Local-to-Global Self-Attention in Vision Transformers

Jinpeng Li[*]
Inception Institute of
Artificial Intelligence (IIAI),
Abu Dhabi, UAE
ljpadam@gmail.com

Yichao Yan[*†]
MoE Key Lab of Artificial Intelligence,
AI Institute,
Shanghai Jiao Tong University, China
yanyichao91@gmail.com

Shengcai Liao[†]
Inception Institute of
Artificial Intelligence (IIAI),
Abu Dhabi, UAE
scliao@ieee.org

Xiaokang Yang
MoE Key Lab of Artificial Intelligence,
AI Institute,
Shanghai Jiao Tong University, China
xkyang@sjtu.edu.cn

Ling Shao
Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
ling.shao@ieee.org

Abstract

Transformers have demonstrated great potential in computer vision tasks. To avoid dense computations of self-attentions in high-resolution visual data, some recent Transformer models adopt a hierarchical design, where self-attentions are only computed within local windows. This design significantly improves the efficiency but lacks global feature reasoning in early stages. In this work, we design a multi-path structure of the Transformer, which enables local-to-global reasoning at multiple granularities in each stage. The proposed framework is computationally efficient and highly effective. With a marginal increasement in computational overhead, our model achieves notable improvements in both image classification and semantic segmentation. Code is available at *https://github.com/ ljpadam/LG-Transformer*.

1 Introduction

Over the past decade, we have witnessed the continuous success of convolutional neural networks (CNNs) in computer vision. Recently, Visual Transformers have also demonstrated strong potential, achieving state-of-the-art performance in several visual tasks, including image classification, object detection, and semantic segmentation. Transformers originally emerged from natural language processing (NLP) models and are considered as more flexible alternatives to CNNs for processing multi-modal inputs.

One of the major differences between a CNN and Transformer is the feature interaction mechanism. In a CNN, convolutional kernels are locally connected to the input feature maps, where features only
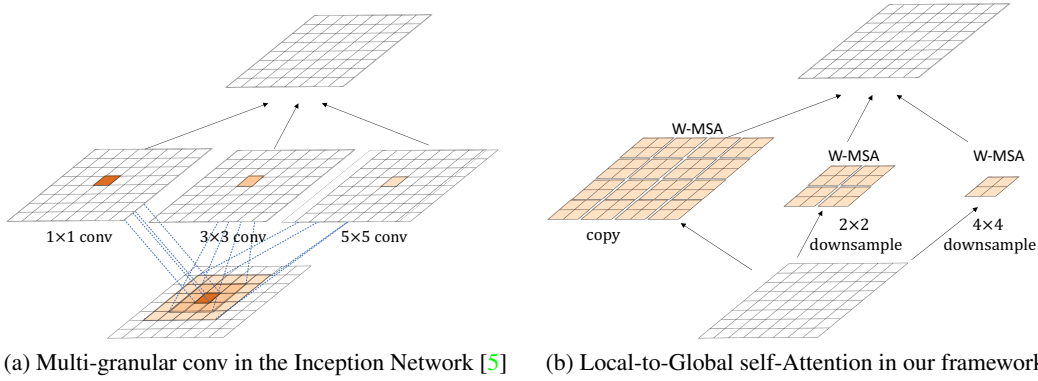
---

[*]*Equal contribution

[†]Corresponding author

(a) Multi-granular conv in the Inception Network [5]   (b) Local-to-Global self-Attention in our framework

Figure 1: Our design is inspired by the multi-granular convolutional structure in CNN models, where 1×1, 3×3 and 5×5 convolution kernels captures different levels of local information. In our framework, feature maps are down-sampled into different scales, and window-based multi-head self attention (W-MSA) is applied to each scale to model local-to-global feature interaction.

interact with their local neighbors. In contrast, Transformer models, such as ViT [1], require global feature reasoning by computing self-attentions among all the tokens. As a result, Transformers are computationally inefficient when processing images with a large number of visual tokens. To address this issue, several recent works, such as Swin Transformer [2], CvT [3], and PVT [4], take inspiration from CNN models, and only compute self-attentions within local windows. This strategy brings significant improvements in efficiency, but abandons global feature reasoning in early stages, which weakens the potential of Transformer models.

