Focal Self-attention for Local-Global Interactions in Vision Transformers

Jianwei Yang[1] **Chunyuan Li**[1] **Pengchuan Zhang**[1] **Xiyang Dai**[2] **Bin Xiao**[2]
Lu Yuan[2] **Jianfeng Gao**[1]
[1]Microsoft Research at Redmond, [2]Microsoft Cloud + AI
{jianwyan,chunyl,penzhan,xidai,bixi,luyuan,jfgao}@microsoft.com

Abstract

Recently, Vision Transformer and its variants have shown great promise on various computer vision tasks. The ability of capturing short- and long-range visual dependencies through self-attention is the key to success. But it also brings challenges due to quadratic computational overhead, especially for the high-resolution vision tasks (*e.g.*, object detection). Many recent works have attempted to reduce the computational and memory cost *and* improve performance by applying either coarse-grained global attentions or fine-grained local attentions. However, both approaches cripple the modeling power of the original self-attention mechanism of multi-layer Transformers, thus leading to sub-optimal solutions. In this paper, we present *focal self-attention*, a new mechanism that incorporates both fine-grained local and coarse-grained global interactions. In this new mechanism, each token attends its closest surrounding tokens at fine granularity and the tokens far away at coarse granularity, and thus can capture both short- and long-range visual dependencies efficiently *and* effectively. With focal self-attention, we propose a new variant of Vision Transformer models, called *Focal Transformer*, which achieves superior performance over the state-of-the-art (SoTA) vision Transformers on a range of public image classification and object detection benchmarks. In particular, our Focal Transformer models with a moderate size of 51.1M and a larger size of 89.8M achieve 83.5% and 83.8% Top-1 accuracy, respectively, on ImageNet classification at 224×224. When employed as the backbones, Focal Transformers achieve consistent and substantial improvements over the current SoTA Swin Transformers [44] across 6 different object detection methods. Our largest Focal Transformer yields **58.7/58.9** box mAPs and **50.9/51.3** mask mAPs on COCO mini-val/test-dev, and **55.4** mIoU on ADE20K for semantic segmentation, creating new SoTA on three of the most challenging computer vision tasks.

1 Introduction

Nowadays, Transformer [60] has become a prevalent model architecture in natural language processing (NLP) [22, 6]. In the light of its success in NLP, there is an increasing effort on adapting it to computer vision (CV) [48, 51]. Since its promise firstly demonstrated in Vision Transformer (ViT) [23], we have witnessed a flourish of full-Transformer models for image classification [57, 63, 67, 44, 80, 59], object detection [9, 91, 84, 20] and semantic segmentation [61, 65]. Beyond these static image tasks, it has also been applied on various temporal understanding tasks, such as action recognition [41, 83, 11], object tracking [15, 62], scene flow estimation [39].

In Transformers, self-attention is the key component making it unique from the widely used convolutional neural networks (CNNs) [38]. At each Transformer layer, it enables the global content-dependent interactions among different image regions for modeling both short- and long-range

Preprint. Under review.

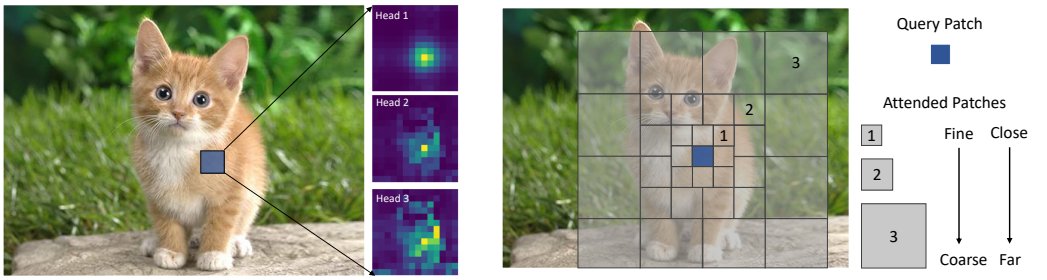


Figure 1: Left: Visualization of the attention maps of the three heads at the given query patch (blue) in the first layer of the DeiT-Tiny model [57]. Right: An illustrative depiction of focal self-attention mechanism. Three granularity levels are used to compose the attention region for the blue query.

dependencies. Through the visualization of full self-attentions[1], we indeed observe that it learns to attend local surroundings (like CNNs) and the global contexts at the same time (See the left side of Fig. 1). Nevertheless, when it comes to high-resolution images for dense predictions such as object detection or segmentation, a global and fine-grained self-attention becomes non-trivial due to the quadratic computational cost with respect to the number of grids in feature maps. Recent works alternatively exploited either a coarse-grained global self-attention [63, 67] or a fine-grained local self-attention [44, 80, 59] to reduce the computational burden. However, both approaches cripple the power of the original full self-attention *i.e.*, the ability to simultaneously model short- and long-range visual dependencies, as demonstrated on the left side of Fig. 1.

In this paper, we present a new self-attention mechanism to capture both local and global interactions in Transformer layers for high-resolution inputs. Considering that the visual dependencies between regions nearby are usually stronger than those far away, we perform the fine-grained self-attention only in local regions while the coarse-grained attentions globally. As depicted in the right side of Fig. 1, a query token in the feature map attends its closest surroundings at the finest granularity as itself. However, when it goes to farther regions, it attends to summarized tokens to capture coarse-grained visual dependencies. The further away the regions are from the query, the coarser the granularity is As a result, it can effectively cover the whole high-resolution feature maps while introducing much less number of tokens in the self-attention computation than that in the full self-attention mechanism. As a result, it has the ability to capture both short- and long-range visual dependencies efficiently. We call this new mechanism *focal self-attention*, as each token attends others in a focal manner. Based on the proposed focal self-attention, a series of Focal Transformer models are developed, by 1) exploiting a multi-scale architecture to maintain a reasonable computational cost for high-resolution images [63, 67, 44, 80], and 2) splitting the feature map into multiple windows in which tokens share the same surroundings, instead of performing focal self-attention for each token [59, 80, 44].

We validate the effectiveness of the proposed focal self-attention via a comprehensive empirical study on image classification, object detection and segmentation. Results show that our Focal Transformers with similar model sizes and complexities consistently outperform the SoTA Vision Transformer models across various settings. Notably, our small Focal Transformer model with 51.1M parameters can achieve 83.5% top-1 accuracy on ImageNet-1K, and the base model with 89.8M parameters obtains 83.8% top-1 accuracy. When transferred to object detection, our Focal Transformers consistently outperform the SoTA Swin Transformers [44] for six different object detection methods. Our largest Focal Transformer model achieves **58.9** box mAP and **51.3** mask mAP on COCO test-dev for object detection and instance segmentation, respectively, and **55.4** mIoU on ADE20K for semantic segmentation. These results demonstrate that the focal self-attention is highly effective in modeling the local-global interactions in Vision Transformers.

2 Method

2.1 Model architecture

To accommodate the high-resolution vision tasks, our model architecture shares a similar multi-scale design with [63, 80, 44], which allows us to obtain high-resolution feature maps at earlier stages. As shown in Fig. 2, an image $I \in \mathcal{R}^{H \times W \times 3}$ is first partitioned into patches of size 4×4, resulting

[1]DeiT-Tiny model, checkpoint downloaded from `https://github.com/facebookresearch/deit`.

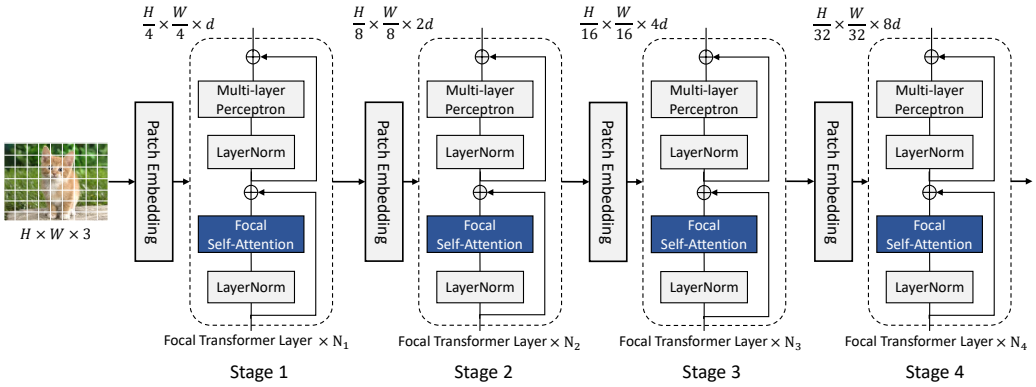


Figure 2: Model architecture for our Focal Transformers. As highlighted in light blue boxes, our main innovation is the proposed focal self-attention mechanism in each Transformer layer.

in $\frac{H}{4} \times \frac{W}{4}$ visual tokens with dimension $4 \times 4 \times 3$. Then, we use a patch embedding layer which consists of a convolutional layer with filter size and stride both equal to 4, to project these patches into hidden features with dimension d. Given this spatial feature map, we then pass it to four stages of focal Transformer blocks. At each stage $i \in \{1, 2, 3, 4\}$, the focal Transformer block consists of N_i focal Transformer layers. After each stage, we use another patch embedding layer to reduce the spatial size of feature map by factor 2, while the feature dimension is increased by 2. For image classification tasks, we take the average of the output from last stage and send it to a classification layer. For object detection, the feature maps from last 3 or all 4 stages are fed to the detector head, depending on the particular detection method we use. The model capacity can be customized by varying the input feature dimension d and the number of focal Transformer layers at each stage $\{N_1, N_2, N_3, N_4\}$.

