FACTORIZATION TRANSFORMER: MODELING LONG RANGE DEPENDENCY WITH LOCAL WINDOW COST

Anonymous authors
Paper under double-blind review

ABSTRACT

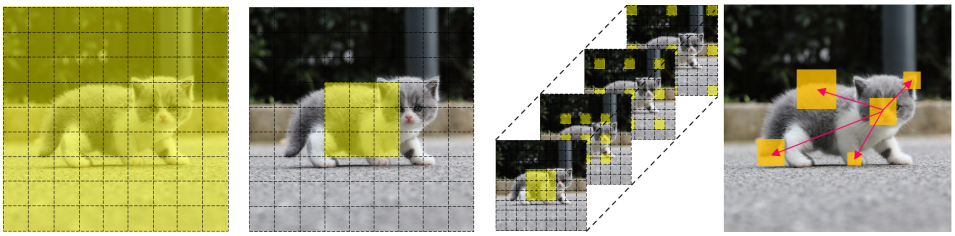
Transformers have astounding representational power but typically consume considerable computation. The current popular Swin transformer reduces the computational cost via a local window strategy. However, this inevitably causes two drawbacks: i) the local window-based self-attention loses global dependency modeling capability; ii) recent studies point out that the local windows impair robustness. This paper proposes a novel factorization self-attention mechanism (FaSA) that enjoys both the advantages of local window cost and long-range dependency. Specifically, we factorize a large area of feature tokens into non-overlapping subsets and obtain a strictly limited number of key tokens enriched of long-range information through cross-set interaction. Equipped with a new mixed-grained multi-head attention that adjusts the granularity of key features in different heads, FaSA is capable of modeling long-range dependency while aggregating multi-grained information at a computational cost equivalent to the local window-based self-attention. With FaSA, we present a family of models named factorization vision transformer (FaViT). Extensive experiments show that our FaViT achieves state-of-the-art performance on both classification and downstream tasks, while demonstrating strong robustness to corrupted and biased data. Compared with Swin-T, our FaViT-B2 significantly improves classification accuracy by 1% and robustness by 6%, and reduces model parameters by 14%. Our code will soon be publicly available on https://github.com/anonymous0519/DeViT.

1 INTRODUCTION

Since the success of Alexnet (Krizhevsky et al., 2012), revolutionary improvements have been achieved via advanced training recipes (He et al., 2019). With AutoML (Zoph & Le, 2017), CNNs have achieved state-of-the-art performance across various vision tasks (Tan & Le, 2019). On the other hand, recently popular transformers have shown superior performance over previous dominated CNNs (Daquan et al., 2020; Gao et al., 2021). The key difference between transformers and CNNs lies in the modeling of long-range dependency. In a vision conventional transformer, the images are divided into a sequence of patches, then processed in parallel, making the computational cost quadratic with the input image resolution. As a result, transformers consume significantly higher cost when compared to CNNs.

Several methods have been proposed to reduce the cost issue. Such as Swin transformer (Liu et al., 2021), where the tokens are divided into several windows, and the self-attention calculation is constrained within the predefined windows. Consequently, the computational cost becomes quadratic with the window size, which is set to be significantly smaller than input resolution. However, the local window inevitably impairs the long-range dependency modeling capability. As shown in Figure 1(b), Swin only calculates the relationship within a local area and loses the long-range dependency. Additionally, the damage of this strategy to model robustness has been demonstrated by recent studies Zhou et al. (2022).

The trade-off between the computational cost and the capability of modeling long-range dependency thus becomes a fundamental problem yet to be explored. We propose a novel self-attention mechanism termed factorization self-attention (FaSA) in this work. Specifically, given an input image, we take each point as a query. For gathering keys, we evenly divide the image into a series of non-overlapping local windows, uniformly sample a fixed number of points in each window through

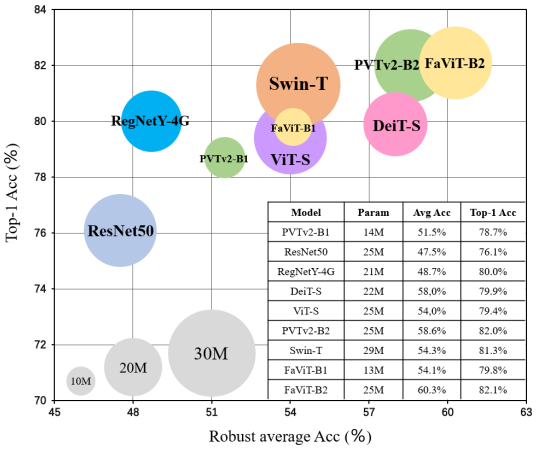


(a) ViT (b) Swin (c) FaSA

Figure 1: Comparison of ViT, Swin, and our FaSA. a) ViT enjoys global attention span but is computationally intensive; b) Swin is efficient but inferior in long-rang dependency modeling; c) FaSA factorizes tokens and hence successfully models long-range dependency at local window cost.

dilated sampling, and fuse the features of sampled points at the same position in different windows. Since the number of keys is strictly limited and each key contains information from multiple windows spanned across the whole image, attending to such a set of keys enables modeling long-range dependency at the local window cost.

Considering that each obtained key fuses multi-point information and hence easily lacks fine-grained details, we further introduce a mixed-grained multi-head attention. Concretely, we gradually increase the local window size in different head groups, and adaptively adjust the dilation rates for point sampling to keep the number of keys in different head groups the same. As a result, the obtained keys fuse features at fewer locations and hence with more fine-grained details. By fusing the attended features from multiple head groups, the long-range and mixed-grained information can be obtained simultaneously.