In this work, we study the necessity of global and local feature interactions in self-attention modules of Vision Transformers and propose a local-to-global multi-path mechanism in self-attention, which we refer to as the LG-Transformer. Fig. 2 provide an outline of our architecture, which follows the macro-structure of locally connected Transformers [2, 3, 4]. In contrast to these models, in the early stages (stage 1-3), we expand the self-attention module into a multi-path structure, where feature maps (tokens) are downsampled so as to cover different local scales. To enable global reasoning, we set the window size equal to the smallest feature scale (i.e., $\frac{H}{32} \times \frac{W}{32}$). After the multi-path self-attention, the local-to-global features are aggregated as a more discriminative representation for the tokens. As the scales of global features are significantly downsampled at each stage, the computational overhead is marginally increased compared with the single-path locally connected baseline Transformer. In the meantime, the multi-path self-attention structure is easy to implement, and it does not bring in novel operations, which yields an efficient and strong Transformer model.

Our design is inspired by the multi-granular convolutional structure in CNN models, such as Inception Networks [5, 6, 7] and High Resolution Networks [8], where feature interactions are computed within local regions of different scales. Different from these CNN frameworks, our model not only computes the local feature interactions, but also enables global feature reasoning, taking advantage of both CNN and Transformer models. A comparison is illustrated in Fig. 1.

LG-Transformer achieves competitive results on both image classification and semantic segmentation. Compared with the state-of-the-art model Swin-Transformer [2], LG-Transformer improves the Top-1 accuracy by 0.9% on ImageNet [9] classification. Meanwhile, it improves the mIoU by 0.8% on ADE20K [10] for semantic segmentation.

2 Related Work

Transformers and Vision Transformers. The pioneering deep neural network architecture, Transformer [11], was firstly proposed to solve sequence transduction tasks in NLP. The original Transformer is built on self-attention modules which excel at modeling long-range information under arbitrary input lengths. Based on this simple and powerful architecture, several Transformer variants [12, 13, 14, 15] achieve dominant positions in many NLP tasks, such as machine translation and summarization.

Motivated by their great success in NLP, self-attention mechanisms and Transformer have also been introduced to computer vision [16, 17, 18, 19, 20]. The early Transformer-based vision models mainly leverage the advantage of global dependencies. For example, the Vision Transformer (ViT) [1] is a pure Transformer-based image classification model which takes image patches as tokens and directly models the global attention on all the tokens. It achieves comparable performance as hand-designed and automatically searched CNN architectures [21, 20]. However, due to the large resource cost of global attention, ViT is more suitable to generate low-resolution outputs, which limits its applications.

To address this issue, a Pyramid Vision Transformer (PVT) [4] is proposed to reduce the resolutions of keys and values with a spatial-reduction attention, which improves the efficiency while still maintaining the capability of aggregating global information. Although global connectivity in Transformer models shows promising results in computer vision, it is undeniable that existing CNN models also demonstrate the importance of local information. Some works try to combine the advantages of CNN and self-attention by appended blocks [22], interleave architecture [23, 24] and parallel paths [25]. LocalViT [24] employs depth-wise convolution layers into transformers to explicitly model the local dependencies. CvT [3] introduces convolution into token embeddings and feature projection to combine local information and global image contexts. Swin-Transformer [2] explores local information by a pure transformer based architecture which splits feature maps into non-overlapping windows, and applies intra-window and cross-window interactions to build hierarchical attentions.