Standard self-attention can capture both short- and long-range interactions at fine-grain, but it suffers from high computational cost when it performs the attention on high-resolution feature maps as noted in [80]. Take stage 1 in Fig. 2 as the example. For the feature map of size $\frac{H}{4} \times \frac{W}{4} \times d$, the complexity of self-attention is $\mathcal{O}((\frac{H}{4} \times \frac{W}{4})^2 d)$, resulting in an explosion of time and memory cost considering $\min(H, W)$ is 800 or even larger for object detection. In the next, we describe how we address this with the proposed focal self-attention.

2.2 Focal self-attention

In this paper, we propose focal self-attention to make Transformer layers scalable to high-resolution inputs. Instead of attending all tokens at fine-grain, we propose to attend the fine-grain tokens only locally, but the summarized ones globally. As such, it can cover as many regions as standard self-attention but with much less cost. In Fig. 3, we show the area of receptive field for standard self-attention and our focal self-attention when we gradually add more attended tokens. For a query position, when we use gradually coarser-grain for its far surroundings, focal self-attention can have significantly larger receptive fields at the cost of attending the same number of visual tokens than the baseline.

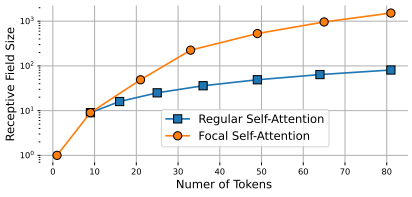


Figure 3: The size of receptive field (y-axis) with the increase of used tokens (x-axis) for standard and our focal self-attention. For focal self-attention, we assume increasing the window granularity by factor 2 gradually but no more than 8. Note that the y-axis is logarithmic.

Our focal mechanism enables long-range self-attention with much less time and memory cost, because it attends a much smaller number of surrounding (summarized) tokens. In practice, however, extracting the surrounding tokens for each query position suffers from high time and memory cost since we need to duplicate each token for all queries that can get access to it. This practical issue has been noted by a number of previous works [59, 80, 44] and the common solution is to partition the input feature map into windows. Inspired by them, we resort to perform focal self-attention at the window level. Given

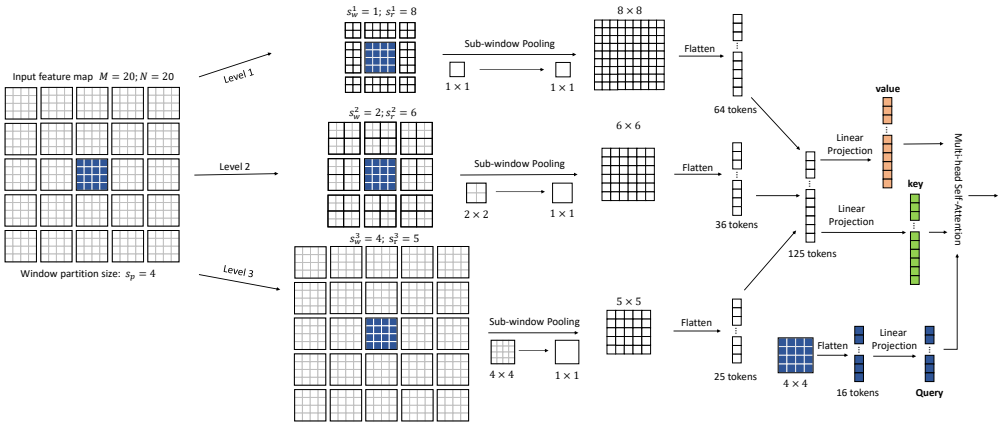


Figure 4: An illustration of our focal self-attention at window level. Each of the finest square cell represents a visual token either from the original feature map or the squeezed ones. Suppose we have an input feature map of size 20×20. We first partition it into 5×5 windows of size 4×4. Take the 4×4 blue window in the middle as the query, we extract its surroundings tokens at multiple granularity levels as its keys and values. For the first level, we extract the 8×8 tokens which are closest to the blue window at the finest grain. Then at the second level, we expand the attention region and pool the surrounding 2×2 sub-windows, which results in 6×6 pooled tokens. At the third level, we attend even larger region covering the whole feature map and pool 4×4 sub-windows. Finally, these three levels of tokens are concatenated to compute the keys and values for the $4 \times 4 = 16$ tokens (queries) in the blue window.

a feature map of $x \in \mathcal{R}^{M \times N \times d}$ with spatial size $M \times N$, we first partition it into a grid of windows with size $s_p \times s_p$. Then, we find the surroundings for each window rather than individual tokens. In the following, we elaborate the window-wise focal self-attention.

2.2.1 Window-wise attention

An illustration of the proposed window-wise focal self-attention is shown in Fig. 4. We first define three terms for clarity:

- **Focal levels** L – the number of granularity levels we extract the tokens for our focal self-attention. In Fig. 1, we show 3 focal levels in total for example.
- **Focal window size** s_w^l – the size of sub-window on which we get the summarized tokens at level $l \in \{1, ..., L\}$, which are 1, 2 and 4 for the three levels in Fig. 1.
- **Focal region size** s_r^l – the number of sub-windows horizontally and vertically in attended regions at level l, and they are 3, 4 and 4 from level 1 to 3 in Fig. 1.

With the above three terms $\{L, s_w, s_r\}$, we can specify our focal self-attention module, proceeded in two main steps:

Sub-window pooling. Assume the input feature map $x \in \mathcal{R}^{M \times N \times d}$, where $M \times N$ are the spatial dimension and d is the feature dimension. We perform sub-window pooling for all L levels. For the focal level l, we first split the input feature map x into a grid of sub-windows with size $s_w^l \times s_w^l$. Then we use a simple linear layer f_p^l to pool the sub-windows spatially by:

$$x^l = f_p^l(\hat{x}) \in \mathcal{R}^{\frac{M}{s_w^l} \times \frac{N}{s_w^l} \times d}, \quad \hat{x} = \text{Reshape}(x) \in \mathcal{R}^{(\frac{M}{s_w^l} \times \frac{N}{s_w^l} \times d) \times (s_w^l \times s_w^l)}, \tag{1}$$

The pooled feature maps $\{x^l\}_1^L$ at different levels l provide rich information at both fine-grain and coarse-grain. Since we set $s_w^l = 1$ for the first focal level which has the same granularity as the input feature map, there is no need to perform any sub-window pooling. Considering the focal window size is usually very small (7 maximally in our settings), the number of extra parameters introduced by these sub-window pooling are fairly negligible.

Attention computation. Once we obtain the pooled feature maps $\{x^l\}_1^L$ at all L levels, we compute the query at the first level and key and value for all levels using three linear projection layers f_q, f_k

and f_v:

$$Q = f_q(x^1), \quad K = \{K^l\}_1^L = f_k(\{x^1, ..., x^L\}), \quad V = \{V^l\}_1^L = f_v(\{x^1, ..., x^L\}) \tag{2}$$

To perform focal self-attention, we need to first extract the surrounding tokens for each query token in the feature map. As we mentioned earlier, tokens inside a window partition $s_p \times s_p$ share the same set of surroundings. For the queries inside the i-th window $Q_i \in \mathcal{R}^{s_p \times s_p \times d}$, we extract the $s_r^l \times s_r^l$ keys and values from K^l and V^l around the window where the query lies in, and then gather the keys and values from all L to obtain $K_i = \{K_i^1, ..., K_i^L\} \in \mathcal{R}^{s \times d}$ and $V_i = \{V_i^1, ..., V_i^L\} \in \mathcal{R}^{s \times d}$, where s is the sum of focal region from all levels, *i.e.*,, $s = \sum_{l=1}^L (s_r^l)^2$. Note that a strict version of focal self-attention following Fig. 1 requires to exclude the overlapped regions across different levels. In our model, we intentionally keep them in order to capture the pyramid information for the overlapped regions. Finally, we follow [44] to include a relative position bias and compute the focal self-attention for Q_i by:

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + B)V_i, \tag{3}$$

where $B = \{B^l\}_1^L$ is the learnable relative position bias. It consists of L subsets for L focal levels. Similar to [44], for the first level, we parameterize it to $B^1 \in \mathcal{R}^{(2s_p-1)\times(2s_p-1)}$, considering the horizontal and vertical position range are both in $[-s_p + 1, s_p - 1]$. For the other focal levels, considering they have different granularity to the queries, we treat all the queries inside a window equally and use $B^l \in \mathcal{R}^{s_r^l \times s_r^l}$ to represent the relative position bias between the query window and each of $s_r^l \times s_r^l$ pooled tokens. Since the focal self-attention for each window is independent of others, we can compute Eq. (3) in parallel. Once we complete it for the whole input feature map, we send it to the MLP block for proceeding computation as usual.

2.2.2 Complexity analysis

We analyze the computational complexity for the two main steps discussed above. For the input feature map $x \in \mathcal{R}^{M \times N \times d}$, we have $\frac{M}{s_w^l} \times \frac{N}{s_w^l}$ sub-windows at focal level l. For each sub-window, the pooling operation in Eq.1 has the complexity of $\mathcal{O}((s_w^l)^2 d)$. Aggregating all sub-windows brings us $\mathcal{O}((MN)d)$. Then for all focal levels, we have the complexity of $\mathcal{O}(L(MN)d)$ in total, which is independent of the sub-window size at each focal level. Regarding the attention computation in Eq. 3, the computational cost for a query window $s_p \times s_p$ is $\mathcal{O}((s_p)^2 \sum_l (s_r^l)^2 d)$, and $\mathcal{O}(\sum_l (s_r^l)^2 (MN)d)$ for the whole input feature map. To sum up, the overall computational cost for our focal self-attention becomes $\mathcal{O}((L + \sum_l (s_r^l)^2)(MN)d)$. In an extreme case, one can set $s_r^L = 2\max(M, N)/s_w^L$ to ensure global receptive field for all queries (including both corner and middle queries) in this layer.