Model	Param	Avg Acc	Top-1 Acc
PVTv2-B1	14M	51.5%	78.7%
ResNet50	25M	47.5%	76.1%
RegNetY-4G	21M	48.7%	80.0%
DeiT-S	22M	58.0%	79.9%
ViT-S	25M	54.0%	79.4%
PVTv2-B2	25M	58.6%	82.0%
Swin-T	29M	54.3%	81.3%
FaViT-B1	13M	54.1%	79.8%
FaViT-B2	25M	60.3%	82.1%

Figure 2: Comparison of Accuracy-robustness trade-off. Our FaViT achieves the best performance in both accuracy and robustness, with fewer parameters (indicated by the circle size).

Based on FaSA, we design a family of models termed factorization vision transformer (FaViT). With the aid of FaSA, FaViT enjoys two essential advantages that are not possible with previous transformers. First, as each local window contains an equal number of tokens, its computational cost is fixed. The long-range dependency is captured with no additional overhead. Secondly, mixed subgroup attention matrix aggregates multi-grained information. Extensive experiments demonstrate the state-of-the-art performance and superior robustness of our FaViT. As depicted in Figure 2, FaViT achieves high classification accuracy and robustness at a similar model size. Notably, compared with the baseline model Swin, FaViT-B2 outperforms Swin-T in all aspects. The robustness significantly improves by 6%, and the classification accuracy improves by 0.8%, while the parameters drop by a considerable 14%. Furthermore, FaViT also exhibits state-of-the-art performance on object detection and semantic segmentation downstream tasks. To sum up, this work makes the following contributions:

- We propose a novel factorization self-attention mechanism, which is capable of modeling long-range dependency while aggregating multi-grained information at local window cost.
- Based on FaSA, we present an efficient factorization vision transformer (FaViT), which exhibits state-of-the-art accuracy and robustness.

2 Related Work

Vision transformers Transformer is originally developed for NLP tasks (Brown et al., 2020; Devlin et al., 2019),but has now achieved great success in several fields (Lee et al., 2019; Cui et al., 2022;

Vasilescu et al., 2021; Ainslie et al., 2020). ViT Dosovitskiy et al. (2021) is proposed for the first time in computer vision. It splits the image into a sequence of tokens for encoding and constructs a convolution-free network structure through self-attention. DeiT Touvron et al. (2021) introduces a series of training strategies to make ViT work on smaller dataset ImageNet-1K replace large-scale JFT-300M. In addition, some works Jiang et al. (2021); Ding et al. (2022); Tu et al. (2022) consummate the ViT and achieve better performance. The HaloNet Vaswani et al. (2021) introduces the idea of a slightly larger window of key than the query. A CrossViT Chen et al. (2021) comes up with a dual branch approach with multi-scale patch size. The above methods verify the feasibility of the transformer-based structure.

Efficient variants A series of methods have recently been proposed to reduce the high cost of ViT, and they can be divided into two strategies. 1. Global fusion tokens self-attention; 2. Local window-based self-attention. PVTv1 Wang et al. (2021) is a representative model using the first strategy. It introduces spatial reduction and tokens fusion to reduce the cost of multi-head attention. The subsequently proposed PVTv2 Wang et al. (2022) improves it by introducing overlapping patch embedding, depth-wise convolution (Chollet, 2017), and linear SRA. A shunted self-attention Ren et al. (2021) comes up to unify multi-scale feature extractions via multi-scale token aggregation. The key point of this strategy is to reduce cost through spatial compression. Nevertheless, the information of small targets and delicate textures will be overwhelmed, destroying fine-grained features.

Local self-attention The local window-based self-attention strategy is to divide sub-regions and only apply self-attention within each. The most representative model among them is Swin (Liu et al., 2021). It introduces a hierarchical sliding window structure and shifts operations across sub-regions. MOA Patel et al. (2022) exploits the neighborhood and global information among all non-local windows. SimViT Li et al. (2022a) integrates spatial structure and cross-window connections of sliding windows into the visual transformer. This strategy significantly reduces computational cost, but its drawbacks cannot be ignored. Lack of long-range dependency limits modeling power, while excessive windowing of tokens reduces robustness. Previous works are unable to restore dependencies while maintaining costs. Therefore, we propose the FaViT for long-range modeling dependency with local window-based self-attention computational cost.

3 METHOD

3.1 PRELIMINARIES

Given an input feature map $\mathbf{X} \in \mathbb{R}^{h \times w \times c}$, the conventional self-attention (SA) used in ViT applies linear embeddings with parameters $\mathbf{W}^{\mathrm{K}}, \mathbf{W}^{\mathrm{Q}}, \mathbf{W}^{\mathrm{V}}$ to embed all the points into key $\mathbf{K} = \mathbf{W}^{\mathrm{K}}\mathbf{X}$, query $\mathbf{Q} = \mathbf{W}^{\mathrm{Q}}\mathbf{X}$, and value $\mathbf{V} = \mathbf{W}^{\mathrm{V}}\mathbf{X}$, respectively. It then performs self-attention as

$$\mathrm{SA}(\mathbf{X}) = \mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{h}})\mathbf{V}, \tag{1}$$

where $\sqrt{h}$ is a scaling factor. Therefore, the computational complexity of ViT can be computed as

$$\Omega(ViT) = 4hwc^2 + 2(hw)^2c, \tag{2}$$

which grows quadratically as the input feature map size increases.

Swin transformer adopts local window-based self-attention (WSA) to reduce the computational cost by evenly partitioning $\mathbf{X}$ into non-overlapping windows and performing self-attention calculation within each window locally. Suppose each window contains $m \times m$ tokens, the computational complexity of Swin is:

$$\Omega(Swin) = 4hwc^2 + 2m^2hwc, \tag{3}$$

which is linear to the input resolution given a fixed m. SA enjoys global attention span but is computationally intensive for high input resolution. In contrast, WSA is more efficient but inferior in modeling long-range dependency, which impairs performance and model robustness. The above limitations drive us to explore a new self-attention mechanism to model long-range dependency at local window cost.