Multi-granular connection in CNN. Multi-granular information has been widely explored in computer vision since the era of hand-crafted features [26, 27, 28, 29]. Although CNN models [21, 20, 30, 31] demonstrate strong feature representation capability due to the end-to-end learning and deep hierarchical architectures, they still benefit from multi-granular connections in many computer vision applications, such as image classification [32, 5], object detection [33] and semantic segmentation [34, 35]. Inception Networks [5, 6, 7] are a serious of classification models built upon multi-granular convolution kernels to improve the classification performance and computation efficiency. HRNet [8, 36] constructs multi-granular feature maps in parallel paths to simultaneously maintain more semantic and spatial information, and it achieves promising performance on several dense prediction tasks. To detect objects with various sizes, pyramid feature maps are generated by connecting bottom-up and top-down pathways in many CNN based detectors [37, 38, 39]. PSP-Net [40] introduces a pyramid pooling module to fuse multi-granular context information for semantic segmentation methods. Hourglass [41] stacks multiple encoder and decoder networks and builds multi-granular connections among them for human pose estimation.

In this work, we take advantage of both the local connectivity in CNNs and global connectivity in Transformers, building a Transformer model with local-to-global attention, yielding a framework with improved performance and meanwhile maintains high efficiency.

3   Method

In this section, we describe the detailed architecture of the proposed LG-Transformer. As illustrated in Fig. 2, our model follows the hierarchical design of several recent visual Transformers [2, 3, 4] for efficient learning. Our framework contains four stages. In the first stage, patch embeddings are applied to the input image, resulting in $\frac{H}{4} \times \frac{W}{4}$ tokens. In each of the subsequent stages, the resolution of feature maps (tokens) is reduced by a factor of 4 (2×2) with a patch merging layer [2], while the feature dimensions are increased by a factor of 2. Finally, the network outputs $\frac{H}{32} \times \frac{W}{32}$ visual tokens.

3.1   Local-to-Global Attention Block

The core design of our framework is the Local-to-Global (LG) Attention block, applied to stages 1-3. As illustrated in the bottom part of Fig. 2, the LG-Attention block contains several parallel attention paths. Take stage 2 for example. After layer normalization (LN), the input feature z^1 with resolution $\frac{H}{8} \times \frac{W}{8}$ is bilinearly down-sampled into feature maps with lower dimensions (*i.e.*, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$). Each feature map is processed with a window based multi-head self-attention with shifted window partitioning (SW-MSA) [2]. Specifically, feature maps are partitioned into non-overlapped windows, and self-attentions are only calculated within each window, thus the computation complexity is significantly reduced compared performing self-attention on the entire feature map. By displacing
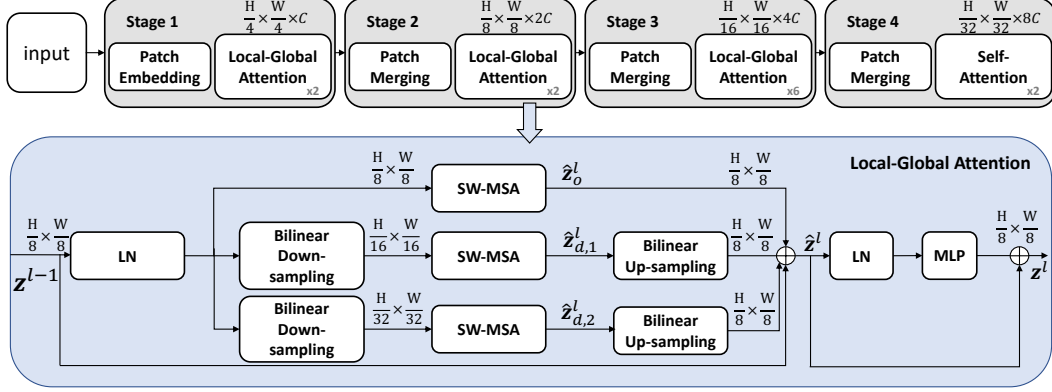
Figure 2: Our Local-to-Global (LG) Transformer contains four stages, each of which includes a patch embedding/merging layer, followed by several LG-Attention blocks. Each LG-Attention block contains several parallel SW-MSA modules to calculate local attentions, gathering local-to-global information with feature interactions.