2.3 Model configuration

We consider three different network configurations for our focal Transformers. Here, we simply follow the design strategy suggested by previous works [63, 67, 44], though we believe there should be a better configuration specifically for our focal Transformers. Specifically, we use similar design to the Tiny, Small and Base models in Swin Transformer [44], as shown in Table 1. Our models take 224×224 images as inputs and the window partition size is also set to 7 to make our models comparable to the Swin Transformers. For the focal self-attention layer, we introduce two levels, one for fine-grain local attention and one for coarse-grain global attention. Expect for the last stage, the focal region size is consistently set to 13 for the window partition size of 7, which means that we expand 3 tokens for each window partition. For the last stage, since the whole feature map is 7×7, the focal region size at level 0 is set to 7, which is sufficient to cover the entire feature map. For the coarse-grain global attention, we set its focal window size same to the window partition size 7, but gradually decrease the focal region size to get $\{7, 5, 3, 1\}$ for the four stages. For the patch embedding layer, the spatial reduction ratio p_i for four stages are all $\{4, 2, 2, 2\}$, while Focal-Base has a higher hidden dimension compared with Focal-Tiny and Focal-Small.

3 Related work

Vision Transformers. The Vision Transformer (ViT) was first introduced in [23]. It applies a standard Transformer encoder, originally developed for NLP [60], to encode image by analogously splitting an

	Output Size	Layer Name	Focal-Tiny	Focal-Small	Focal-Base
stage 1	56×56	Patch Embedding	$p_1 = 4; c_1 = 96$	$p_1 = 4; c_1 = 96$	$p_1 = 4; c_1 = 128$
	56×56	Transformer Block	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 7\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 7\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 7\} \end{bmatrix} \times 2$
stage 2	28×28	Patch Embedding	$p_2 = 2; c_2 = 192$	$p_2 = 2; c_2 = 192$	$p_2 = 2; c_2 = 256$
	28×28	Transformer Block	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 5\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 5\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 5\} \end{bmatrix} \times 2$
stage 3	14×14	Patch Embedding	$p_3 = 2; c_3 = 384$	$p_3 = 2; c_3 = 384$	$p_3 = 2; c_3 = 512$
	14×14	Transformer Block	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 3\} \end{bmatrix} \times 6$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 3\} \end{bmatrix} \times 18$	$\begin{bmatrix} s^0_{w,r} = \{1, 13\} \\ s^1_{w,r} = \{7, 3\} \end{bmatrix} \times 18$
stage 4	7×7	Patch Embedding	$p_4 = 2; c_4 = 768$	$p_4 = 2; c_4 = 768$	$p_4 = 2; c_4 = 1024$
	7×7	Transformer Block	$\begin{bmatrix} s^0_{w,r} = \{1, 7\} \\ s^1_{w,r} = \{7, 1\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 7\} \\ s^1_{w,r} = \{7, 1\} \end{bmatrix} \times 2$	$\begin{bmatrix} s^0_{w,r} = \{1, 7\} \\ s^1_{w,r} = \{7, 1\} \end{bmatrix} \times 2$

Table 1: Model configurations for our focal Transformers. We introduce three configurations Focal-Tiny, Focal-Small and Focal-Base with different model capacities.

image into a sequence of visual tokens. It has demonstrated superior performance to convolutional neural networks (CNNs) such as the ResNet [34] on multiple image classification benchmarks, when trained with sufficient data [23] and careful data augmentation and regularization [57]. These advancements further inspired the applications of transformer to various vision tasks beyond image classification, such as self-supervised learning [16, 10, 40], object detection [9, 91, 84, 20] and semantic segmentation [61, 65, 86]. Apart from the downstream tasks, another line of work focus on improving the original vision transformers from different perspectives, such as data-efficient training [57], improved patch embedding/encoding [18, 75, 32], integrating convolutional projections into transformers [67, 74], multi-scale architectures and efficient self-attention mechanisms for high-resolution vision tasks [63, 67, 44, 80, 17]. We refer the readers to [37, 31, 37] for comprehensive surveys. This paper focuses on improving the general performance of vision transformer with the proposed focal self-attention mechanism. In the following, we particularly discussed the most related works regarding attention mechanisms.

Efficient global and local self-attention. Transformer models usually need to cope with a large number of tokens, such as long documents in NLP and high-resolution images in CV. Recently, various efficient self-attention mechanisms are proposed to overcome the quadratic computational and memory cost in the vanilla self-attention. On one hand, a number of works in both NLP and CV resort to coarse-grained global self-attention by attending the downsampled/summarized tokens, while preserving the long-range interactions [50, 47, 63, 67, 32]. Though this approach can improve the efficiency, it loses the detailed context surrounding the query tokens. On the other hand, the local fine-grained attention, *i.e.*, attending neighboring tokens within a constant window size, is another solution for both language [3, 78, 1] and vision [59, 44, 80]. In this paper, we argue that both types of attentions are important and the full-attention ViT models indeed have learned both of them, as shown in Fig. 1 left. This is also supported by the recent advanced CNN models [36, 66, 64, 71, 2, 8, 52], which showed that global attention or interaction can effectively improve the performance. Our proposed focal self-attention is the first to reconcile the global and local self-attention in a single transformer layer. It can capture both local and global interactions as vanilla full attention but in more efficient and effective way, particularly for high-resolution inputs.

4 Experiments

4.1 Image classification on ImageNet-1K

We compare different methods on ImageNet-1K [21]. For fair comparison, we follow the training recipes in [57, 63]. All models are trained for 300 epochs with a batch size 1024. The initial learning rate is set to 10^{-3} with 20 epochs of linear warm-up starting from 10^{-5}. For optimization, we use AdamW [45] as the optimizer with a cosine learning rate scheduler. The weight decay is set to 0.05 and the maximal gradient norm is clipped to 5.0. We use the same set of data augmentation and regularization strategies used in [57] after excluding random erasing [87], repeated

Model	#Params.	FLOPs	Top-1 (%)
ResNet-50 [34]	25.0	4.1	76.2
DeiT-Small/16 [57]	22.1	4.6	79.9
PVT-Small [63]	24.5	3.8	79.8
ViL-Small [80]	24.6	5.1	82.0
CvT-13 [67]	20.0	4.5	81.6
Swin-Tiny [44]	28.3	4.5	81.2
Focal-Tiny (Ours)	29.1	4.9	82.2
ResNet-101 [34]	45.0	7.9	77.4
PVT-Medium [63]	44.2	6.7	81.2
CvT-21 [67]	32.0	7.1	82.5
ViL-Medium [80]	39.7	9.1	83.3
Swin-Small [44]	49.6	8.7	83.1
Focal-Small (Ours)	51.1	9.1	83.5
ResNet-152 [34]	60.0	11.0	78.3
ViT-Base/16 [23]	86.6	17.6	77.9
DeiT-Base/16 [57]	86.6	17.5	81.8
PVT-Large [63]	61.4	9.8	81.7
ViL-Base [80]	55.7	13.4	83.2
Swin-Base [44]	87.8	15.4	83.4
Focal-Base (Ours)	89.8	16.0	**83.8**

Table 2: Comparison of image classification on ImageNet-1K for different models. Except for ViT-Base/16, all other models are trained and evaluated on 224×224 resolution.

Backbone	RetinaNet	Mask R-CNN	
	AP^b	AP^b	AP^m
ResNet-50 [34]	36.3	38.0	34.4
PVT-Small	40.4	40.4	37.8
ViL-Small [80]	41.6	41.8	38.5
Swin-Tiny [44]	42.0	43.7	39.8
Focal-Tiny (Ours)	**43.7** (+1.7)	**44.8** (+1.1)	**41.0** (+1.3)
ResNet-101 [34]	38.5	40.4	36.4
ResNeXt101-32x4d [70]	39.9	41.9	37.5
PVT-Medium [63]	41.9	42.0	39.0
ViL-Medium [80]	42.9	43.4	39.7
Swin-Small [44]	45.0	46.5	42.1
Focal-Small (Ours)	**45.6** (+0.6)	**47.4** (+0.9)	**42.8** (+0.7)
ResNeXt101-64x4d [70]	41.0	42.8	38.4
PVT-Large [63]	42.6	42.9	39.5
ViL-Base [80]	44.3	45.1	41.0
Swin-Base [44]	45.0	46.9	42.3
Focal-Base (Ours)	**46.3** (+1.3)	**47.8** (+0.9)	**43.2** (+0.9)

Table 3: Comparisons with CNN and Transformer baselines and SoTA methods on COCO object detection. The box mAP (AP^b) and mask mAP (AP^m) are reported for RetinaNet and Mask R-CNN trained with $1\times$ schedule. More detailed comparisons with $3\times$ schedule are in Table 4.

augmentation [4, 35] and exponential moving average (EMA) [49]. The stochastic depth drop rates are set to 0.2, 0.2 and 0.3 for our tiny, small and base models, respectively. During training, we crop images randomly to 224×224, while a center crop is used during evaluation on the validation set.

In Table 2, we summarize the results for baseline models and the current state-of-the-art models on image classification task. We can find our Focal Transformers consistently outperforms other methods with similar model size (#Params.) and computational complexity (GFLOPs). Specifically, Focal-Tiny improves over the Transformer baseline DeiT-Small/16 by 2.0%. Meanwhile, using the same model configuration (2-2-6-2) and a few extra parameters and computations, our Focal-Tiny improves over Swin-Tiny by 1.0 point (81.2% $\rightarrow$ 82.2%). When we increase the window size from 7 to 14 to match the settings in ViL-Small [80], the performance can be further improved to 82.5%. For small and base models, our Focal Transformers still achieves slightly better performance than the others. Notably, our Focal-Small with 51.1M parameters can reach 83.5% which is better than all counterpart small and base models using much less parameters. When further increasing the model size, our Focal-Base model achieves 83.8%, surpassing all other models using comparable parameters and FLOPs. We refer the readers to our appendix for more detailed comparisons.