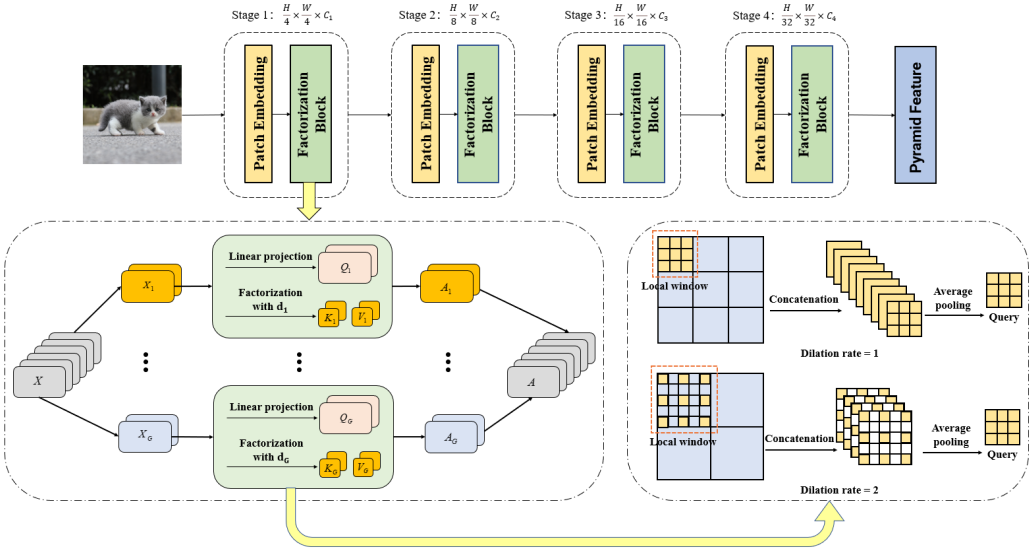


Figure 3: Overall architecture of FaViT. We divide the input image into local windows and model long-range dependency through cross-window interaction.

3.2 FACTORIZATION SELF-ATTENTION

Figure 3 illustrates the overall architecture of the proposed factorization self-attention mechanism (FaSA). We first uniformly divide the input feature map $\mathbf{X} \in \mathbb{R}^{h \times w \times c}$ into multiple groups. Each group independently runs the self-attention mechanism to acquire long-range features at a specific granularity level. Regarding the self-attention in each group, we take all the points of the feature map as query tokens and gather key tokens in three steps: *i) factorization*, evenly divide the entire feature map into a series of local windows; *ii) dilated sampling*, uniformly sample a fixed number points in each local window; *iii) cross-window fusion*, fuse the features of the sampled points at the same position in different windows. Hence the resulting fused key tokens contain long-range information from multiple windows spanned across the whole feature map.

While the fusion of multi-point information can easily lead to the loss of fine-grained details in the obtained key tokens. We hence introduce mixed-grained multi-head attention by gradually enlarging the local window size in different head groups. To keep the number of sampled points unchanged, we also adaptively increase the corresponding dilation rate for point sampling. The resulting key token fuses features from fewer points and thus has more fine-grained information. By fusing the attended features from multiple head groups, the proposed FaSA models long-range dependency while aggregating multi-grained information at a computational cost equivalent to the local window-based self-attention. Next, we elaborate on each step.

Head grouping. Given the input feature map $\mathbf{X}$, we first uniformly divide it into multiple groups along the channel dimension.

$$\mathbf{X} = \left\{ \mathbf{X}_i \in \mathbb{R}^{h \times w \times c'}, | \; i = 1, \cdots, \mathrm{G} \right\}, \tag{4}$$

where $c' = c/G$. We take the divided features as the input of different attention head groups, in which factorization self-attention is performed independently to capture long-range information with different granularities.

Gathering queries. We gather the query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) in each attention head group individually. For the i-th head group, we take each point of $\mathbf{X}_i$ as a query and obtain query features as

$$\mathbf{Q}_i = \mathbf{W}_i^{\mathrm{Q}} \mathbf{X}_i \in \mathbb{R}^{h \times w \times c'}, \tag{5}$$

where $\mathbf{W}_i^{\mathrm{Q}} \in \mathbb{R}^{c \times c'}$ is a learnable linear embedding implemented by 1×1 convolution.

Gathering keys. The acquisition of the keys largely determines the attention span and computational cost of self-attention. To model long-range dependency while retaining the local window cost, we gather the keys in the following three steps.

Factorization. We first uniformly divide $\mathbf{X}_i$ into multiple non-overlapping local windows. For simplicity, we consider $h = w$, such that the length and width of each local window are equal, denotes as s_i.

$$\mathbf{X}_i = \left\{ \mathbf{X}_i^j \in \mathbb{R}^{s_i \times s_i \times c'}, | \, j = 1, \cdots, M_i \right\}. \tag{6}$$

We gradually enlarge the size of local window across different head groups, resulting in decreasing M_i.

Dilated sampling. We next uniformly sample a fixed number n points in each local window. For the i-th head group, the sampled point set is

$$\mathbf{P}_i = \left\{ \mathbf{P}_i^j \in \mathbb{R}^{\mathrm{n} \times c'}, | \, j = 1, \cdots, M_i \right\} \in \mathbb{R}^{M_i \times n \times c'}. \tag{7}$$

Since different head groups have distinct local window sizes, in order to keep the number of points sampled from the local window of different head groups the same, we apply dilated sampling with increasing dilation rate across head groups. The dilation rate for the i-th head group is computed as

$$d_i = (s_i - 1)/(\sqrt{n} - 1). \tag{8}$$

Hence, the sampled points are uniformly distributed inside each local window. As the head group index i increases, the interval between the sampled points become larger.