the window between consecutive attention blocks, SW-MSA enables local patches to communicate with its different neighbors. However, the feature interaction is still limited to a local region. In this work, we set the window size to $[\frac{H}{32}, \frac{W}{32}]$, such that the feature map with the lowest resolution can be processed with global self-attention. After the SW-MSA modules, the downsampled features maps are upsampled to the same resolution and combined, before being fed into the LN and MLP layers. The LG-Attention block can be calculated as:

$$\hat{\mathbf{z}}_o^l = \text{SW-MSA}(\text{LN}(\mathbf{z}^{l-1})), \tag{1}$$

$$\hat{\mathbf{z}}_{d,1}^l = \text{SW-MSA}(\text{BD}_1(\text{LN}(\mathbf{z}^{l-1}))), \tag{2}$$

$$\hat{\mathbf{z}}_{d,2}^l = \text{SW-MSA}(\text{BD}_2(\text{LN}(\mathbf{z}^{l-1}))), \tag{3}$$

$$\hat{\mathbf{z}}^l = \hat{\mathbf{z}}_o^l + \text{BU}_1(\hat{\mathbf{z}}_{d,1}^l) + \text{BU}_2(\hat{\mathbf{z}}_{d,2}^l) + \mathbf{z}^{l-1}, \tag{4}$$

$$\mathbf{z}^l = \text{MLP}(\text{LN}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l, \tag{5}$$

where BD and BU denote bilinear downsampling and bilinear upsampling, respectively. $\mathbf{z}^{l-1}$ is the input feature, $\hat{\mathbf{z}}_o^l$, $\hat{\mathbf{z}}_{d,1}^l$ and $\hat{\mathbf{z}}_{d,2}^l$ are the intermediate features containing local-to-global information, $\hat{\mathbf{z}}^l$ is the aggregation of these features, and $\mathbf{z}^l$ is the output feature.

3.2 Computational Complexity

Due to the downsampling operations in the LG-Attention blocks, the computational complexity is marginally increased compared with the Swin-Transformer baseline. Suppose each local window contains $M \times M$ tokens. The complexities of a SW-MSA module and our LG-Attention block (with parallel down-sampled ratios of 2 and 4) based on a feature map containing $h \times w$ tokens are[3]:

$$\Omega(\text{SW-MSA}) = 4hwC^2 + 2M^2hwC, \tag{6}$$

$$\Omega(\text{LG-Att}) = 5.25hwC^2 + 2.625M^2hwC. \tag{7}$$

Although two novel branches are added to the network, LG-Attention block only increases the computations about 0.3 times compared with SW-MSA. Further, the complexity is a linear function of hw, making it significantly lower than the complexity of the global self-attention, which is proportional to $(hw)^2$.

---

[3]We omit the computation of Softmax and window shift.

	Downsp. Rate (output size)	Layer Name	LG-T	LG-S
stage 1	4× (56×56)	Patch Embeding	C=4×4, D=96, LN	C=4×4, D=96, LN
		LG-Att	S=[4×, 16×, 32×] W=7×7, D=96, H=3 × 2	S=[4×, 16×, 32×] W=7×7, D=96, H=3 × 2
stage 2	8× (28×28)	Patch Merging	C=2×2, D=192, LN	C=2×2, D=192, LN
		LG-Att	S=[8×, 16×, 32×] W=7×7, D=192, H=6 × 2	S=[8×, 16×, 32×] W=7×7, D=192, H=6 × 2
stage 3	16× (14×14)	Patch Merging	C=2×2, D=384, LN	C=2×2, D=384, LN
		LG-Att	S=[16×, 32×] W=7×7, D=384, H=12 × 6	S=[16×, 32×] W=7×7, D=384, H=12 × 18
stage 4	32× (7×7)	Patch Merging	C=2×2, D=768, LN	C=2×2, D=768, LN
		SW-MSA	S=[32×,] W=7×7, D=768, H=24 × 2	S=[32×,] W=7×7, D=768, H=24 × 2

