NomMer: Nominate Synergistic Context in Vision Transformer for Visual Recognition

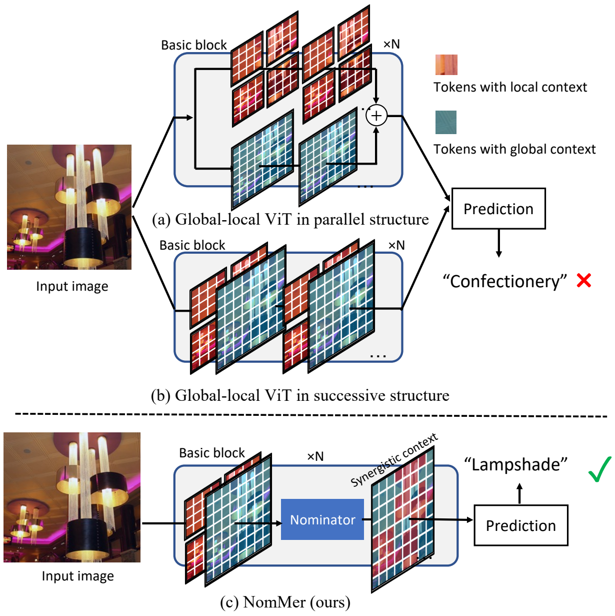
Hao Liu[†*] Xinghua Jiang[*] Xin Li Zhimin Bao Deqiang Jiang Bo Ren
Tencent YouTu Lab
{ivanhliu, clarkjiang, fujikoli, zhiminbao, dqiangjiang, timren}@tencent.com

Abstract

*Recently, Vision Transformers (ViT), with the self-attention (SA) as the de facto ingredients, have demonstrated great potential in the computer vision community. For the sake of trade-off between efficiency and performance, a group of works merely perform SA operation within local patches, whereas the global contextual information is abandoned, which would be indispensable for visual recognition tasks. To solve the issue, the subsequent global-local ViTs take a stab at marrying local SA with global one in parallel or alternative way in the model. Nevertheless, the exhaustively combined local and global context may exist redundancy for various visual data, and the receptive field within each layer is fixed. Alternatively, a more graceful way is that global and local context can adaptively contribute per se to accommodate different visual data. To achieve this goal, we in this paper propose a novel ViT architecture, termed NomMer, which can dynamically **Nom**inate the synergistic global-local context in vision transfor**Mer**. By investigating the working pattern of NomMer, we further explore what context information is focused. Beneficial from this "dynamic nomination" mechanism, without bells and whistles, the NomMer can not only achieve 84.5% Top-1 classification accuracy on ImageNet with only 73M parameters, but also show promising performance on dense prediction tasks, i.e., object detection and semantic segmentation. The code and models are publicly available at https://github.com/TencentYoutuResearch/VisualRecognition-NomMer.*

1. Introduction

In computer vision, Convolutional Neural Networks (CNNs) [6, 9, 11, 21, 23, 35] have served as the *de facto* gold standard for many years. Recently, the Vision Transformer (ViT) [8] and its variants [25, 31] have been proposed challenging the *status quo*. Thanks to the global information communication and content-dependent learning nature of Self-Attention (SA), which is substantively different from local behavior in CNN, ViTs have shown superior performance on many visual recognition tasks.



(a) Global-local ViT in parallel structure

(b) Global-local ViT in successive structure

(c) NomMer (ours)

Figure 1. Illustration of motivation of the proposed NomMer. (a) Global-local ViTs in parallel structure. (b) Global-local ViTs in successive structure. Previous global-local ViTs only focus on fusing global and local contexts while the modulation on them lacks, where the redundant information may have negative impact when recognizing various cases. (c) Our NomMer. When recognizing an object, our method can dynamically yield synergistic context from global-local context through nominator.

However, the ViTs reasoning global dependency of split feature patch embeddings (*a.k.a*, tokens) is computationally expensive. To attack the issues, many recent works, such as SwinT [15], TNT-ViT [38] and HaloNet [27], introduce CNN-like inductive bias (*e.g.*, locality, translation equivari-

[*] Equal contribution. [†] Contact person.

ance) by building the token relations via self-attentions only within local windows and hierarchically aggregating them in a bottom-up manner. These local SA-based ViTs significantly improve the data efficiency, whereas the global contextual relation, especially in the early stage, is abandoned. To remedy this shortage, as illustrated in Fig. 1 (a) and (b), two kinds of "global-local ViTs" take a compromise, where both local and global visual dependencies are incorporated for each token. The first kind [12, 18, 37] marries the context aggregated by local SA or CNN with the one captured by global SA in parallel structure, such as multi-granular connection [24], in the basic block of hierarchical ViT. On the other hand, the local and global context are aggregated alternatively in the second kind [32,39]. However, no matter in the parallel or successive manner, these global-local ViTs only focus on fusing global-local context while the modulation on it lacks, where the redundant information may have negative impact when recognizing various cases. For example, the "lampshade" is mis-recognized as "confectionery" by previous method, probably due to redundant contextual clues, such as colored lights, misleading the model.

According to a pioneering research [22], the human visual system can simultaneously process both *peripheral* and *foveal* vision when perceiving the real world scene, which also exhibits the intriguing property of modulation under various scenes. As a result, the redundant information can be naturally ignored. Specifically, foveal vision focuses on a local region of interest with more visual details, such as fine-grained details, colors or textures of objects in the scene, while peripheral vision refers to the one viewed at large angles but containing rough global scene information.