Cross-window fusion. Previous studies have proved that computing self-attention only inside a local window impairs the modeling ability and robustness due to the lack of cross-window information interaction Liu et al. (2021). To enhance cross-window interaction and at the same time reduce the number of key tokens, we introduce a novel cross-window fusion strategy. Specifically, we first perform feature embedding for each sampled point

$$\begin{aligned} \mathbf{K}_i' &= \mathbf{W}_i^{\mathrm{K}} \mathbf{P}_i \in \mathbb{R}^{M_i \times n \times c'}, \\ \mathbf{V}_i' &= \mathbf{W}_i^{\mathrm{V}} \mathbf{P}_i \in \mathbb{R}^{M_i \times n \times c'}, \end{aligned} \tag{9}$$

where $\mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{c' \times c'}$ are learnable linear embeddings implemented by two separate 1×1 convolutions.

Next, we fuse the features of the sampled points at the same position in different windows to obtain the final key and value features.

$$\begin{aligned} \mathbf{K}_i &= \boldsymbol{\sigma}(\mathbf{K}_i') \in \mathbb{R}^{n \times c'}, \\ \mathbf{V}_i &= \boldsymbol{\sigma} \mathbf{V}_i' \in \mathbb{R}^{n \times c'}, \end{aligned} \tag{10}$$

where $\boldsymbol{\sigma}(\cdot)$ is a symmetric aggregation function. Hereby we implement it as a simple form, ***i.e.***, average pooling.

As a result, each fused feature is enriched of long-range information by combining the information of M_i points that are evenly distributed over the whole feature map. In addition, as the head group index i increases, M_i decreases such that the fused feature aggregates information at fewer locations and hence with more fine-grained details.

Mixed-grained multi-head attention. Given the gathered queries and keys for each head group, we perform self-attention individually.

$$\mathbf{X}_i' = Softmax(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{h_i}}) \mathbf{V}_i \in \mathbb{R}^{h \times w \times c'}, \tag{11}$$

where $\sqrt{h_i}$ is a scaling factor. We then combine the attended features from all the head groups to obtain the final output.

$$\mathbf{X}' = \boldsymbol{\delta}(\mathbf{X}_i', | i = 1, \cdots, G) \in \mathbb{R}^{h \times w \times c}, \tag{12}$$

where $\boldsymbol{\delta}(\cdot)$ denotes concatenation operation along the feature channel dimension. Consequently, FaSA is capable of modeling long-range dependency while aggregating multi-grained information.

Complexity analysis. The computational complexity of factorization self-attention can be calculated as:

$$\Omega(FaSA) = 4hwc^2 + 2nhwc. \tag{13}$$

Since n is a pre-set fixed value, the computational complexity of FaSA is linear w.r.t. the input image size. As a result, FaSA is capable of modeling long-range dependency at a computational cost equivalent to the local window-based self-attention.

3.3 ARCHITECTURE DETAILS

We present four model variants of FaViT with distinct model sizes by adopting different parameter settings. More details of the model architecture can be found in Table 1.

Table 1: Architecture variants for FaViT. P_i: patch size, H_i: number of heads, E_i: MLP expansion ratio, D_i: dilation rate set, C_i: output feature dimension, B_i: number of blocks.

FaViT-B0	FaViT-B1	FaViT-B2	FaViT-B3
$P_1 = 4; H_1 = 1; E_1 = 8; D_1 = [1, 8]$			
$C_1, B_1 = 32, 2$	$C_1, B_1 = 64, 2$	$C_1, B_1 = 64, 2$	$C_1, B_1 = 96, 2$
$P_2 = 2; H_2 = 2; E_2 = 6; D_2 = [1, 4]$			
$C_2, B_2 = 64, 2$	$C_2, B_2 = 128, 2$	$C_2, B_2 = 128, 3$	$C_2, B_2 = 192, 3$
$P_3 = 2; H_3 = 4; E_3 = 4; D_3 = [1, 2]$			
$C_3, B_3 = 128, 6$	$C_3, B_3 = 256, 6$	$C_3, B_3 = 256, 18$	$C_3, B_3 = 384, 14$
$P_4 = 2; H_4 = 8; E_4 = 4; D_4 = [1]$			
$C_4, B_4 = 256, 2$	$C_4, B_4 = 512, 2$	$C_4, B_4 = 512, 3$	$C_4, B_4 = 768, 3$

4 EXPERIMENTS

We evaluate the effectiveness and generalizability of FaViT on image classification, object detection, and semantic segmentation. We also test the robustness of FaViT by learning under image corruptions, label noise, and class imbalance. Ablation studies are provided to validate our design choices.

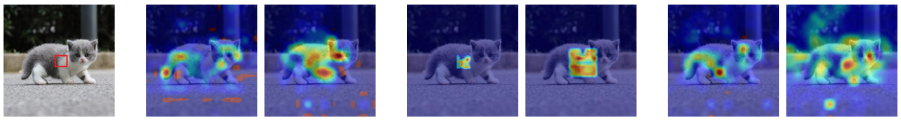
4.1 IMAGE CLASSIFICATION

Table 2: Classification results on ImageNet-1K. All models are trained from scratch using the same training strategies.