Table 1: Detailed architecture of LG-T and LG-S. C denotes the concatenation rate. D is the number of dimensions of the embedding. S are the scale rates of parallel attention paths in the LG-Attention block. W denotes the window size of SW-MSA. H is the number of heads in the multi-head attention.

Network	Image Size	Params (M)	FLOPs (G)	Top-1 Acc. (%)	Top-5 Acc. (%)
CNN Based Models					
MobileNetV1 [42]	224 × 224	4.2	0.6	70.6	–
MobileNetV2 (1.4) [43]	224 × 224	6.9	0.6	74.7	–
ResNet-18 [21]	224 × 224	11.7	1.8	69.8	89.1
ResNet-50 [21]	224 × 224	25.6	4.1	76.1	92.9
DenseNet-169 [31]	224 × 224	14.2	3.4	75.6	92.8
RegNet-4GF [44]	224 × 224	20.7	4.0	80.0	–
RegNet-16GF [44]	224 × 224	84	16	82.9	–
EfficientNet-B4 [45]	380 × 380	19.3	4.5	82.9	96.4
EfficientNet-B6 [45]	528 × 528	43	19	84.0	96.8
Transformers Based Models					
DeiT-T [46]	224 × 224	5.7	1.3	72.2	91.1
DeiT-S [46]	224 × 224	22.1	4.6	79.8	95.1
CrossViT-S [47]	224 × 224	26.7	5.6	81.0	–
T2T-ViT-14 [48]	224 × 224	22	5.2	81.5	–
TNT-S [49]	224 × 224	23.8	5.2	81.3	–
CoaT Mini [50]	224 × 224	10	6.8	80.8	–
CoaT-Lite Small [50]	224 × 224	20	4.0	81.9	–
PVT-Small [4]	224 × 224	24.5	3.8	79.8	–
CPVT-Small-GAP [51]	224 × 224	23	4.6	81.5	–
Swin-T [2]	224 × 224	29	4.5	81.3	–
LG-T (**ours**)	224 × 224	32.6	4.8	**82.1**	**95.8**
T2T-ViT-19 [48]	224 × 224	39.2	8.9	81.9	–
PVT-Medium [4]	224 × 224	44.2	6.7	81.2	–
Swin-S [2]	224 × 224	50	8.7	83.0	–
Swin-B [2]	224 × 224	88	15.4	**83.3**	–
LG-S (**ours**)	224 × 224	61.0	9.4	**83.3**	**96.2**

Table 2: Image classification results on ImageNet datasets.

3.3 Architecture Variants

We build two variants of our framework, denoted as LG-T and LG-S, where the only difference is the number of LG-Attention blocks in each stage. The detailed architecture of our framework is reported in Table 1. Except for the LG-Attention blocks, LG-Transformer basically employs the same settings as Swin-Transformer [2].

Stages with LG-Att	Params (M)	FLOPs (G)	Top-1 Acc. (%)
–	4.5	28.3	78.44
1	4.5	28.4	78.68
1∼2	4.6	29.0	78.89
1∼3	4.8	32.6	79.72