Inspired by the above preliminary and observations, in this paper, we regard the locally aggregated context in global-local ViT model as the foveal vision, while the globally aggregated one is treated as the peripheral-like visual information. Moreover, we propose a more effective context leverage strategy that the useful dependency information is nominated from local and global context dynamically. To achieve this nomination process, we need to solve the following two non-trivial problems: (i) *How to make nominated global and local context work in harmony when processing different visual cases?* (ii) *How to preserve the information at most without increasing computational cost evidently when reasoning global dependency?*

For the first problem, a straightforward way of nomination is to directly hard-sample from global or local context. Unfortunately, this sample process is indifferentiable, which makes the model suffer from the gradient lost problem. To overcome this issue, we coin a novel learnable Synergistic Context Nominator (SCN) to dynamically yield the global-local context with synergy for each spatial location, which is vividly shown by Fig. 1 (c). Additionally, as for the second global context reasoning problem, most of previous methods [12, 18] adopt the pooling or bilinear downsampling before conducting global SA-based context aggregation to strike favorable efficiency/performance tradeoffs. Nevertheless, this naive computational simplification may remove both redundant and salient information, leading to the detriment on performance. Considering there exists tremendous redundancy in natural images containing most smooth signals with high frequency noise, we build a Compressed Global Context Aggregator (CGCA) upon Discrete Cosine Transform (DCT) [1] to reason the global dependency with redundancy reduced from the frequency domain but without increasing computational complexity. This global context reasoning mechanism is also surprisingly consistent with the working behavior of human visual system [22] when processing peripheral vision.

Based on the SCN and CGCA submodules, we carve out the basic Transformer blocks and stack them into our NomMer framework, which can dynamically **Nom**inate the synergistic context in vision transfor**Mer** for various visual data. We have experimentally verified the effectiveness of our proposed method on image classification task as well as dense prediction tasks, *i.e.*, object detection and semantic segmentation. Our contributions can be summarized as follows:

1) We propose a novel learnable Synergistic Context Nominator (SCN) in terms of aggregated context, which is in stark contrast to previous ViTs with global-local context greedly aggregated.

2) We coin a novel Compressed Global Context Aggregator (CGCA) more effective to reduce global redundancy and to capture global correlations.

3) We propose a novel ViT framework, termed NomMer, which enables the nominated global-local context complementary with each other for various cases and tasks. We also investigate the working behavior of SCN in NomMer.

4) Thanks to the "nomination" mechanism, the NomMer can achieve 84.5% Top-1 classification accuracy on ImageNet with only 73M parameters. With fewer parameters, our small and tiny models can still achieve 83.7% and 82.6% accuracy respectively. We also witness the promising performance of NomMer on dense prediction tasks, *i.e.*, object detection and semantic segmentation.

2. Related Work

2.1. Vision Transformer

Global Vision Transformer. The vanilla Vision Transformer (ViT) [8] greedily reasons the global dependency of visual tokens throughout the whole model. However, it may suffer from the slow model convergence and expensive computational cost. To alleviate it, DeiT [25] exploits distillation technique while PVT [31] introduces the pyramid structure [13] into ViT.

Local Vision Transformer. Due to the lacking of inductive bias and the high computational complexity of SA calculated in global range, the inherent shortages of global ViT cannot be totally eliminated. Therefore, many subsequent ViTs alternatively propose to limit the token relation building by Self-Attention (SA) only within local regions. Among, SwinT [15] coins basic block with the successive WMHSA and shifted ones to perform within- and cross-window information communication. TNT-ViT [38] proposes to represent the local structure by recursively aggregating neighboring Tokens into one Token. In the similar spirit, HaloNet [27] introduces a non-centered local attention and extends it with "haloing" operation. Although the local SA-based ViTs significantly reduce the model complexity and improve the data efficiency, whereas the global contextual relation, especially in the early stage, is abandoned.

Global-Local Vision Transformer. To overcome the "global contextual information lost" issue, many global-local ViTs [12, 18, 32, 37, 39] attempt to seek an equilibrium between the *fully global* and *fully local* contextual information leverage. Specifically, the literature [37] proposes a focal SA incorporates both fine-grained local and coarse-grained global interactions. Conformer [18] fuses local features and global representations under different resolutions based on the Feature Coupling Unit. Other than the above global-local ViTs with parallel structure, NesT [39] stacks the local SA and CNN-bsed global aggregation modules alternatively while the basic block consisting of successive global SA and local CNN are designed in the CVT [32].

2.2. Redundancy Reduction Methods

The design of local [27, 38] and global-local ViTs [39] can be regarded as the redundancy reduction in terms of architecture. By contrast, our proposed architecture can dynamically harmonize the local and global context through nomination mechanism. Moreover, OctConv [6] is a pioneering successful attempt on feature redundancy reduction via separating the CNN feature into low- and high-frequencies, which is realized by down-sampling and up-sampling. DRConv [5] and DynamicViT [20] are another two representative works. The DRConv proposes to dynamically select the CNN filters whereas there is still local context exploited, while the DynamicViT dynamically sparsifies tokens, which may underperform on dense prediction tasks due to the attenuation of fine-grained local interactions. The DCTransformer [17] transits the view of solving the problem into frequency domain and demonstrates the sparse representations can carry sufficient information for generating images. Similarly, the work [36] also converts the input image into frequency domain for visual understanding. The intriguing property of frequency domain motivates our method aggregating global context at feature-level in frequency domain, which is different from above both works only operating input images.