Model	#Param	FLOPs	Top-1 Acc (%)
PVTv2-B0 (Wang et al., 2022)	3.7M	0.6G	70.5
FaViT-B0	**3.4M**	**0.6G**	**71.5**
ResNet18 (He et al., 2016)	**11.7M**	**1.8G**	69.8
PVTv1-T (Wang et al., 2021)	13.2M	1.9G	75.1
PVTv2-B1 (Wang et al., 2022)	14.0M	2.1G	78.7
FaViT-B1	13.0M	2.4G	**79.4**
ConvNeXt-T (Liu et al., 2022)	29.0M	4.5G	82.1
DeiT-S (Touvron et al., 2021)	**22.1M**	4.6G	79.9
T2T-ViT-14 (Yuan et al., 2021)	22.0M	5.2G	81.5
ViL-S (Zhang et al., 2021)	24.6M	5.1G	82.0
CrossViT-15 (Chen et al., 2021)	27.4M	5.8G	81.5
TNT-S (Han et al., 2021)	23.8M	5.2G	81.5
DW-ViT-T (Ren et al., 2022)	30.0M	5.2G	82.0
Swin-T (Liu et al., 2021)	29.0M	4.5G	81.3
FaViT-B2	24.9M	**4.5G**	**82.1**
ConvNeXt-S (Liu et al., 2022)	50.0M	8.7G	83.1
Focal-S (Yang et al., 2021)	51.1M	9.4G	**83.6**
SimViT-M (Li et al., 2022a)	51.3M	10.9G	83.3
MOA-S (Patel et al., 2022)	**39.0M**	9.4G	83.5
PVTv2-B3 (Wang et al., 2022)	45.2M	**6.9G**	83.2
Swin-S (Liu et al., 2021)	50.0M	8.7G	83.0
FaViT-B3	48.8M	8.5G	83.4

Main results. We test FaViT on classification dataset ImageNet-1K (Russakovsky et al., 2015). For a fair comparison with priors Wang et al. (2022), we train the model for 300 epochs using AdamW optimizer with an initial learning rate 0.001. Table 2 shows that our lightweight FaViT-B0 and FaViT-B1 achieve 71.5% and 79.4% accuracy, respectively, outperforming previous models with similar number of parameters. Compared to Swin transformer baseline, our FaViT model family exhibits better performance with fewer parameters and FLOPs. In particular, FaViT-B2 achieves a high accuracy of 82.1%, which surpasses Swin-T counterpart by a large margin of 1% while significantly reducing the parameters by 14%. The superior accuracy and efficiency achieved by FaViT should be credited to its capability of modeling long-range dependency with local window cost.

Visualization. FaViT achieves high classification accuracy due to its attention span across both short and long ranges. We visualize the attention spans for a given token area from the first and the second stages of different models in Fig. 4. The Swin transformer baseline focuses only on the small neighborhood of the token, which leads to degraded accuracy and robustness. In comparison, ViT has longer-range attention span owing to the adopted global attention mechanism, but at the cost of quadratically increased computational complexity with the increase of input image size. Our FaViT well takes the complementary strengths of both models. It enjoys a large attention span similar to that of ViT while with only local a local window cost as Swin transformer, achieving an ideal trade-off between dependency modeling capability and computational cost.



(a) Token area (b) ViT (c) Swin (d) FaViT

Figure 4: Visualizing the attention span for a given token area (red box) by different models.

Efficiency. We compare the computational complexity of our FaViT with other priors in Fig. 5 by reporting the number of FLOPs under varying input image sizes. When the input image size is small,

Table 3: Object detection and instance segmentation results on COCO 2017. All models are pre-trained on ImageNet-1K and fine-tuned with 1× schedule.

Model	#Param	RetinaNet 1× AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#Param	Mask R-CNN 1× AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
PVTv2-B0 (Wang et al., 2022)	13.0M	37.2	57.2	39.5	**23.1**	40.4	**49.7**	23.5M	**38.2**	**60.5**	40.7	**36.2**	**57.8**	**38.6**
FaViT-B0	**12.7M**	**37.4**	**57.2**	**39.8**	22.9	**40.6**	49.5	**23.2M**	37.9	59.6	**41.0**	35.4	56.6	37.8
ResNet18 (He et al., 2016)	**21.3M**	31.8	49.6	33.6	16.3	34.3	43.2	**31.2M**	34.0	54.0	36.7	31.2	51.0	32.7
PVTv2-B1 (Wang et al., 2022)	23.8M	41.2	**61.9**	43.9	25.4	44.5	54.3	33.7M	41.8	64.3	45.9	38.8	61.2	41.6
FaViT-B1	22.7M	**41.4**	61.8	**44.0**	**25.7**	**44.9**	**55.0**	32.6M	**42.4**	**64.4**	**46.3**	**38.9**	**61.2**	**41.8**
Twins-S (Chu et al., 2021)	**34.4M**	43.0	64.2	46.3	28.0	46.4	57.5	**44.0M**	43.4	66.0	47.3	40.3	63.2	43.4
Swin-T (Liu et al., 2021)	38.5M	41.5	62.1	44.2	25.1	44.9	55.5	47.8M	42.2	64.6	46.2	39.1	61.6	42.0
FaViT-B2	34.6M	**44.4**	**65.0**	**47.7**	**27.7**	**48.2**	**58.8**	44.6M	**45.4**	**67.1**	**49.4**	**41.0**	**64.0**	**44.1**
PVTv1-M (Wang et al., 2021)	**53.9M**	41.9	63.1	44.3	25.0	44.9	57.6	**63.9M**	42.0	64.4	45.6	39.0	61.6	42.1
Swin-S (Liu et al., 2021)	59.8M	44.5	65.7	47.5	27.4	48.0	59.9	69.1M	44.8	66.6	48.9	40.9	63.4	44.2
FaViT-B3	58.9M	**46.0**	**66.7**	**49.1**	**28.4**	**50.3**	**62.0**	68.3M	**47.1**	**68.0**	**51.4**	**42.7**	**65.9**	**46.1**

e.g., 224×224, all the models have comparable number of FLOPs. As the input image size increases, DeiT suffers a dramatic rise in FLOPs. Similarly, PVTv2 also exhibits a quadratic increase. In comparison, the number of FLOPs for our FaViT grows linearly with the increase of image size owning to the adopted local window strategy. The superiority in computational complexity of our FaViT is highlighted when the input image is large. Specially, when the image size is 1000×1000, FaViT-B2 has only 25% and 50% FLOPs compared to Diet-S and PVTv2-B2, respectively. Compared to Swin-T, our FaViT-B2 has the same computational cost but achieves higher accuracy.