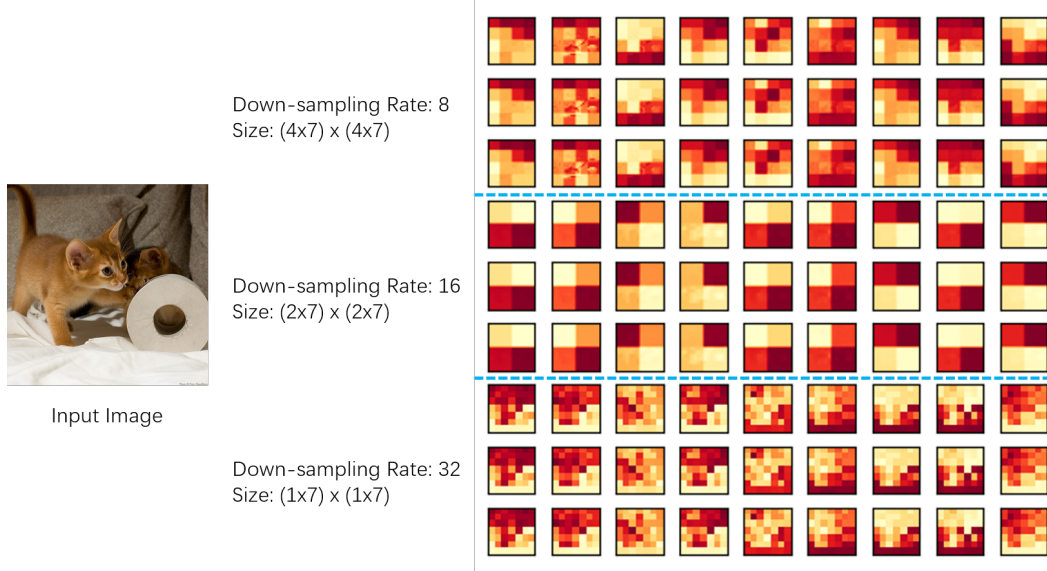Table 3: Ablation study by inserting LG Attention blocks into different stages.

Figure 3: Visualization of feature maps in stage 2 of LG-T. The LG-Attention block contains three paths, which generate feature maps of the input image with downsampling rates of 8 (local attention), 16 (mid-level attention), and 32 (global attention). Zoom in for better visualization.

4 Experimental Results

In this section, we report the experimental results of our LG-Transformer on two computer vision tasks, *i.e.*, image classification, and semantic segmentation. Comparisons with state-of-the-art models are given in each subsection. We also provide a thorough ablation study on model components and structure variations.

4.1 Image Classification

Experiment Settings. To achieve fair comparison with previous works, most of our experiment settings follows Swin-Transformer [2]. Experiments on image classification are conducted on the ImageNet-1K dataset [9] which contains 1,000 classes, and 1.28M and 50K images for the training set and validation set, respectively. Resolutions of input images are set to 224×224 for both training and evaluation. AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is use as optimizer for all of our models. We train LG-T and LG-S for 300 epochs in the comparison with state-of-the-art methods, and 100 epochs in the ablation studies. The initial learning rate is 0.001 with a cosine scheduler for learning rate decay, and a linear warm-up with 20 epochs is employ to stabilize training. All models are trained on a cluster with 8 V100 GPUs. We set Batch size to 128 on each GPU, and it takes near 4 days and 6 days to train a LG-T and a LG-S with 300 epochs, respectively. Data augmentation is used in the training phase, and we employ the same parameters as in [2]. We evaluate the test images based on a center cropped image without any test-time augmentation.

Comparative Results. Table 2 presents the comparative results on ImageNet classification. When comparing LG-Transformer with state-of-the-art Vision Transformers, we observe that our framework achieves competitive or better results under similar computational overheads. For example, LG-T

(a) Training loss of LG-T and Swin-T.

(b) Validation accuracy of LG-T and Swin-T.

(c) Training loss of LG-T with different numbers of LG-Att.