4.2 Object detection and instance segmentation

We benchmark our models on object detection with COCO 2017 [43]. The pretrained models are used as visual backbones and then plug into two representative pipelines, RetinaNet [42] and Mask R-CNN [33]. All models are trained on the 118k training images and results reported on 5K validation set. We follow the standard to use two training schedules, $1\times$ schedule with 12 epochs and $3\times$ schedule with 36 epochs. For $1\times$ schedule, we resize image's shorter side to 800 while keeping its longer side no more than 1333. For $3\times$ schedule, we use multi-scale training strategy by randomly resizing its shorter side to the range of $[480, 800]$. Considering this higher input resolution, we adaptively increase the focal sizes at four stages to $(15, 13, 9, 7)$, to ensures the focal attention covers more than half of the image region (first two stages) to the whole image (last two stages). With the focal size increased, the relative position biases are accordingly up-sampled to corresponding sizes using bilinear interpolation. During training, we use AdamW [45] for optimization with initial learning rate 10^{-4} and weight decay 0.05. Similarly, we use 0.2, 0.2 and 0.3 stochastic depth drop rates to regularize the training for our tiny, small and base models, respectively. Since Swin Transformer does not report the numbers on RetinaNet, we train it by ourselves using their official code with the same hyper-parameters with our Focal Transformers.

Backbone	#Params (M)	FLOPs (G)	RetinaNet 3x schedule + MS						Mask R-CNN 3x schedule + MS					
			AP^b	AP^b_{50}	AP^b_{75}	AP_S	AP_M	AP_L	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50 [34]	37.7/44.2	239/260	39.0	58.4	41.8	22.4	42.8	51.6	41.0	61.7	44.9	37.1	58.4	40.1
PVT-Small[63]	34.2/44.1	226/245	42.2	62.7	45.0	26.2	45.2	57.2	43.0	65.3	46.9	39.9	62.5	42.8
ViL-Small [80]	35.7/45.0	252/174	42.9	63.8	45.6	27.8	46.4	56.3	43.4	64.9	47.0	39.6	62.1	42.4
Swin-Tiny [44]	38.5/47.8	245/264	45.0	65.9	48.4	29.7	48.9	58.1	46.0	68.1	50.3	41.6	65.1	44.9
Focal-Tiny (Ours)	39.4/48.8	265/291	**45.5**	**66.3**	**48.8**	**31.2**	**49.2**	**58.7**	**47.2**	**69.4**	**51.9**	**42.7**	**66.5**	**45.9**
ResNet101 [34]	56.7/63.2	315/336	40.9	60.1	44.0	23.7	45.0	53.8	42.8	63.2	47.1	38.5	60.1	41.3
ResNeXt101-32x4d [70]	56.4/62.8	319/340	41.4	61.0	44.3	23.9	45.5	53.7	44.0	64.4	48.0	39.2	61.4	41.9
PVT-Medium [63]	53.9/63.9	283/302	43.2	63.8	46.1	27.3	46.3	58.9	44.2	66.0	48.2	40.5	63.1	43.5
ViL-Medium [80]	50.8/60.1	339/261	43.7	64.6	46.4	27.9	47.1	56.9	44.6	66.3	48.5	40.7	63.8	43.7
Swin-Small [44]	59.8/69.1	335/354	46.4	67.0	50.1	31.0	50.1	60.3	48.5	70.2	53.5	43.3	67.3	46.6
Focal-Small (Ours)	61.7/71.2	367/401	**47.3**	**67.8**	**51.0**	**31.6**	**50.9**	**61.1**	**48.8**	**70.5**	**53.6**	**43.8**	**67.7**	**47.2**
ResNeXt101-64x4d [70]	95.5/102	473/493	41.8	61.5	44.4	25.2	45.4	54.6	44.4	64.9	48.8	39.7	61.9	42.6
PVT-Large[63]	71.1/81.0	345/364	43.4	63.6	46.1	26.1	46.0	59.5	44.5	66.0	48.3	40.7	63.4	43.7
ViL-Base [80]	66.7/76.1	443/365	44.7	65.5	47.6	29.9	48.0	58.1	45.7	67.2	49.9	41.3	64.4	44.5
Swin-Base [44]	98.4/107	477/496	45.8	66.4	49.1	29.9	49.4	60.3	48.5	69.8	53.2	43.4	66.8	46.9
Focal-Base (Ours)	100.8/110.0	514/533	**46.9**	**67.8**	**50.3**	**31.9**	**50.3**	**61.5**	**49.0**	**70.1**	**53.6**	**43.7**	**67.6**	**47.0**

Table 4: COCO object detection and segmentation results with RetinaNet [42] and Mask R-CNN [34]. All models are trained with $3\times$ schedule and multi-scale inputs (MS). The numbers before and after "/" at column 2 and 3 are the model size and complexity for RetinaNet and Mask R-CNN, respectively.

In Table 3, we show the performance for both CNN-based models and the current Transformer-based state-of-the-arts methods. The bbox mAP (AP^b) and mask mAP (AP^m) are reported. Our Focal Transformers outperform the CNN-based models consistently with the gap of 4.8-7.1 points. Compared with the other methods which also use multi-scale Transformer architectures, we still observe substantial gains across all settings and metrics. Particularly, our Focal Transformers brings 0.7-1.7 points of mAP against the current best approach Swin Transformer [44] at comparable settings. Different from the other multi-scale Transformer models, our method can simultaneously enable short-range fine-grain and long-range coarse-grain interactions for each visual token, and thus capture richer visual contexts at each layer for better dense predictions. To have more comprehensive comparisons, we further train them with $3\times$ schedule and show the detailed numbers for RetinaNet and Mask R-CNN in Table 4. For comprehension, we also list the number of parameters and the associated computational cost for each model. As we can see, even for $3\times$ schedule, our models can still achieve 0.3-1.1 gain over the best Swin Transformer models at comparable settings.

To further verify the effectiveness of our proposed Focal Transformers, we follow [44] to train four different object detectors including Cascade R-CNN [7], ATSS [81], RepPoints [72] and Sparse R-CNN [55]. We use Focal-Tiny as the backbone and train all four models using $3\times$ schedule. The box mAPs on COCO validation set are reported in Table 5. As we can see, our Focal-Tiny exceeds Swin-Tiny by 1.0-2.3 points on all methods. These significant and consistent improvements over different detection methods in addition to RetinaNet and Mask R-CNN suggest that our Focal Transformer can be used as a generic backbone for a variety of object detection methods.

Method	Backbone	#Param	FLOPs	AP^b	AP^b_{50}	AP^b_{75}
C. Mask R-CNN [7]	R-50	82.0	739	46.3	64.3	50.5
	Swin-T	85.6	742	50.5	69.3	54.9
	Focal-T	86.7	770	**51.5** (+1.0)	**70.6**	**55.9**
ATSS [81]	R-50	32.1	205	43.5	61.9	47.0
	Swin-T	35.7	212	47.2	66.5	51.3
	Focal-T	36.8	239	**49.5** (+2.3)	**68.8**	**53.9**
RepPointsV2 [72]	R-50	43.4	431	46.5	64.6	50.3
	Swin-T	44.1	437	50.0	68.5	54.2
	Focal-T	45.4	491	**51.2** (+1.2)	**70.4**	**54.9**
Sparse R-CNN [55]	R-50	106.1	166	44.5	63.4	48.2
	Swin-T	109.7	172	47.9	67.3	52.3
	Focal-T	110.8	196	**49.0** (+1.1)	**69.1**	**53.2**

Table 5: Comparison with ResNet-50, Swin-Tiny across different object detection methods. We use Focal-Tiny as the backbone and train all models using $3\times$ schedule.

Besides the instance segmentation results above, we further evaluate our model on semantic segmentation, a task that usually requires high-resolution input and long-range interactions. We benchmark our method on ADE20K [88]. Specifically, we use UperNet [68] as the segmentation method and our Focal Transformers as the backbone. We train three models with Focal-Tiny, Focal-Small, Focal-Base, respectively. For all models, we use a standard recipe by setting the input size to 512×512 and train the model for 160k iterations with batch size 16. In Table 7, we show the comparisons to previous works. As we can see,

Method	#Param	FLOPs	mini-val		test-dev	
			AP^b	AP^m	AP^b	AP^m
X101-64x4d [70]	155M	1033G	52.3	46.0	-	-
EfficientNet-D7 [56]	77M	410G	54.4	-	55.1	-
GCNet* [8]	-	1041G	51.8	44.7	52.3	45.4
ResNeSt-200 [79]	-	-	52.5	-	53.3	47.1
Copy-paste [28]	185M	1440G	55.9	47.2	56.0	47.4
BoTNet-200 [52]	-	-	49.7	-	-	-
SpineNet-190 [24]	164M	1885G	52.6	-	52.8	-
CenterNet2 [90]	-	-	-	-	56.4	-
Swin-L (HTC++) [44]	284M	1470G	57.1	49.5	57.7	50.2
Swin-L (DyHead) [19]	213M	965G	56.2	-	-	-
Swin-L† (HTC++) [44]	284M	-	58.0	50.4	58.7	51.1
Swin-L† (DyHead) [19]	213M	-	58.4	-	58.7	-
Swin-L† (QueryInst) [26]	-	-	56.1	-	56.1	-
Focal-L (HTC++) (Ours)	265M	1165G	57.0	49.9	-	-
Focal-L (DyHead) (Ours)	229M	1081G	56.4	-	-	-
Focal-L† (HTC++) (Ours)	265M	-	58.1	**50.9**	58.4	**51.3**
Focal-L† (DyHead) (Ours)	229M	-	**58.7**	-	**58.9**	-

Table 6: Comparison with state-of-the-art methods on COCO object detection and instance segmentation. The numbers are reported on 5K val set and test-dev. Augmented HTC [13] (denoted by HTC++) and DyHead [19] are used as the detection methods. † means multi-scale evaluation.

