–6408.

[53] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," in *International Conference on Learning Representations*, 2022.

[54] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.