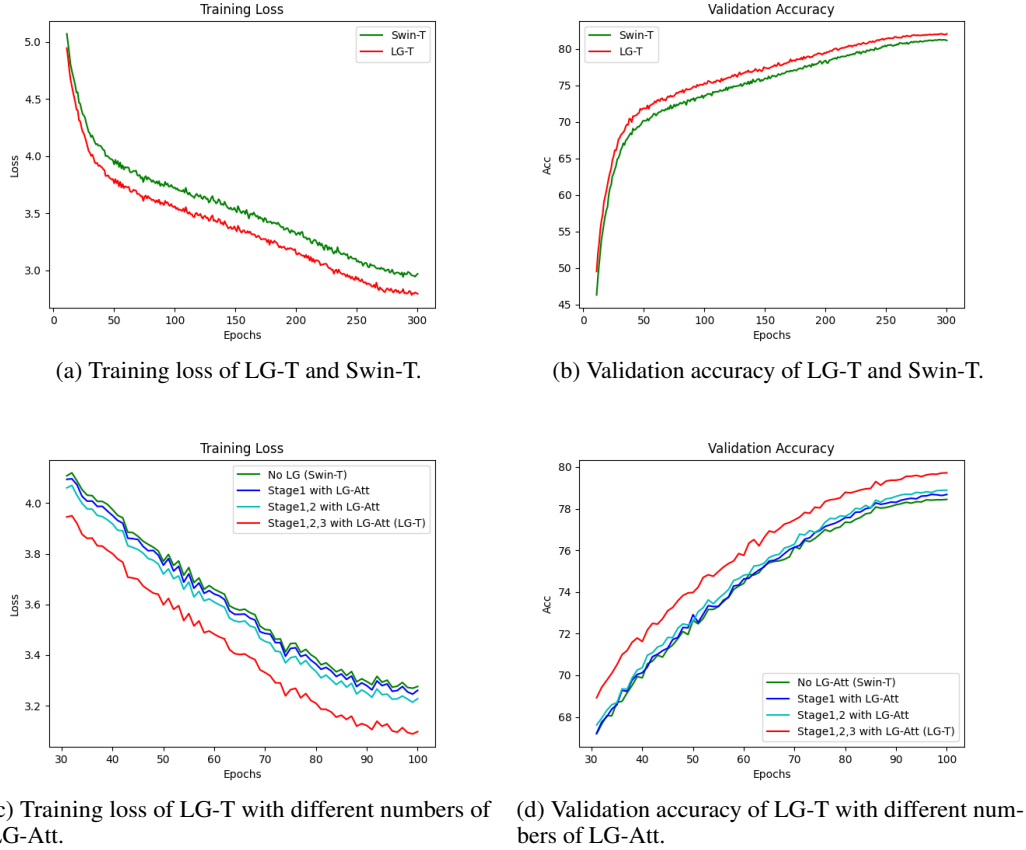(d) Validation accuracy of LG-T with different numbers of LG-Att.

Figure 4: Training losses and validation accuracies on ImageNet.

achieves 82.1% top-1 accuracy, outperforming Swin-T by 0.8%, PVT-Small by 2.3%, T2T-ViT-14 by 0.6%, and DeiT-S by 2.3%. Meanwhile, LG-S also outperforms its Transformer counterparts, *e.g.*, T2T-ViT-19, PVT-Medium, Swin-S. Notably, LG-S achieves comparable performance to Swin-B, with significantly fewer parameters and lower FLOPs.

Further, the performance of LG-Transformer is significantly better than the widely utilized CNN models, *i.e.*, MobileNet, ResNet, and DenseNet. Notably, the LG-Transformers achieve comparable performance to the state-of-the-art CNN models, *i.e.*, RegNet and EfficientNet.

Scale $\frac{1}{16}$	Scale $\frac{1}{32}$	Params (M)	FLOPs (G)	Top-1 Acc. (%)
		4.5	28.3	78.44
✓		4.6	28.7	78.74
	✓	4.7	32.2	79.60
✓	✓	4.8	32.6	79.72

Table 4: Ablation study for LG-Attention paths.