Backbone	Method	#Param	FLOPs	mIoU	+MS
ResNet-101	DANet [46]	69M	1119G	45.3	-
ResNet-101	ACNet [27]	-	-	45.9	-
ResNet-101	DNL [73]	69M	1249G	46.0	-
ResNet-101	UperNet [68]	86M	1029G	44.9	-
HRNet-w48 [54]	OCRNet [77]	71M	664G	45.7	-
ResNeSt-200 [79]	DLab.v3+ [14]	88M	1381G	48.4	-
Swin-T [44]	UperNet [68]	60M	945G	44.5	45.8
Swin-S [44]	UperNet [68]	81M	1038G	47.6	49.5
Swin-B [44]	UperNet [68]	121M	1188G	48.1	49.7
Twins-SVT-L [17]	UperNet [68]	133M	-	48.8	50.2
MiT-B5 [69]	SegFormer [69]	85M	-	51.0	51.8
ViT-L/16† [23]	SETR [85]	308M	-	50.3	-
Swin-L‡ [44]	UperNet [68]	234M	3230G	52.1	53.5
ViT-L/16‡ [23]	Segmenter [53]	334M	-	51.8	53.6
Swin-L‡ [44]	K-Net [82]	-	-	-	54.3
Swin-L‡ [44]	PatchDiverse [29]	234M	-	53.1	54.4
VOLO-D5 [76]	UperNet [68]	-	-	-	54.3
Focal-T (Ours)	UperNet [68]	62M	998G	45.8	47.0
Focal-S (Ours)	UperNet [68]	85M	1130G	48.0	50.0
Focal-B (Ours)	UperNet [68]	126M	1354G	49.0	50.5
Focal-L‡ (Ours)	UperNet [68]	240M	3376G	**54.0**	**55.4**

Table 7: Comparison with SoTA methods for semantic segmentation on ADE20K [88] val set. Both single- and multi-scale evaluations are reported at the last two columns. ‡ means pretrained on ImageNet-22K.

our tiny, small and base models consistently outperforms Swin Transformers with similar size on single-scale and multi-scale mIoUs.

4.3 Comparing with system-level SoTA methods

To compare with SoTA at the system level, we build Focal Large model by increasing the hidden dimension in Focal-Base from 128 to 196 while keeping all the others the same, similar to Swin Transformers. To achieve the best performance, the common practice is to pretrain on ImageNet-22K and then transfer the model to down stream tasks [67, 44]. However, due to the limited resources, we partially initialize our model with the pretrained Swin Transformer checkpoint[2], considering our network architecture is similar with Swin Transformers except for the window shift and focal self-attention. Specifically, we reuse the parameters in Swin-Large model but remove the window shift operation and randomly initialize our own window pooling layer in Eq. (1) and local-to-global relative position bias in Eq. (3). Then, we finetune our model on ImageNet-1K to learn the focal-specific parameters. The resulting model is used as the backbone and further finetuned on object detection and semantic segmentation tasks.

Comparison with SoTA detection systems. For object detection on COCO, we first follow Swin Transformer to also use HTC [13] as the detection method in that it reported SoTA performance on COCO detection when using Swin Transformer as the backbone. For fair comparison, we also use soft-NMS [5], instaboost [25] and a multi-scale training strategy with shorter side in range $[400, 1400]$ while the longer side is no more than 1600. We train the model using AdamW [45] with base learning rate 1e-4 and weight decay 0.1. The model is trained using standard $3\times$ schedule. The box and mask mAPs on COCO validation set and test-dev are reported in Table 6. We show both single-scale evaluation and multi-scale evaluation results. Our Focal-Large model with multi-scale test achieve 58.1 box mAP and 50.9 mask mAP on mini-val set, which are better than the reported numbers for Swin-Large in [44]. When evaluating our model on the test-dev set, it achieves 58.4 box mAP and 51.3 mask mAP, which is slightly better than Swin Transformer. Note that because our model does not include global self-attention layer used in Swin Transformer at the last stage, it has smaller model size and fewer FLOPs. More recently, DyHead [19] achieves new SoTA on COCO, when combined with Swin-Large. We replace the Swin-Large model with our Focal-Large model, and use the same $2\times$ training schedule as in [19]. We report the box mAPs for both mini-val and

[2]Pretrained models are available at https://github.com/microsoft/Swin-Transformer

Model	W-Size	FLOPs	Top-1 (%)	AP^b	AP^m
Swin-Tiny	7	4.5	81.2	43.7	39.8
	14	4.9	82.1	44.0	40.5
Focal-Tiny	7	4.9	82.2	44.9	41.1
	14	5.2	82.3	45.5	41.5

Table 8: Impact of different window sizes (W-Size). We alter the default size 7 to 14 and observe consistent improvements for both methods.

Model	W-Shift	Top-1 (%)	AP^b	AP^m
Swin-Tiny	-	80.2	38.8	36.4
	✓	81.2	43.7	39.8
Focal-Tiny	-	82.2	44.8	41.0
	✓	81.9	44.9	41.1

Table 9: Impact of window shift (W-Shift) on Swin Transformer and Focal Transformer. Tiny models are used.

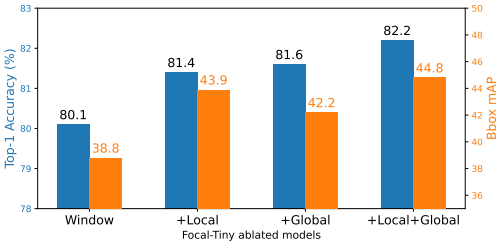


Figure 5: Ablating Focal-Tiny model by adding local, global and both interactions, respectively. Blue bars are for image classification and orange bars indicate object detection performance. Both local and global interactions are essential to obtain good performance. Better viewed in color.

Depths	Model	#Params.	FLOPs	Top-1 (%)	AP^b	AP^m
2-2-2-2	Swin	21.2	3.1	78.7	38.2	35.7
	Focal	21.7	3.4	79.9	40.5	37.6
2-2-4-2	Swin	24.7	3.8	80.2	41.2	38.1
	Focal	25.4	4.1	81.4	43.3	39.8
2-2-6-2	Swin	28.3	4.5	81.2	43.7	39.8
	Focal	29.1	4.9	82.2	44.8	41.0

Table 10: Impact of the change of model depth. We gradually reduce the number of transformer layers at the third stage from original 6 to 4 and further 2. It apparently hurts the performance but our Focal Transformers has much slower drop rate than Swin Transformer.

test-dev. Our Focal-Large clearly bring substantial improvements over both metrics, reaching new SoTA on both metrics.

Comparison with SoTA semantic segmentation systems. We further use the pretrained Focal-Large model as the backbone for semantic segmentation. We follow the same setting as in [44]. Specifically, we use input image size 640×640 and train the model for 160k iterations with a batch size of 16. We set the initial learning to 6e-5 and use a polynomial learning rate decay. The weight decay is set to 0.01. For multi-scale evaluation, we use the same scaling ratios $[0.5, 0.75, 1.0, 1.25, 1.5, 1.75]$ as in previous works. In Table 7, we see that our Focal-Large achieves significantly better performance than Swin-Large. In both single-scale and multi-scale evaluation, Focal-Large has more than 1 point mIoU improvement, which presents a new SoTA for semantic segmentation on ADE20K. These encouraging results verify the effectiveness of our proposed focal self-attention mechanism in capturing long-range dependencies required by dense visual prediction tasks.

4.4 Ablation studies

We conduct ablation studies to inspect the model's capacity from different aspects. Focal-Tiny is considered on both image classification and object detection tasks.

Effect of varying window size. Above we have demonstrated that both short- and long-range interactions are necessary. Based on this, a natural question is that whether increasing the window size can further help the model learning giving an enlarged receptive field. In Table 8, we show the performance of Swin-Tiny and Focal-Tiny with window size 7 and 14. Clearly, a larger window size brings gain for both methods on all three metrics and our Focal-Tiny model consistently outperforms Swin-Tiny using both window sizes. Comparing the second and third row, we find our model beats Swin even using much smaller window size (7 *v.s.* 14). We suspect the long-range interactions in our model is the source of this gain.

The necessity of window shift. In [44], the authors proposed window shift operations to enable the cross-window interactions between two successive layers. In contrast, the visual tokens in our Focal Transformer can always communicate with those in other windows at both fine- and coarse-grain. A natural question is whether adding the window shift to our Focal Transformers can further lead to

improvements. To investigate this, we remove the window shift from Swin Transformer while add it to our Focal Transformer. As shown in Table 9, Swin Transformer shows a severe degradation after removing the window shift. However, our Focal Transformer is even hurt on classification task. These results indicate that the window shift is not a necessary ingredient in our model. As such, our model can get rid of the constraint in Swin Transformer that there should be an even number of layers in each stage for the alternative window shift operation.

Contributions of short- and long-range interaction. We attempt to factorize the effect of short-range fine-grain and long-range coarse-grain interactions in our Focal Transformers. We ablate the original Focal-Tiny model to: a) Focal-Tiny-Window merely performing attention inside each window; b) Focal-Tiny-Local attending the additional fine-grain surrounding tokens and c) Focal-Tiny-Global attending the extra coarse-grain squeezed tokens. We train them using the same setting as Focal-Tiny and report their performance on image classification and object detection using Mask R-CNN $1\times$ schedule. As we can see from Fig. 5, Focal-Tiny-Window suffers from significant drop on both image classification (82.2→80.1) and object detection (44.8→38.3). This is expected since the communication across windows are totally cut off at each Transformer layer. After we enable either the local fine-grain or global coarse-grain interactions (middle two columns), we observe significant jumps. Though they prompt richer interactions from different paths, finally both of them enable the model to capture more contextual information. When we combine them together, we observe further improvements on both tasks. This implies that these two type of interactions are complementary to each other and both of them should be enabled in our model. Another observation is that adding long-range tokens can bring more relative improvement for image classification than object detection and vice versa for local tokens. We suspect that dense predictions like object detection more rely on fine-grained local context while image classification favors more the global information.