3. Methodology

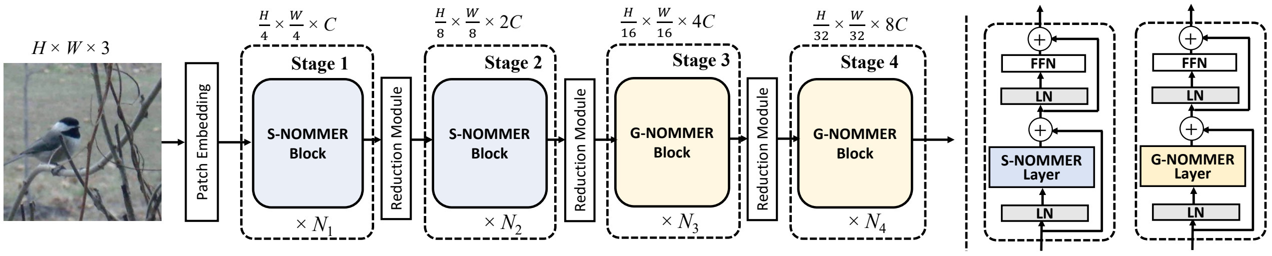
3.1. Architecture Overview

The overview of the proposed architecture is shown in Fig. 2 (a), which follows the hierarchical design in other ViT models [15,18,31,37]. It mainly consists of four stages with "Reduction Modules" dispersed between adjacent stages to down-sample the token numbers and increase the channels by factor 2. Each reduction module is composed of a 3×3 convolutional layer and a max pooling layer with stride 2.

Before being sent to the first stage, the image in size of $H \times W \times 3$ is split into patches of $4 \times 4 \times 3$ size, which are then projected by "patch embedding" consisting of convolutional layer with $4 \times 4 \times C$ kernel and 4 stride size into $\frac{H}{4} \times \frac{W}{4}$ visual tokens in C dimension. We coin two kinds of NomMer basic blocks, *i.e.*, Synergistic NomMer (S-NomMer) and Global NomMer (G-NomMer) blocks in our proposed architecture. Stage 1 and 2 contains N_1 and N_2 S-NomMer blocks, which are responsible for capturing the synergistic context from global-local context. As the feature maps in stage 3 and 4 become relatively small in spatial size ($\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$), the global context and local context could become homogeneous and the information is highly abstracted. Therefore, inside the Stage 3 and 4, we build the G-NomMer layers upon Global Self-Attention (SA), which pay more attention on capturing the semantic information in global range. As is shown in Fig. 2 (b), both S-NomMer and G-NomMer layers are all equipped with Feed-Forward Networks (FFN) [28], residual connection and Layer Normalization (LN) [28]. As the S-NomMer layer is the core component of our method, we will elaborate it in the next subsections.

3.2. Synergistic NomMer Layer

To implement the "dynamic nomination" of S-NomMer layer serving as the core ingredient in our proposed ViT model, our design mainly concentrates on two vital aspects: 1) *generating global and local context*; 2) *nominating synergistic context from them for each visual tokens*. Generally, as illustrated in Fig. 3, S-NomMer layer has three trainable submodules, *i.e.*, Compressed Global Context Aggregator (CGCA), Local Context Aggregator (LCA) and Synergistic Context Nominator (SCN). The CGCA provides rough global context while LCA yields more detailed contextual information within local region. By evaluating the contributions of local or global context aggregated to each tokens, SCN flexibly modulates them and build the nomination map to mask out the features with synergistic context.



(a) Architecture

(b) S-NOMMER and G-NOMMER blocks

Figure 2. (a) Architecture of the proposed NomMer; (b) NomMer basic block consisting of NomMer layer, Layer Normalization (LN) and Feed-Forward Network(FFN) with residual connections.

3.2.1 Compressed Global Context Aggregator

As introduced in Sec. 1, we expect the peripheral-like global context can provide the rough visual clue in terms of the scene, where more fine-grained information can be supplemented from local context. Therefore, the feature maps including visual tokens for global context aggregation should simultaneously satisfy: *sparse* and *informative*. To obtain the sparse representation with the spatial redundancy reduced, we propose a novel Compressed Global Context Aggregator (CGCA) to convert the spatial feature maps containing visual tokens into frequency domain and selectively preserve the low frequency information for global context reasoning.

Unlike many previous methods, such as OctConv [6], conducting down-sampling on features, we aim at seeking the best trade-off between redundancy reduction and useful information preservation. The idea behind our method is mostly inspired by the spirit of JPEG codec [29] leveraging the discrete cosine transform (DCT) to separate spatial frequencies from image. Specifically, as shown by "Compressed Global Context Aggregator" branch in Fig. 3, the input feature $\mathbf{F} \in \mathbb{R}^{D\times D\times C}$ is partitioned into blocks of $N \times N$ size $\{\mathbf{F}^{(m,n)}, m = 1,2,...,\frac{D}{N}, n = 1,2,...,\frac{D}{N}\}, \mathbf{F}^{(m,n)} \in \mathbb{R}^{N\times N\times C}$, where each channel of $\mathbf{F}^{(m,n)}$ is applied by 2D-DCT to obtain a serious of corresponding frequency blocks $\{\mathbf{f}^{(m,n)}, m = 1,2,...,\frac{D}{N}, n = 1,2,...,\frac{D}{N}\}, \mathbf{f}^{(m,n)} \in \mathbb{R}^{N\times N\times C}$, which is denoted as:

$$f(i,j) = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1} c(u)c(v) \cdot F_{u,v} \cos\left[\frac{(i+0.5)\pi}{N}u\right] \cos\left[\frac{(j+0.5)\pi}{N}v\right], \quad (1)$$

$$c(\lambda) = \begin{cases} \sqrt{\frac{1}{N}}, & \lambda = 0 \\ \sqrt{\frac{2}{N}}, & \lambda \neq 0 \end{cases}, \quad (2)$$

where $F_{u,v}$ represent the pixel with index (u,v) in $\mathbf{F}^{(m,n)}$ while the i and j are indexes of horizontal and vertical spatial frequencies in frequency block $\mathbf{f}^{(m,n)}$. $c(\cdot)$ is the normalization scale factor ensuring the orthogonality.

Afterwards, the redundancy reduction is achieved by the low-frequency perceiver (LFP) module. In detail, LFP first drops the high-frequencies of each frequency block in proportion α, which is set to 0.5 in default:

$$\hat{\mathbf{f}}^{(m,n)} = \{\mathbf{f}^{(m,n)}(i,j)\}, i,j \in \{1,2,...,l\}, \quad (3)$$

$$l = \lfloor \alpha N \rfloor. \quad (4)$$

then the low frequency map is obtained by flattening each $\hat{\mathbf{f}}^{(m,n)}$ into a vector in $l^2 \cdot C$ dimension. To further extract useful frequencies while reducing dimensions, a convolutional layer with $1 \times 1 \times C$ kernel is applied to obtain the compressed frequency map $\hat{\mathbf{f}} \in \mathbb{R}^{\frac{D}{N}\times\frac{D}{N}\times C}$. As the redundancy is only reduced in the frequency domain while the spatial information still preserved, it is feasible to perform global context aggregation by using global multi-head self-attention (G-MHSA):

$$\mathbf{f}^{(\hat{G})} = \mathit{Conv}(\mathit{G\text{-}MHSA}(\hat{\mathbf{f}})), \quad (5)$$

where $\mathbf{f}^{(\hat{G})} \in \mathbb{R}^{\frac{D}{N}\times\frac{D}{N}\times N^2\cdot C}$ and "$\mathit{Conv}(\cdot)$" is a convolutional layer with $1 \times 1 \times N^2 \cdot C$ kernel size. Then, $\mathbf{f}^{(\hat{G})}$ is reshaped to the tensor with shape $\frac{D}{N} \times \frac{D}{N} \times N \times N \times C$, where each block has the same shape $(N \times N \times C)$ with $\mathbf{F}^{(m,n)}$. To project the frequency maps with compressed global context back to the spatial domain, the 2D-IDCT (Inverse DCT) is conducted within each channel of each block $\hat{\mathbf{f}}^{(G)}_{(m,n)}$ according to:

$$F^{(G)}_{u,v} = \sum_{u=0}^{N-1}\sum_{v=0}^{N-1} c(u)c(v) \cdot f^{(G)}(i,j) \cos\left[\frac{(i+0.5)\pi}{N}u\right] \cos\left[\frac{(j+0.5)\pi}{N}v\right], \quad (6)$$

where $F^{(G)}_{u,v}$ represent the pixel with index (u,v) in a restored spatial feature block while the $f^{(G)}(i,j)$ are spatial frequencies in one frequency block of $\mathbf{f}^{(\hat{G})}$. $c(\cdot)$ is the normalization scale factor given in Eqn. (2). Finally, spatial feature with compressed global context $\mathbf{F}^{(G)}$ is yielded.

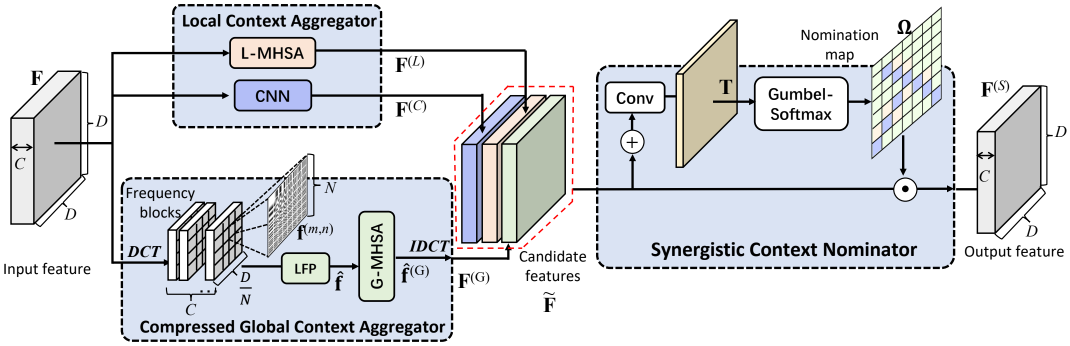


Figure 3. Illustration of detailed Synergistic NomMer layer. Best viewed in color.