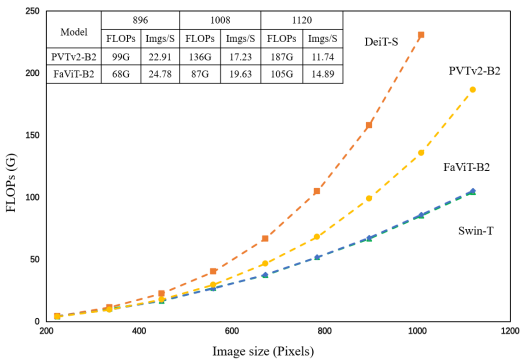


Figure 5: FLOPs w.r.t. input size. The FLOPs and Throughputs when the input image size is 224×224 are listed at the upper left corner.

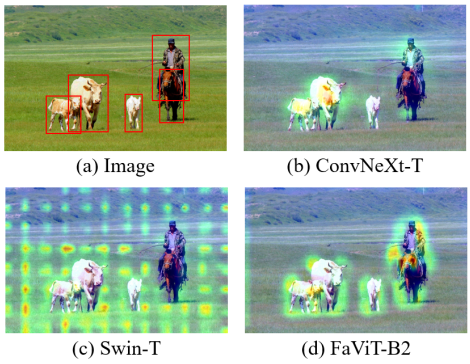


Figure 6: Visualizing attention maps for detecting multiple objects.

4.2 Object Detection

Settings and results. We evaluate FaViT for object detection and instance segmentation on COCO2017 (Lin et al., 2014) using RetinaNet Lin et al. (2020) and Mask R-CNN He et al. (2020), respectively. We pretrain FaViT on ImageNet-1K and fine-tune it for 12 epochs with initial learning rate 0.0001. Table 3 shows that under the respective size levels, FaViT-B0 to B3 all exhibit state-of-the-art performance. Compared with Swin baseline, FaViT benefits from the long-range dependency and achieves higher accuracy on both two detector. Notably, FaViT-B2 has stronger processing power on objects of various scales, such as AP and AP_L are improved by 2.9% and 3.3%, respectively. For instance segmentation, FaViT-B2 also achieves higher accuracy and significantly improves AP^m by 1.9%.

Visualization. Due to the long-range dependency, FaViT enjoys advantages in detecting various scale objects. We draw attention to heatmaps of FaViT-B2 under the multi-object challenge in Figure 6. Furthermore, compare it with ConvNeXt-T and Swin-T of similar model sizes. The ability of ConvNeXt-T to focus on multiple objects simultaneously is weak, and individual targets will be lost. Swin's attention heatmap has an apparent grid shape, which is relatively disorganized. By contrast, FaViT's attention is more evenly distributed over the entire object. We can observe the

Table 5: Robustness against image corruptions on ImageNet-C (%).

Model	#Param	Retention	Blur				Noise				Digital				Weather			
			Motion	Fafoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Contr	Satur	JEPG	Pixel	Bright	Snow	Fog	Frost
Mobile Setting (<15M)																		
MobileNetV2	4M	49.2	33.4	29.6	21.3	32.9	24.4	21.5	23.7	32.9	57.6	49.6	38.0	62.5	28.4	45.2	37.6	28.3
EfficientNet-B0	5M	54.7	36.4	26.8	**26.9**	**39.3**	**39.8**	**38.1**	**47.1**	39.9	**65.2**	58.2	**52.1**	**69.0**	37.3	**55.1**	44.6	37.4
PVTv2-B0	3M	58.9	30.8	24.9	24.0	35.8	33.1	35.2	44.2	50.6	59.3	50.8	36.6	61.9	38.6	50.7	45.9	41.8
ResNet18	**11M**	32.8	29.6	28.0	22.9	32.0	22.7	17.6	20.8	27.7	30.8	52.7	46.3	42.3	58.8	24.1	41.7	28.2
PVTv2-B1	13M	65.4	45.7	41.3	30.5	43.9	48.1	46.2	46.6	55.0	57.6	68.6	59.9	50.2	71.0	49.8	56.8	53.0
FaViT-B0	**3M**	**59.2**	**38.1**	**31.6**	24.8	37.4	38.3	35.6	39.9	**45.2**	47.9	**60.8**	51.6	38.9	**63.2**	38.5	**44.6**	**42.5**
FaViT-B1	13M	**68.1**	**48.2**	**43.2**	**30.7**	**45.6**	**53.8**	**52.4**	**52.6**	**58.7**	**59.6**	**70.1**	**61.7**	**53.5**	**72.1**	**50.9**	**57.1**	**54.7**
GPU Setting (20M+)																		
ResNet50	25M	62.5	42.1	40.1	27.2	42.2	42.2	36.8	41.0	50.3	51.7	69.2	59.3	51.2	71.6	38.5	53.9	42.3
ViT-S	25M	67.6	49.7	45.2	**38.4**	48.0	50.2	47.6	49.0	57.5	58.4	70.1	61.6	57.3	72.5	51.2	50.6	57.0
DeiT-S	**22M**	72.6	52.6	48.9	38.1	51.7	57.2	55.0	54.7	60.8	63.7	71.8	64.0	58.3	73.6	55.1	61.1	60.7
PVTv1-S	25M	66.9	54.3	48.4	34.7	46.4	51.7	51.7	50.0	55.8	57.6	69.4	60.7	53.7	49.5	**62.3**	55.2	53.1
PVTv2-B2	25M	71.5	54.3	48.4	34.7	50.7	61.2	60.7	59.5	**64.5**	**65.5**	**73.5**	65.5	58.8	**75.2**	56.7	**67.8**	**62.7**
Swin-T	29M	66.8	49.5	45.0	31.7	47.6	54.7	51.6	52.6	58.4	62.1	71.4	62.2	54.4	73.4	60.0	64.7	60.2
FaViT-B2	25M	**73.4**	**55.9**	**49.8**	34.8	**51.7**	**62.6**	**62.1**	**61.2**	65.3	64.9	73.3	**66.1**	**64.8**	75.0	55.3	62.9	59.5