Model capacity against model depth. Considering our focal attention prompts local and global interactions at each Transformer layer, one question is that whether it needs less number of layers to obtain similar modeling capacity as those without global interactions. To answer this, we conduct the experiments by reducing the number of Transformer layers at stage 3 in Swin-Tiny and Focal-Tiny from the original 6 to 4 and 2. In Table 10, we show the performance and model complexity for each variant. First, we can find our model outperforms Swin model consistently with the same depth. More importantly, using two less layers, our model achieves comparable performance to Swin Transformer. Particularly, Focal-Tiny with 4 layers achieves 81.4 on image classification which is even better than original Swin-Tiny model with 6 layers (highlighted in gray cells). Though we do not explore different architectures for our Focal Transformer, these results suggest that we can potential find even more efficient *and* effective architectures.

5 Conclusion

In this paper, we have presented focal self-attention to enable efficient local-global interactions in vision Transformers. Different from previous works, it performs the local self-attention at fine-grain and global self-attention at coarse-grain, which results in an effective way to capture richer context in both short and long-range at a reasonable cost. By plugging it into a multi-scale transformer architecture, we propose Focal Transformers, which demonstrates its superiority over the SoTA methods on both image classification, object detection and segmentation. With these extensive experimental results, the proposed focal attention is shown as a generic approach for modeling local-global interactions in vision Transformers for various vision tasks.

Limitations and future work. Though extensive experimental results showed that our focal self-attention can significantly boost the performance on both image classification and dense prediction tasks, it does introduce extra computational and memory cost, since each query token needs to attend the coarsened global tokens in addition to the local tokens. Developing some practical or methodological techniques to reduce the cost would be necessary to make it more applicable in realistic scenarios. Our ablation study on the number of used Transformer layers in Table 10 indeed shed light on the potential way to reduce the cost via reducing the number of transformer layers. However, we merely scratched the surface and further study on this aspect is still needed. In Focal Transformers, we chose multi-scale architecture as the base so that it can work for high-resolution prediction tasks. However, we believe our focal attention mechanism is also applicable to monolithic vision Transformers and Transformers in both vision and language domain. We leave this as a promising direction to further explore in the future.

References

[1] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Philip Pham, Anirudh Ravula, and Sumit Sanghai. Etc: Encoding long and structured data in transformers. *arXiv preprint arXiv:2004.08483*, 2020.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3286–3295, 2019.

[3] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[4] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019.

[5] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[6] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[8] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[11] Shuning Chang, Pichao Wang, Fan Wang, Hao Li, and Jiashi Feng. Augmented transformer with adaptive graph for temporal action proposal generation. *arXiv preprint arXiv:2103.16024*, 2021.

[12] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[13] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019.

[14] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[15] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. *arXiv preprint arXiv:2103.15436*, 2021.

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[17] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting spatial attention design in vision transformers. *arXiv preprint arXiv:2104.13840*, 2021.

[18] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *Arxiv preprint 2102.10882*, 2021.

[19] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.

[20] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. *arXiv preprint arXiv:2011.09094*, 2020.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2020.

[25] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–691, 2019.

[26] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries, 2021.

[27] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6757, 2019.

[28] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020.

[29] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification, 2021.

[30] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:22104.01136*, 2021.

[31] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

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

[33] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

[36] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[37] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[38] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[39] Bing Li, Cheng Zheng, Silvio Giancola, and Bernard Ghanem. Sctn: Sparse convolution-transformer network for scene flow estimation. *arXiv preprint arXiv:2105.04447*, 2021.

[40] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021.

[41] Xiangyu Li, Yonghong Hou, Pichao Wang, Zhimin Gao, Mingliang Xu, and Wanqing Li. Trear: Transformer-based rgb-d egocentric action recognition. *arXiv preprint arXiv:2101.03904*, 2021.

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

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

[44] 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. *arXiv preprint arXiv:2103.14030*, 2021.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307, 2017.

[47] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844, 2019.

[48] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[49] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[50] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

[51] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[52] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition, 2021.

[53] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021.

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[55] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. *arXiv preprint arXiv:2011.12450*, 2020.

[56] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[57] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[58] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021.

[59] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *arXiv preprint arXiv:2012.00759*, 2020.

[62] Ning Wang, Wengang Zhou, Jie Wang, and Houqaing Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. *arXiv preprint arXiv:2103.11681*, 2021.

[63] 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. *arXiv preprint arXiv:2102.12122*, 2021.

[64] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[65] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020.

[66] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[67] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[68] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.

[70] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[71] Jianwei Yang, Zhile Ren, Chuang Gan, Hongyuan Zhu, and Devi Parikh. Cross-channel communication networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1297–1306, 2019.

[72] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019.

[73] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020.

[74] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv preprint arXiv:2103.11816*, 2021.

[75] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[76] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv preprint arXiv:2106.13112*, 2021.

[77] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[78] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*, 2020.

[79] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[80] 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. *arXiv preprint arXiv:2103.15358*, 2021.

[81] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020.

[82] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation, 2021.

[83] Jiaojiao Zhao, Xinyu Li, Chunhui Liu, Shuai Bing, Hao Chen, Cees GM Snoek, and Joseph Tighe. Tuber: Tube-transformer for action detection. *arXiv preprint arXiv:2104.00969*, 2021.

[84] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020.

[85] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[86] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[87] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[88] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[89] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[90] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021.

[91] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Appendix

A.1 Image classification

We present the exhaustive comparison with previous works in Table 11. We compare our method with both CNN-based and Transformer-based methods. We categorize different methods into groups based on two properties:

- **Scale** – the scale of feature maps in a model. It can be either a single-scale or multi-scale. In single-scale models, all feature maps have the same size across different stages. For multi-scale models, there are usually feature maps with different resolutions with the proceeding stages.
- **Locality** – the locality of operations in a model. It can be either global or local. Local operations can be a convoluional layer in CNN models or a transformer layer which conducts local self-attention. However, global operations such as the standard self-attention, produce the output feature map by gather information from all inputs.

Based on this criterion, all CNN models are natural multi-scale because their feature map sizes gradually decrease at different stages. Recently, a number of works attempt to integrate the global operations into CNNs by introducing squeeze-and-excitation (SE) layer [36], channel-wise attention layer [66] and even self-attention layer [2, 52]. As we can see, the combination of local and global operations significantly improve the performance for image classification. Particularly, BotNet-S1-110 achieves 82.8 top-1 accuracy with moderate number of parameters (61.6M).

On the contrary, Transformers [60] by nature performs global self-attention by which each visual token can interact with all others. Even without multi-scale design as in CNNs, a number of Transformer-based works such as TNT [32], DeepViT [89] and CaiT [58] achieve superior performance to CNN models with comparable model size and computational cost. To accommodate the high resolution feature maps, some recent works replace global self-attention with more efficient local self-attention and demonstrate comparable performance on image classification while much promising results on dense prediction tasks such as object detection and semantic segmentation [44].

In this paper, we present focal attention which is the first to combine global self-attention and local self-attention in an efficient way. Replacing either the global self-attention or the local self-attention with our focal self-attention, we achieve better performance than both. These results along with the CNN models augmented by local and global computations demonstrate that combining local and global interactions are more effective than either of them. In the table, we also report the speed for different methods. Using the same script provided by [44], we run the test on a single Tesla-V100 with batch size 64. Accordingly, our Focal Transformer has slower running speed though it has similar FLOPs as Swin Transformer. This is mainly due to two reasons: 1) we introduce the global coarse-grain attention, and it introduces the extra computations; 2) though we conduct our focal attention on the windows, we swill observe that extracting the surrounding tokens around local windows and the global tokens across the feature map are time-consuming.

A.2 Object detection and segmentation

For completeness, we report the full metrics for RetinaNet and Mask R-CNN trained with 1x schedule in Table 12. As we can see, our Focal Transformers consistently outperform previous works including the state-of-the-art Swin Transformers on all metrics. We observe that our models trained with 1x schedule generally have more gain against the previous best models than 3x schedule (+1.2 *v.s.* +0.8 and +1.0 *v.s.* +0.7 box mAP for RetinaNet and Mask R-CNN, respectively). This indicates that our models have faster learning convergences compared with previous works. Compared with the local-attention based methods, *e.g.*, Swin Transformer, integrating the long-range interactions can help capture more visual dependencies and thus help the model to learn faster.