3.2.2 Local Context Aggregator

Local Context Aggregator (LCA) plays as the role of aggregating local context with more foveal vision-like details serving as the complementary visual information, which is expected to collaborate with compressed global context. To reach this goal, as illustrated in Fig. 3, we adopt two kinds of LCAs in this work, *i.e.*, Local-MHSA (L-MHSA) and CNN. As suggested by many prevalent researches [18, 39], CNN introduced in the ViT model can provide more inductive bias helping model convergence. Essentially, the CNN is the content-independent aggregator while the MHSA is the content-dependent one. The combination of CNN and MHSA can be deemed as a trade-off between leveraging pre-defined inductive bias or learning it from data. However, most of previous methods adopting parallel [18] or subsequent [39] combination manner can only have single acquisition *w.r.t.* the inductive bias within one layer. Contrastively, a more elegant way is to determine the inductive bias usage according to the specific region, which motivates our design of LCA providing both CNN feature $\mathbf{F}^{(C)} \in \mathbb{R}^{D\times D\times C}$ and L-MHSA feature $\mathbf{F}^{(L)} \in \mathbb{R}^{D\times D\times C}$ for dynamic nomination. More concretely, we inherit the window-based self-attention (W-MSA) with M window size from SwinT [15] as L-MHSA. On the other hand, we adopt "Bottleneck" [9] with structure $\{\text{Conv}_{1\times1}$, $\text{Conv}_{3\times3}$, $\text{Conv}_{1\times1}\}$ as the CNN aggregator.

3.2.3 Synergistic Context Nominator

Once the compressed global and local context features obtained, they are treated as the candidate features $\tilde{\mathbf{F}} = \{\mathbf{F}^{(L)}, \mathbf{F}^{(C)}, \mathbf{F}^{(G)}\} \in \mathbb{R}^{3\times D\times D\times C}$ (highlighted by red dotted boundary in Fig. 3) for the proposed Synergistic Context Nominator (SCN). The aim of SCN is to pick up the most valuable context with synergy for each spatial location, where the detailed process is described in Fig. 3. The candidate features are first fused by element-wise addition operation and passed to a convolutional layer with $1\times1\times3$ kernel to obtain a tensor $\mathbf{T}$ of shape $D\times D\times 3$, where each channels in the last axis describes the probability of nominating either one type context feature from three candidates. To further acquire the nomination map, we can perform the element-wise hard sampling on the tensor $\mathbf{T}$:

$$\omega_{i,j} = argmax(\tau_{i,j,1}, \tau_{i,j,2}, \tau_{i,j,3}), \tag{7}$$

where $\tau_{i,j,c}$ is the (i,j)-th element in c-th channel of $\mathbf{T}$, $\omega_{i,j}$ is the channel index of nominated context feature at (i,j) spatial location. Correspondingly, the nomination map $\mathbf{\Omega} \in \mathbb{R}^{D\times D\times 3}$ can be determined. For example, if the L-MHSA context feature is nominated at position (i,j), the corresponding nomination vector $\mathbf{\Omega}_{i,j}$ is $[1,0,0]$.

One challenge lying in the proposed SCN module is that the hard sampling process is not differentiable whereas the weights in SCN need to be updated during training. To tackle this issue, we introduce a reparameterization method, termed Gumbel Softmax trick [10], which allows the gradients to back-propagate through the discrete sampling process. In the end, the feature map $\mathbf{F}^{(S)} \in \mathbb{R}^{D\times D\times C}$ with nominated synergistic context is output by masking the nomination map on the candidate feature set:

$$\mathbf{f}^{(S)}_{i,j} = \sum\nolimits_{p=1}^{3} \mathbf{\Omega}_{i,j}\tilde{\mathbf{f}}_{p,i,j}, \tag{8}$$

where $\mathbf{f}^{(S)}_{i,j}$ is the (i,j)-th feature vector of $\mathbf{F}^{(S)}$ and p is the type index of context feature at location (i,j).

3.3. NomMer Variants

To fully explore the potential of NomMer with different configurations, we build several variants of it, *i.e.*, NomMer-T, NomMer-S and NomMer-B, which refer to tiny, small and base model separately. The detailed configurations are given in supplementary material.

4. Experiments

4.1. Image Classification on ImageNet-1K

Experimental setting. We compare our proposed NomMer against several baselines on ImageNet-1K [7]. For a fair comparison, we follow the experimental settings in [25]. Concretely, all our models are pre-trained for 300 epochs with the input size of 224^2. The initial learning rate and batch size are set to 10^{-3} and 1,024, respectively. For optimization, AdamW [16] optimizer with a cosine learning rate scheduler is used. The weight decay is set to 0.05 and the maximal gradient norm is clipped to 5.0. We also inherit the data augmentation and regularization techniques from [25]. The stochastic depth drop rates are set to 0.1, 0.3 and 0.5 for tiny, small and base models individually. When reporting the results of 384^2 input, we fine-tune the models with a total batch size of 512 for 30 epochs. The learning rate and weight decay are 10^{-5} and 10^{-8}.