position and outline of multiple objects simultaneously. The comparison of heatmap visualizations proves the superiority of FaViT.

4.3 Semantic segmentation

We test our FaViT for semantic segmentation on ADE20K (Zhou et al., 2018). We evaluate different backbone models using Semantic FPN Kirillov et al. (2019) framework and the same fine-tuning strategy Liu et al. (2021). Table 4 depicts that our FaViT families of different model sizes consistently perform better than their corresponding Swin transformer counterparts. Noticeably, FaViT-B2 outperforms Swin-T by a large margin of 3.5% in mIoU, with fewer parameters and FLOPs.

Table 4: Segmentation results on ADE20K.

Model	#Param	FLOPs	mIoU (%)
PVTv2-B0 (Wang et al., 2022)	7.6M	25.0G	37.2
FaViT-B0	**7.3M**	**24.6G**	**37.2**
ResNet18 (He et al., 2016)	**15.5M**	**32.2G**	32.9
EFormer-L1 (Li et al., 2022b)	16.0M	33.0G	38.9
FaViT-B1	16.7M	33.9G	**42.0**
Twins-S (Chu et al., 2021)	**28.3M**	**37.5G**	43.2
Swin-T (Liu et al., 2021)	31.9M	46.0G	41.5
FaViT-B2	28.7M	45.2G	**45.0**
CAT-B (Lin et al., 2022)	55.0M	76.9G	44.9
Swin-S (Liu et al., 2021)	53.2M	70.0G	45.2
FaViT-B3	**52.4M**	**66.7G**	**47.2**

4.4 Robustness Analysis

We evaluate the robustness of FaViT to image corruptions, label noise, and class imbalance.

Robustness against image corruptions. We test on ImageNetC Hendrycks & Dietterich (2019) that comprises various corrupted images by introducing blur, natural noise, digital noise, and severe weather condition. All models are pre-trained on ImageNet-1K without further fine-tuning Zhou et al. (2022). Table 5 indicates that our FaViT exhibits stronger robustness under both Mobile and GPU settings compared to CNN and transformer based priors. To reduce the impact of the model representation, we propose a new metric named accuracy retention, which is the ratio between the accuracy on the corrupted images and the clean images. This metric reflects how much of the accuracy could be preserved when testing on corrupted images and the accuracy on the clean images are thus normalized. Noticeably, FaViT-B2 significantly surpasses Swin-T by 6.6% in accuracy retention and performs better when faced with almost all types of image corruptions. The results indicate that FaViT successfully gathers long-range contextual information which is critical to improving the robustness against image corruptions.

Robustness against label noise. We test the robustness against real-world label noise on Clothing1M Xiao et al. (2015) and WebVision Li et al. (2017). For fair comparison, we use the same training strategy Li et al. (2020) for all models. Table 6 illustrates that our FaViT-B2 achieves the best accuracy on both datasets. Specially, FaViT-B2 achieves a top-1

Table 6: Robustness against label noise on Clothing1M and Webvision.

Model	#Param	Clothing1M	Webvision	
		Test Acc (%)	Top-1 Acc (%)	Top-5 Acc (%)
PVTv2-B2	24.5M	69.89	65.28	85.72
Shunted-S	**22.4M**	70.04	67.44	**86.24**
Swin-T	29.0M	69.12	60.84	82.48
FaViT-B2	24.9M	**70.82**	**67.72**	85.80

Table 8: Ablation studies on CIFAR100.

(a) Apply to other frameworks.

Model	#Param	FLOPs	Acc (%)
Swin-T	28.3M	4.36G	61.8
Swin-FaSA-T	**26.2M**	**4.0G**	**66.1**
PVTv2-B0	3.7M	0.6G	63.5
PVTv2-FaSA-B0	**3.1M**	**0.6G**	**64.0**
PVTv2-B1	14.0M	2.1G	70.8
PVTv2-FaSA-B1	**11.5M**	**2.1G**	**73.1**

(b) Impact of dilation rate.

Model	#Param	FLOPs	Acc (%)
FaSA-low	3.4M	0.6G	64.9
FaSA-high	3.4M	0.6G	64.5
FaSA	3.4M	0.6G	**65.2**

(c) Imapct of global features.

Model	Acc (%)	Param w.r.t image size			
		448	672	896	1120
FaViT-B2$_0$	75.7	18G	**41G**	**72G**	**113G**
FaViT-B2$_{1/4}$	75.7	18G	43G	80G	135G
FaViT-B2$_{1/8}$	**75.8**	**18G**	42G	76G	124G
FaViT-B2$_{1/16}$	75.8	18G	41G	74G	118G

accuracy of 67.72% on WebVision, significantly outperforming the Swin-T baseline by 6.9%. The results clearly evidence the strong robustness of our FaViT to real-world label noise.