A.3 Model inspections

Learning speed comparison. As we briefly discussed earlier, our model shows faster learning speed on object detection task. In Fig. 6, we show the top-1 validation accuracy of our models and Swin Transformers for image classification task. Accordingly, our Focal Transformers have much faster learning speed as well. For example, Focal-Tiny has 75.7% top-1 accuracy at 100-th epoch while Swin-Tiny has 73.9% top-1 accuracy. Similarly, Focal-Small achieves 78.3% at 100-th epoch, which is 2.0 point higher than Swin-Small. Even for the base models, this gap is still maintained for a

Architecture	Scale	Locality	Model	#Params. (M)	FLOPs (G)	Top-1 (%)
Convolutional Neural Network	Multiple	Local	ResNet-50 [34]	25.0	4.1	76.2
			ResNet-101 [34]	45.0	7.9	77.4
			ResNet-152 [34]	60.0	11.0	78.3
		Local +Global	SE-ResNet-50 [36]	28.1	8.2	77.5
			SE-ResNet-101 [36]	49.3	15.6	78.4
			SE-ResNet-152 [36]	66.8	23.1	78.9
			CBAM-ResNet-50 [66]	28.1	3.9	77.3
			CBAM-ResNet-101 [66]	49.3	7.6	78.5
			AttAug-ResNet-50 [2]	25.8	8.3	77.7
			AttAug-ResNet-101 [2]	45.4	16.1	78.1
			AttAug-ResNet-152 [2]	61.6	23.8	79.1
			BotNet-S1-59 [52]	33.5	7.3	81.7
			BotNet-S1-110 [52]	54.7	10.9	82.8
Transformer	Single	Global	ViT-B/16 [23]	86.6	17.6	77.9
			ViT-L/16 [23]	307	190.7	76.5
			DeiT-S/16 [57]	22.0	4.6	79.9
			DeiT-B/16 [57]	86.6	17.5	81.8
			TNT-S [32]	23.8	5.2	81.3
			TNT-B [32]	65.6	14.1	82.8
			CPVT-S [18]	23.0	4.6	81.5
			CPVT-B [18]	88.0	17.6	82.3
			DeepViT-S [89]	27.0	6.2	82.3
			DeepViT-L [89]	55.0	12.5	83.1
			CaiT-S36 [58]	68.0	13.9	83.3
			LeViT-256 [30]	18.9	1.1	81.6
			LeViT-384 [30]	39.1	2.3	82.6
	Multiple	Global	T2T-ViT-19 [75]	39.2	8.9	81.9
			T2T-ViT-24 [75]	64.1	14.1	82.3
			CrossViT-S [12]	26.7	5.6	81.0
			CrossViT-B [12]	104.7	21.2	82.2
			PVT-S [63]	24.5	3.8	79.8
			PVT-M [63]	44.2	6.7	81.2
			PVT-L [63]	61.4	9.8	81.7
			CvT-13 [67]	20.0	4.5	81.6
			CvT-21 [67]	32.0	7.1	82.5
		Local	ViL-S [80]	24.6	5.1	82.0
			ViL-M [80]	39.7	9.1	83.3
			ViL-B [80]	55.7	13.4	83.2
			Swin-T [44]	28.3	4.5	81.2
			Swin-S [44]	49.6	8.7	83.1
			Swin-B [44]	87.8	15.4	83.4
		Local +Global	Twins-SVT-S [17]	24.0	2.8	81.3
			Twins-SVT-B [17]	56.0	8.3	83.1
			Twins-SVT-L [17]	99.2	14.8	83.3
			Focal-T (Ours)	29.1	4.9	82.2
			Focal-S (Ours)	51.1	9.1	83.5
			Focal-B (Ours)	89.8	16.0	**83.8**

Table 11: Full comparison of image classification on ImageNet-1k for different model architectures. We split the methods into two super-groups which use CNNs or Transformers as the main skeleton. Note that they are inclusive to each other in some methods.

long duration until the end of the training. We attribute this faster learning speed to the long-range interactions introduce by our focal attention mechanism in that it can help to capture the global information at very beginning.

Attention scores for different token types. In our main submission, we have shown both local and global attentions are important. Here, we study how much local and global interactions occur at each layer. Using Focal-Tiny trained on ImageNet-1K as the target, we show in Fig. 7 the summed up attention scores for three type of tokens: 1) local tokens inside the window; 2) local tokens surrounding the window and 3) global tokens after the window pooling. To compute these scores, we average over the all local windows and then also take the average over all heads. Finally, we sum up the attention scores that belongs to the aforementioned three type of tokens. These attention scores

Backbone	#Params (M)	FLOPs (G)	RetinaNet 1x schedule						Mask R-CNN 1x schedule					
			AP^b	AP^b_{50}	AP^b_{75}	AP_S	AP_M	AP_L	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
ResNet50 [34]	37.7/44.2	239/260	36.3	55.3	38.6	19.3	40.0	48.8	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small[63]	34.2/44.1	226/245	40.4	61.3	43.0	25.0	42.9	55.7	40.4	62.9	43.8	37.8	60.1	40.3
ViL-Small [80]	35.7/45.0	252/174	41.6	62.5	44.1	24.9	44.6	56.2	41.8	64.1	45.1	38.5	61.1	41.4
Swin-Tiny [44]	38.5/47.8	245/264	42.0	63.0	44.7	26.6	45.8	55.7	43.7	66.6	47.7	39.8	63.3	42.7
Focal-Tiny (Ours)	39.4/48.8	265/291	**43.7**	**65.2**	**46.7**	**28.6**	**47.4**	**56.9**	**44.8**	**67.7**	**49.2**	**41.0**	**64.7**	**44.2**
ResNet101 [34]	56.7/63.2	315/336	38.5	57.8	41.2	21.4	42.6	51.1	40.4	61.1	44.2	36.4	57.7	38.8
ResNeXt101-32x4d [70]	56.4/62.8	319/340	39.9	59.6	42.7	22.3	44.2	52.5	41.9	62.5	45.9	37.5	59.4	40.2
PVT-Medium [63]	53.9/63.9	283/302	41.9	63.1	44.3	25.0	44.9	57.6	42.0	64.4	45.6	39.0	61.6	42.1
ViL-Medium [80]	50.8/60.1	339/261	42.9	64.0	45.4	27.0	46.1	57.2	43.4	65.9	47.0	39.7	62.8	42.1
Swin-Small [44]	59.8/69.1	335/354	45.0	66.2	48.3	27.9	48.8	59.5	46.5	68.7	51.3	42.1	65.8	45.2
Focal-Small (Ours)	61.7/71.2	367/401	**45.6**	**67.0**	**48.7**	**29.5**	**49.5**	**60.3**	**47.4**	**69.8**	**51.9**	**42.8**	**66.6**	**46.1**
ResNeXt101-64x4d [70]	95.5/102	473/493	41.0	60.9	44.0	23.9	45.2	54.0	42.8	63.8	47.3	38.4	60.6	41.3
PVT-Large[63]	71.1/81.0	345/364	42.6	63.7	45.4	25.8	46.0	58.4	42.9	65.0	46.6	39.5	61.9	42.5
ViL-Base [80]	66.7/76.1	443/365	44.3	65.5	47.1	28.9	47.9	58.3	45.1	67.2	49.3	41.0	64.3	44.2
Swin-Base [44]	98.4/107	477/496	45.0	66.4	48.3	28.4	49.1	60.6	46.9	69.2	51.6	42.3	66.0	45.5
Focal-Base (Ours)	100.8/110.0	514/533	**46.3**	**68.0**	**49.8**	**31.7**	**50.4**	**60.8**	**47.8**	**70.2**	**52.5**	**43.2**	**67.3**	**46.5**

Table 12: COCO object detection and segmentation results with RetinaNet [42] and Mask R-CNN [34] trained with 1x schedule. This is a full version of Table 3. The numbers before and after "/" at column 2 and 3 are the model size and complexity for RetinaNet and Mask R-CNN, respectively.

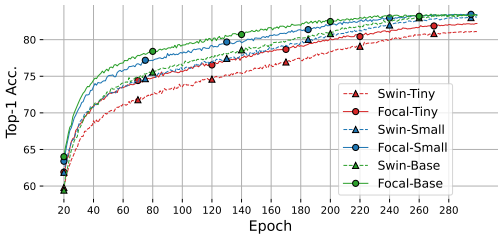


Figure 6: Training curves (Top-1 validation Acc.) for image classification with Swin Transformers and our Focal Transformers.

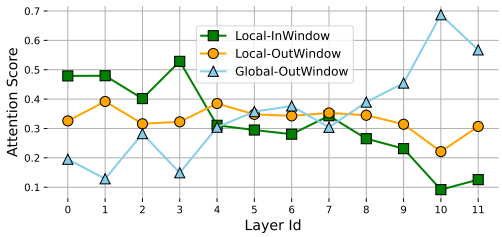
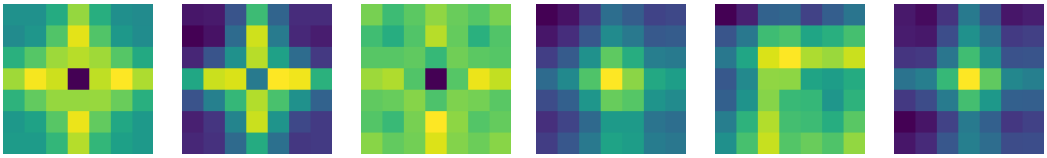


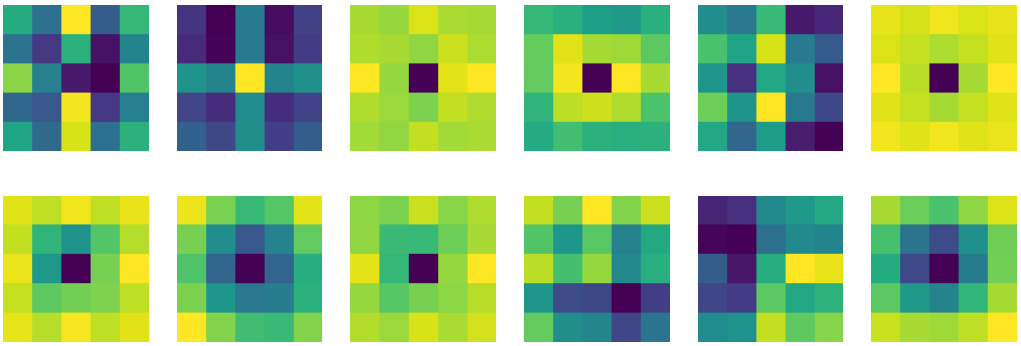
Figure 7: Summed up attention score at each layer for: a) local tokens inside window; b) local tokens surrounding window and c) global tokens.

are further averaged over the whole ImageNet-1K validation set. In Fig. 7, we can see a clear trend that the global attention becomes stronger when it goes to upper layers, while the local attention inside a window is weakened gradually. This indicates that: 1) our model heavily relies on both short- and long-range interactions. Neither of them are neglected in the model at all layers and stages; 2) the gradually strengthened global and weakened local attentions indicate that model tends to focus on more local details at earlier stages while on more global context at the later stages.