Performance. In Tab. 1, we compare our NomMer with state-of-the-art CNN and Transformer architectures on image classification task. The results show that the proposed NomMer consistently outperforms other approaches with similar model size and computational budgets. Compared with CNN-based RegNetY [19], our models achieve improvements ranging from 1.6% to 2.6% under three configurations with input image in size of 224^2. Compared with the ViT models, the proposed NomMer also witnesses the superior performance. More specifically, our NomMer-B with 73M parameters achieves a 84.5% ImageNet Top-1 accuracy, surpassing DeiT-B [25], Swin-B [15] and Conformer-B [18] by 2.7%, 1.2% and 0.4%, respectively. Besides, the lightweight version (NomMer-T) also achieves the best performance. When finetuned on the 384^2 images, the similar trend is also observed. In addition, we further verify the effectiveness of our proposed method pre-trained on the larger ImageNet-21K dataset (see supplementary material).

4.2. Object Detection on COCO

Experimental setting. To verify NomMer's versatility, we benchmark it on object detection with COCO 2017 [14]. The models pretrained on ImageNet-1K [7] are used for initializing the backbone of Cascade Mask R-CNN [3] framework. Similar to SwinT [15], we follow 3× schedule training with 36 epochs for a fair comparison. During training, multi-scale training strategy is employed to randomly resize image's shorter side to the range of [480, 800]. And we use AdamW [16] for optimization with initial learning rate 10^{-4} and weight decay 0.05. In the similar spirit, 0.1, 0.3 and 0.5 stochastic depth drop rates are set to regularize the training for tiny, small and base models.

Performance. The box and mask mAPs on COCO validation set are summarized in Tab. 2, from which we can see that NomMer significantly boosts the AP^b and AP^m. In details, the box's mAP and mask's mAP of NomMer-B are 0.8% and 0.6% higher than that of the strong baseline Swin-B [15], which demonstrates the importance of global representations for high level tasks in our method. When evaluating our tiny model, it surpasses the second best method Focal-T [37] by 0.3%, which also suggests that NomMer can also perform well on the prediction tasks with fewer parameters. To further investigate the versatility of the proposed model when it works in different detection frame-

ImageNet-1K 224^2 trained models			
Method	#param.	FLOPs	Top-1
	(M)	(G)	(%)
RegNetY-4G [19]	21	4.0	80.0
RegNetY-8G [19]	39	8.0	81.7
RegNetY-16G [19]	84	16.0	82.9
NFNet-F0 [2]	72	12.4	83.6
DeiT-S [25]	22	4.6	79.8
DeiT-B [25]	86	17.5	81.8
PVT-S [31]	25	3.8	79.8
PVT-M [31]	44	6.7	81.2
PVT-L [31]	61	9.8	81.7
Swin-T [15]	29	4.5	81.3
Swin-S [15]	50	8.7	83.0
Swin-B [15]	88	15.4	83.3
T2T-ViT$_t$-14 [38]	22	6.1	81.7
T2T-ViT$_t$-19 [38]	39	9.8	82.2
T2T-ViT$_t$-24 [38]	64	15.0	82.6
LG-T [12]	33	4.8	82.1
LG-S [12]	61	9.4	83.3
Focal-T [37]	29	4.9	82.2
Focal-S [37]	51	9.1	83.5
Focal-B [37]	90	16.0	83.8
Conformer-T [18]	24	5.2	81.3
Conformer-S [18]	38	10.6	83.4
Conformer-B [18]	83	23.3	84.1
NesT-T [39]	17	5.8	81.5
NesT-S [39]	38	10.4	83.3
NesT-B [39]	68	17.9	83.8
CvT-13 [32]	20	4.5	81.6
CvT-21 [32]	32	7.1	82.5
CaiT-S [26]	68	13.9	84.0
NomMer-T	22	5.4	**82.6**
NomMer-S	42	10.1	**83.7**
NomMer-B	73	17.6	**84.5**
ImageNet-1K 384^2 finetuned models			
ViT-B/16 [8]	86	49.3	77.9
DeiT-B [25]	86	55.4	83.1
Swin-B [15]	88	47.0	84.2
T2T-ViT$_t$-14 [38]	22	17.1	83.3
CvT-13 [32]	20	16.3	83.0
CvT-21 [32]	32	24.9	83.3
CaiT-S [26]	68	48.0	85.4
NomMer-T	22	17.2	**83.9**
NomMer-S	42	33.1	**84.6**
NomMer-B	73	56.2	84.9

Table 1. Comparison of different backbones on ImageNet-1K classification.

works, we also conduct a series of experiments to compare NomMer with other SOTAs. A more detailed description is given in supplementary material.

Method	#param. (M)	FLOPs (G)	AP^b (%)	AP^b_{50} (%)	AP^b_{75} (%)	AP^m (%)	AP^m_{50} (%)	AP^m_{75} (%)
Res50 [9]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
X101-32 [35]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
X101-64 [35]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-T [15]	86	745	50.5	69.3	54.9	43.7	66.6	47.1
Swin-S [15]	107	838	51.8	70.4	56.3	44.7	67.9	**48.5**
Swin-B [15]	145	982	51.9	70.9	56.5	45.0	68.4	48.7
Focal-T [37]	87	770	51.5	70.6	55.9	-	-	-
NomMer-T	80	755	**51.8**	**70.8**	**56.0**	**44.7**	67.6	**48.1**
NomMer-S	99	851	**52.4**	**71.5**	**56.8**	**45.1**	**68.8**	**48.5**
NomMer-B	130	1006	**52.7**	**71.6**	**57.2**	**45.6**	**68.9**	**49.3**

Table 2. Results on COCO object detection and instance segmentation with Cascade Mask R-CNN.