Robustness against class imbalance. We test the robustness against class imbalance on the long-tailed iNaturalist dataset (Horn et al., 2018) in Table 7.All the models are pre-trained on ImageNet-1K and fine-tuned for 100 epochs with initial learning rate 0.0001. Our FaViT-B1 achieves the best accuracy compared to other priors, demonstrating good robustness when learning form long-tailed data. The above results prove that our FaViT enjoys strong robustness to both data corruptions and biases and hence demonstrates good promise to benefit real applications. Table 7 shows that our FaViT performs close to the state-of-the-art at similar FLOPs.

Table 7: Robustness against class imbalance on iNaturalist 2018.

Model	FLOPs	Top-1 Acc (%)
ResMLP-12 (Touvron et al., 2022)	3.0G	60.2
Inception-V3 (Horn et al., 2018)	2.5G	60.2
LeViT-192 (Graham et al., 2021)	**0.7G**	60.4
ResNet-50 (Cui et al., 2019)	4.1G	**64.1**
FaViT-B1	2.4G	64.2

4.5 Ablation Study

We perform ablation studies on CIFAR100 (Krizhevsky, 2009). All model variants are trained from scratch for 100 epochs with initial learning rate 0.001.

Effectiveness of FaSA. We evaluate the effectiveness and generalizability of the proposed factorization self-attention mechanism by applying it to other transformer based backbones. Concretely, we simply replace the original self-attention in Swin transformer and PVTv2 with FaSA while keeping other network structures unchanged. Table 8(a) shows that our FaSA consistently improves various backbones and at the same time reduces the number of parameters and FLOPs. In particular, it significantly improves the accuracy of Swin-T and PVTv2-B1 by 4.3% and 2.3%, respectively, and reduces the number of parameters of PVTv2-B1 by 18%. The results clearly evidence the superiority of our FaSA over other popular self-attention mechanisms.

Impact of dilation rate. In FaSA, we aggregate multi-grained information which captured by grouped features. We build two models that use only a single fine-grained information. FaSA-low represents that the dilation rate for each group is set to 1. The extracted query has the lowest fine-grained information. FaSA-high indicates that the local window is similar in size to the feature map. Table 8(b) shows that FaSA has the best performance.

Optimization structure with global features. Our FaSA introduces dilation rates to increase the local window size and model long-range but not global dependencies. We argue that introducing an appropriate amount of global features may help to improve model performance. We split part of channels and extract global features from it using the method in Wang et al. (2022). Table 8(c) shows introducing global features improves model performance while increases computational cost. When extracting global features from 1/8 channel and FaSA handles the rest, the performance and cost reach an ideal trade-off.

5 Conclusion

This paper proposes a novel factorization self-attention (FaSA) to explore the optimal trade-off between computational cost and the ability to model long-range dependency. We introduce a dilation rate set to implement the factorization operation. With the aid of FaSA, long-range dependencies will be modeled at the local window equivalent computational cost. Extensive experiments show that the proposed model achieves state-of-the-art performance and superior robustness.

REFERENCES

Joshua Ainslie, Santiago Ontañón, Chris Alberti, Vaclav Cvicek, Zachary Kenneth Fisher, Philip Pham, Anirudh Ravula, Sumit K. Sanghai, Qifan Wang, and Li Yang. Etc: Encoding long and structured inputs in transformers. In *EMNLP*, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 347–356, 2021.

François Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019.

Yutao Cui, Jiang Cheng, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13598–13608, 2022.

Zhou Daquan, Qibin Hou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Rethinking bottleneck structure for efficient mobile network design. In *ECCV*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Mingyu Ding, Bin Xiao, Noel C. F. Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. *ArXiv*, abs/2204.03645, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xinyu Zhang, Ming-Hsuan Yang, and Philip H. S. Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:652–662, 2021.

Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herv'e J'egou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12239–12249, 2021.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *NeurIPS*, 2021.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2020.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 558–567, 2019.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. In *NeurIPS*, 2021.

Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6392–6401, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 2019.

Gang Li, Di Xu, Xingyi Cheng, Lingyu Si, and Changwen Zheng. Simvit: Exploring a simple vision transformer with sliding windows. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022a.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *ArXiv*, abs/2002.07394, 2020.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *ArXiv*, abs/1708.02862, 2017.

Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, S. Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *ArXiv*, abs/2206.01191, 2022b.

Hezheng Lin, Xingyi Cheng, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Qing Song, and Wei Yuan. Cat: Cross attention in vision transformer. *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. 2022.

Krushi Patel, Andrés M. Bur, Fengju Li, and Guanghui Wang. Aggregating global features into local vision transformer. *ArXiv*, abs/2201.12903, 2022.

Pengzhen Ren, Changlin Li, Guangrun Wang, Yun Xiao, and Qing Du Xiaodan Liang Xiaojun Chang. Beyond fixation: Dynamic window visual transformer. *ArXiv*, abs/2203.12856, 2022.

Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. *arXiv preprint arXiv:2111.15193*, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Herv'e J'egou. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Conrad Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *ArXiv*, abs/2204.01697, 2022.

M. Alex O. Vasilescu, Eric Kim, and Xiao-Song Zeng. Causalx: Causal explanations and block multilinear factor analysis. *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10736–10743, 2021.

Ashish Vaswani, Prajit Ramachandran, A. Srinivas, Niki Parmar, Blake A. Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12889–12899, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 548–558, 2021.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *ArXiv*, abs/2106.13797, 2022.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691–2699, 2015.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal attention for long-range interactions in vision transformers. In *NeurIPS*, 2021.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 538–547, 2021.

Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2978–2988, 2021.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2018.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and José Manuel Álvarez. Understanding the robustness in vision transformers. In *ICML*, 2022.

Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *ArXiv*, abs/1611.01578, 2017.