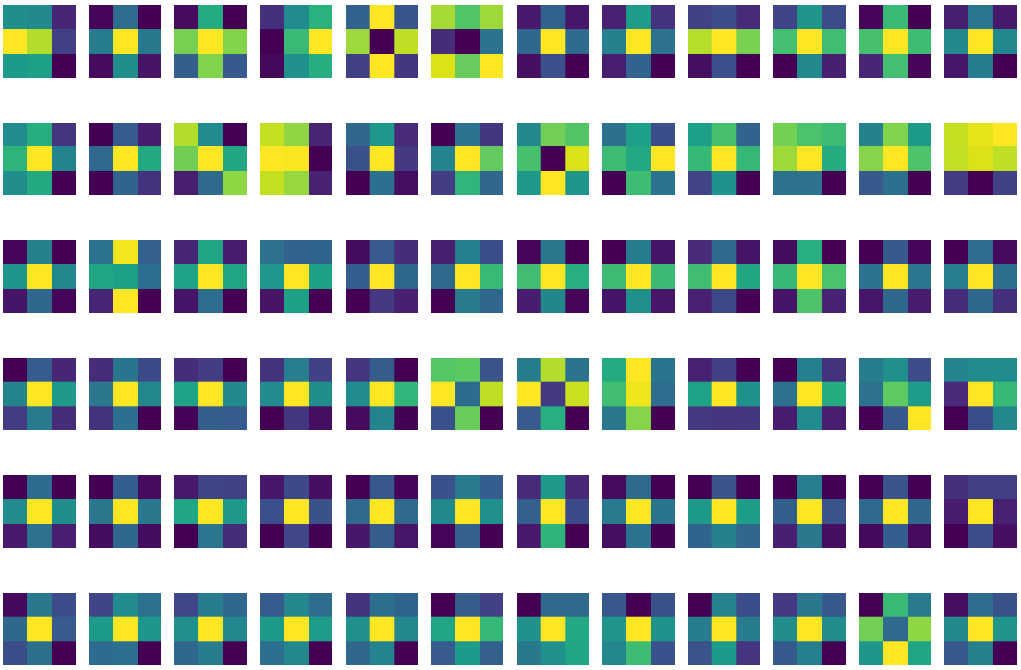
Local-to-global relative position bias. We further inspect what our model learns for the local to global relative position bias introduced in Eq. (3). This relative position bias is a good indicator on how the model put its attention weight on local and global regions. In our Focal Transformers, the focal region sizes at four stages are $(7, 5, 3, 1)$ and $(15, 13, 9, 7)$ for image classification and object detection, respectively. In Fig. 8 and Fig. 9, we visualize the learned relative position bias matrices for all heads and all layers in our Focal-Tiny model trained on ImageNet-1K and COCO, respectively. Surprisingly, though all are randomly initialized, these relative position biases exhibit some interesting patterns. At the first stage of image classification model, all three heads learn to put much less attention on the center window at first layer while focus more on the center at the second layer. For object detection model, however, they are swapped so that the first layer focus more on the center part while the second layer learns to extract the global context from surrounding. As a result, these the two layers cooperate with each other to extract both local and global information. At the second stage of both models, we observe similar property that the two consecutive layers have both local and global interactions. Compared with image classification model, the object detection model has more focus on the center regions. We suspect this is because object detection needs to extract more fine-grained information at local regions to predict the object category and location. At the third stage, we can see there is a fully mixture of local and global attentions in both models. Surprisingly, though randomly initialized, some of the heads automatically learn to disregard the center window pooled token which has much redundancy with the fine-grained tokens inside the center window.



(a) Stage 1, left 3 for first layer, right 3 for second layer, size=7 $\times$ 7

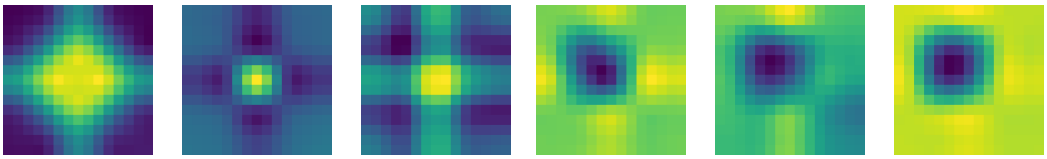


(b) Stage 2, top row for first layer and bottom row for second layer, 6 heads, size=5 $\times$ 5

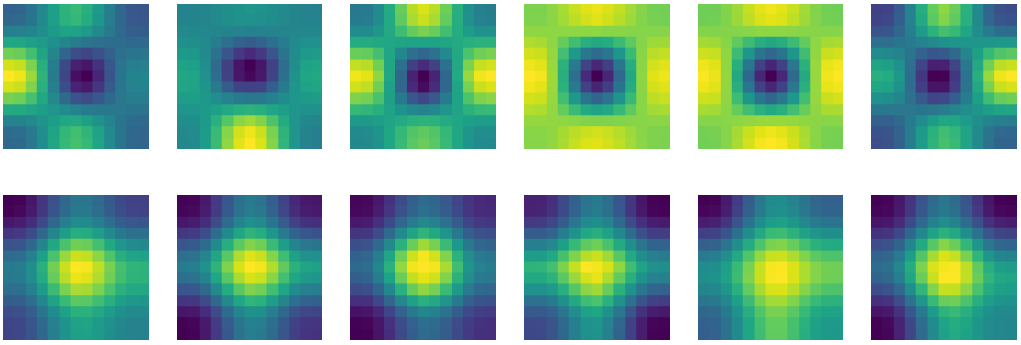


(c) Stage 3, 6 layers from top to bottom row, 12 heads, size=3 $\times$ 3

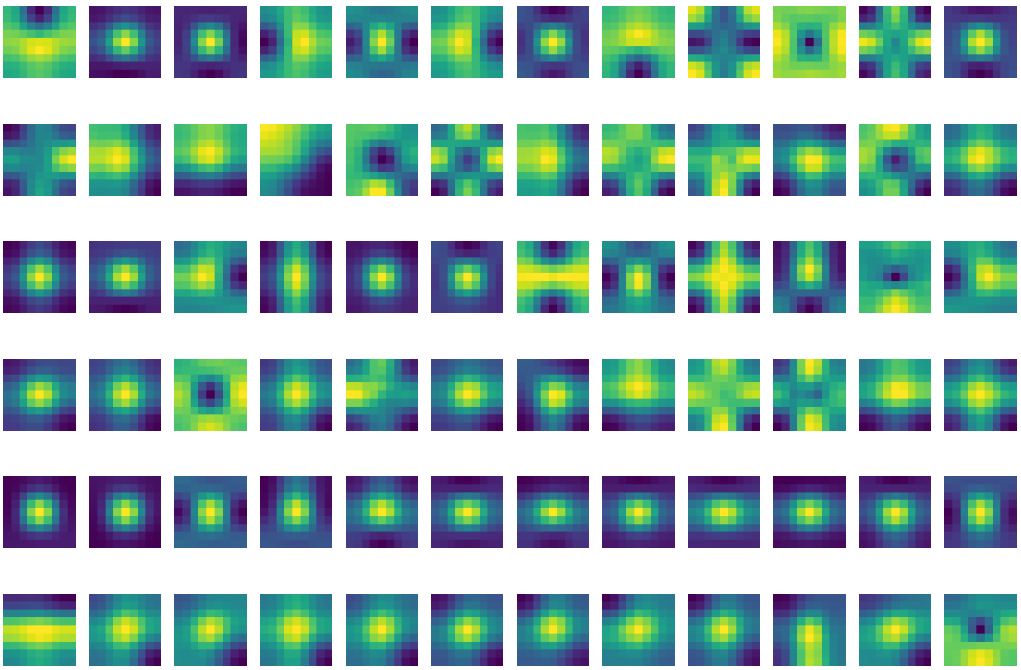
Figure 8: Learned relative position bias between local window and the global tokens in Focal-Tiny trained on ImageNet-1K. From top to bottom, we show the learned relative position bias for all heads at (a) stage 1, (b) stage 2 and (c) stage 3. Since the focal region size is 1 for stage 4 in classification models, we only show the first three stages.



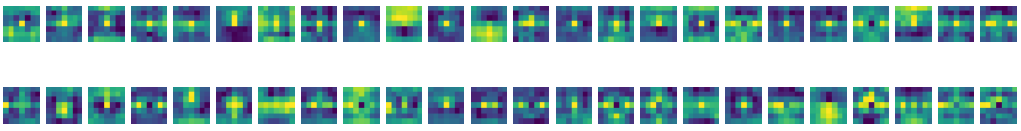
(a) Stage 1, left 3 for first layer and right 3 for second layer, 3 heads, size=15×15



(b) Stage 2, top row for first layer and bottom row for second layer, 6 heads, size=13×13



(c) Stage 3, 6 layers from top to bottom row, 12 heads, size=9×9



(d) Stage 4, top row for first layer and bottom row for second layer, 24 heads, size=7×7

Figure 9: Learned relative position bias between local window and the global tokens in Focal-Tiny for object detection trained on COCO. From top to bottom, we show the relative position bias for different heads at (a) stage 1, (b) stage 2, (c) stage 3 and (d) stage 4.