4.3. Semantic Segmentation on ADE20K

Experimental setting. For another dense prediction task, Semantic Segmentation, we further evaluate our model on the ADE20K [40] dataset. Specifically, our NomMer serves as the backbone of UperNet [33] which is a prevalent segmentation method. Unless explicitly specified, we use a standard recipe by setting the image size to 512^2 and train the models for 160k iterations with batch size 16.

Performance. The results of Upernet with various backbones on the ADE20K [40] dataset are reported in the Tab. 3, where both single-scale and multi-scale evaluation results are included. Obviously, our method significantly outperforms previous state-of-the-arts under different configurations. Under single-scale setting, our NomMer separately achieves 50.0%, 48.7%, and 46.1% mIoU with base, small and tiny model configurations, which are 1.0%, 0.7%, 0.3% higher than the strong baseline Focal [37] counterparts. And we can also observe the consistent performance improvement under the multi-scale evaluation. Conclusively, our NomMer can steadily improve the performance of various visual recognition tasks owing to the synergistic context nomination.

4.4. Ablation Study

To better investigate the effectiveness of different aspects in our proposed S-NomMer layer, we conduct extensive ablation studies on both classification and downstream tasks of which the results are summarized in Tab. 4.

Effect of Local Context Aggregator. We validate the effectiveness of the LCA combing both L-MHSA and CNN. Compared with the performance of “-w/o Global&CNN” only equipped with L-MHSA, the additional CNN (“-w/o Global”) consistently increases 0.3% accuracy on ImageNet [7], 0.2% box’s mAP and 0.3% mask’s mAP on COCO [14], and 0.1% mIoU on ADE20K [40]. This confirms the benefit of inductive bias provided by CNN and also shows the combination can capture the salient fine-grained features through synergy.

Method	#param. (M)	FLOPs (G)	mIoU (%)	+MS (%)
Res101 [9]	86	1029	44.9	-
Swin-T [15]	60	945	44.5	45.8
Swin-S [15]	81	1038	47.6	49.5
Swin-B [15]	121	1188	48.1	49.7
Focal-T [37]	62	998	45.8	47.0
Focal-S [37]	85	1130	48.0	50.0
Focal-B [37]	126	1354	49.0	50.5
LG-T [12]	64	957	45.3	-
NomMer-T	54	954	**46.1**	**47.3**
NomMer-S	73	1056	**48.7**	**50.4**
NomMer-B	107	1220	**50.0**	**51.0**

Table 3. Performance comparison of different backbones with UperNet framework on the ADE20K segmentation task.

Effect of Compressed Global Context Aggregator. As is shown in Tab. 4, we find the performance of all tasks can be consistently improved by integrating the global context into local ones, even if the global context is aggregated from simple max-pooling features. When the aggregation is performed in the frequency domain through our DCT-based CGCA, the average accuracy of each task is further boosted, which indicates our learnable CGCA can well strike the trade-off between redundancy reduction and useful information preservation for the visual recognition tasks.

Effect of Synergistic Context Nominator. To verify the capability of another core component SCN in our architecture, we compare the performance of our NomMer with and without the SCN equipped on each task. As demonstrated in Tab. 4, our model can obtain the best results by adopting nominator, with at least 0.4% improvement on different tasks than the non-nominator version in which various types of contexts are directly fused through element-wise addition. These results further prove the effectiveness of the synergistic context learned by SCN.

4.5. Qualitative Analysis

To further investigate the working pattern of our proposed NomMer, in Fig. 4, we visualize the synergistic nomination maps from intermediate layers of S-NomMer block from base model on classification task. We surprisingly observe that nomination maps exhibit several intriguing properties. In low-level nomination maps (“Layer 1_1”) the CNN context features are always predominant, which is

Method	Local		Global		Nominator	ImageNet	COCO		ADE20K
	L-MHSA	CNN	Pool	DCT	Gumbel	Top-1(%)	AP^b	AP^m	mIoU(%)
NomMer-T* w/o Global&CNN	✓	✗	✗	✗	✗	81.4	50.4	43.5	44.7
NomMer-T* w/o Global	✓	✓	✗	✗	✗	81.7	50.6	43.8	44.8
NomMer-T* w/o Gumbel&DCT	✓	✓	✓	✗	✗	82.0	50.9	44.0	45.1
NomMer-T* w/o Gumbel	✓	✓	✗	✓	✗	82.2	51.2	44.3	45.4
NomMer-T	✓	✓	✗	✓	✓	**82.6**	**51.8**	**44.7**	**46.1**

Table 4. Ablation study of NomMer on three benchmarks based on the NomMer-T architecture.

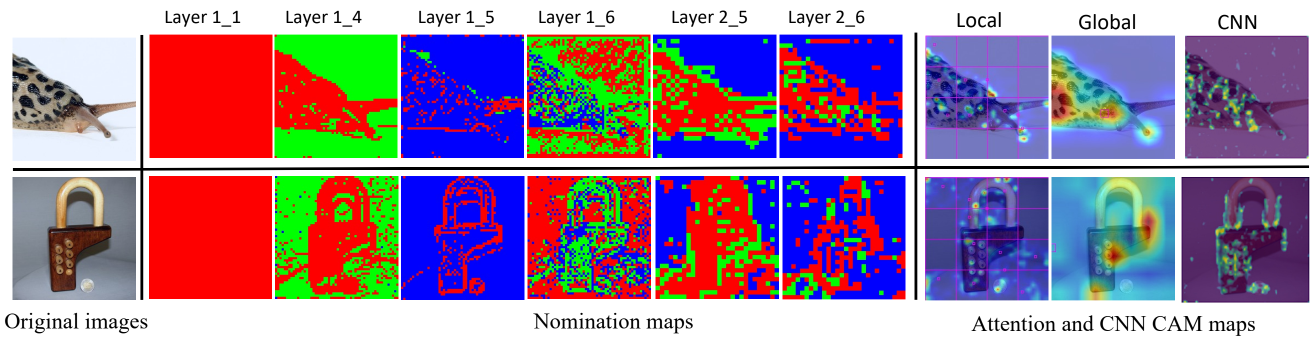


Figure 4. Visualization of nomination maps, attention maps and CNN CAM [30] maps of nominated features from NomMer-B on classification task. **Red**: CNN context $\mathbf{F}^{(C)}$. **Green**: Local context $\mathbf{F}^{(L)}$. **Blue**: Compressed global context $\mathbf{F}^{(G)}$. "Layer B_L" stands for that map is from the L-th NomMer layer of the B-th NomMer block. The pink hollow boxes in attention maps represent the locations of nominated local or global context features. Best viewed in color and zoom in.

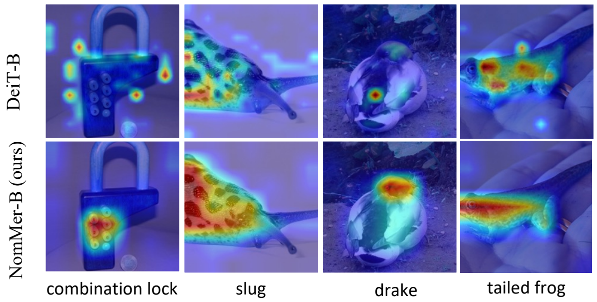


Figure 5. Class activation attention maps on classification task. Best viewed in color.

consistent with the conclusion given in research [34] , that early convolutions help transformers see better. Along with the model going deeper, nomination maps present various synergistic patterns of context in different layers. Specifically, the local context features aggregated by CNN and L-MHSA contribute most in "Layer 1_4", while the adjacent "Layer 1_5" focuses on the global context accompanied with local CNN context. These phenomena demonstrate that our model can successfully obtain synergistic context with redundancy reduced, where the synergy not only happens between local and global context, but also cross layers.

Moreover, one can also observe that the global context features are inclined to be nominated in the smooth regions, *e.g.*, "Layer 1_5" with less color and texture variations. We attribute this behavior to the missing of salient details in smooth regions where model needs to refer more information from larger scope, which is illustrated by the global attention maps in Fig. 4 from "Layer 2_5". Comparatively, there are more fine-grained texture or patterns involved in the local context aggregated by CNN and L-MHSA, which can be demonstrated by the local attention maps and CNN CAM maps (from "Layer 2_5") obtained by algorithm [30].

Thanks to the nomination mechanism, NomMer-B can capture more discriminative features from synergistic context than other ViTs, such as DeiT [25] only exploiting global context. By visualizing attention maps of class activation based on method [4], in Fig. 5, we find that the attentions of NomMer-B often concentrate on the exclusive features of object, such as the keyboard of " combination lock" and the dot pattern of "slug" while DeiT-B presents more unstable activation maps.

5. Conclusion

In this work, we introduced a novel vision transformer architecture solving visual recognition tasks by nominating synergistic context. Extensive experiments on image classification and dense prediction tasks demonstrated its superiority over state-of-the-arts. The visualization of the nomination maps learned by our method can also provide an empirical guidance for the design of architecture, which will be further explored in our future work.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 2

[2] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International Conference on Machine Learning*, pages 1059–1071. PMLR, 2021. 6

[3] 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. 6

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 8

[5] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8064–8073, 2021. 3

[6] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3435–3444, 2019. 1, 3, 4

[7] 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. 5, 6, 7

[8] 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. In *International Conference on Learning Representations*, 2020. 1, 2, 6

[9] 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. 1, 5, 7

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[12] Jinpeng Li, Yichao Yan, Shengcai Liao, Xiaokang Yang, and Ling Shao. Local-to-global self-attention in vision transformers. *arXiv preprint arXiv:2107.04735*, 2021. 2, 3, 6, 7

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[14] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 7

[15] 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. 1, 3, 5, 6, 7

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[17] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 3

[18] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. *arXiv preprint arXiv:2105.03889*, 2021. 2, 3, 5, 6

[19] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10428–10436, 2020. 6

[20] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *arXiv preprint arXiv:2106.02034*, 2021. 3

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[22] Emma EM Stewart, Matteo Valsecchi, and Alexander C Schütz. A review of interactions between peripheral and foveal vision. *Journal of vision*, 20(12):2–2, 2020. 2

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 1

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2, 6, 8

[26] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 6

[27] 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. 1, 3

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[29] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992. 4

[30] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 8

[31] 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. 1, 2, 3, 6

[32] 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. 2, 3, 6

[33] 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. 7

[34] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *arXiv preprint arXiv:2106.14881*, 2021. 8

[35] 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. 1, 7

[36] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020. 3

[37] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2, 3, 6, 7

[38] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, 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. 1, 3, 6

[39] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*, 2021. 2, 3, 5, 6

[40] 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. 7