Ablation Studies. To evaluate the effectiveness of the proposed LG-Attention block, we carry out a series of ablation studies.[4] First, as shown in Table 3, by inserting the LG-Attention block into different stages, we can observe that the performance of LG-T improves continuously (from 78.4% to 79.7%) when LG-Attention is employed in more stages, while the parameters and FLOPs are only marginally increased (from 28.3 GFLOPs to 32.6 GFLOPs).

---

[4]We report results by training 100 epochs.

| ADE20K | | val | | |
Method	Backbone	mIoU	#param.	FLOPs
DANet [52]	ResNet-101	45.2	69M	1119G
ACNet [53]	ResNet-101	45.9	-	-
DNL [54]	ResNet-101	46.0	69M	1249G
OCRNet [36]	ResNet-101	45.3	56M	923G
OCRNet [36]	HRNet-w48	45.7	71M	664G
DeepLab.v3+ [55]	ResNet-101	44.1	63M	1021G
DeepLab.v3+ [55]	ResNeSt-101	46.9	66M	1051G
DeepLab.v3+ [55]	ResNeSt-200	48.4	88M	1381G
SETR [56]	T-Large	50.3	308M	-
UperNet [57]	ResNet-101	44.9	86M	1029G
UperNet	DeiT-S	44.0	52M	1099G
UperNet	Swin-T	46.1	60M	945G
UperNet	Swin-T*	44.5	60M	945G
UperNet	LG-T	45.3	64M	957G

Table 5: Results on the ADE20K val set. ∗ indicates the results from the public codes of Swin-Transformer [2]. Our models are implemented based on this code base.

Second, we compare different combinations of the attention paths with downsampling rates 16 and 32. As illustrated in Table 4, LG-Attention achieves the best performance when combining both types of attention.

Furthermore, Fig. 4 shows the training losses and validation accuracies of different models. From Fig. 4(a)(b), we find that GL-T converges faster, while consistently achieving a lower training loss and higher validation accuracy. When comparing the loss curves and validation accuracies of the model variants in Fig. 4(c)(d), we find that adding LG-Attention blocks to more stages constantly improves the validation performance and training convergence.

In summary, these results validate the importance of the proposed multi-path attention mechanism and demonstrate the necessity of individual attention paths.

Qualitative Results. To better illustrate how the local-to-global attention blocks work, we provide some qualitative results containing the feature maps from stage 2 of LG-T in Fig. 3. We show three levels of feature maps, with downsampling rates of 8 (local attention), 16 (mid-level attention), and 32 (global attention). The basic building block of the feature maps are 7×7 feature blocks. For example, the feature maps with a downsampling rate of 8 contains 4×4 feature blocks. Within each block, only the 7×7 local features interact with each other. Feature maps with a downsampling rate of 16 contain 4 blocks of 7 windows. Although the features in each block can represent more local information, each block is still processed individually. Attentions in the global feature maps a downsampling rate of 32 can cover the whole image, thus containing global semantic information. Notably, there is clear discontinuity across the borders of the local attention windows in the feature maps of downsampling rates of 8 and 16, showing the limitation of local attentions.

4.2 Semantic Segmentation

Experimental Settings. We conduct experiments of semantic segmentation on the ADE20K dataset [58], which includes 150 classes, and 20K, 2K images for training, and validation, respectively. We follow most of the settings as [2]. We resize and crop images to 512×512 resolutions for training, and resize images to 2048×512 for evaluation. Our model is pretrained on ImageNet-1K. AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used as optimizer, and weight decay is set to 0.01. Our model is trained for 160K iterations. The initial learning rate is 6×10^{-5} using a poly scheduler for learning rate decay. Batch size is set to 2 on each GPU. It takes near 1 day to train UperNet with LG-T, on a cluster with 8 V100 GPUs. We use the same training augmentation methods as [2], and no test-time augmentation is employed.

Results Analysis. We apply our model to semantic segmentation by combining LG-T with UperNet [57]. As shown in Table 5, our model achieves improvements compared with ResNet-101 (↑0.4%), DeiT-S (↑1.3%) and Swin-T* (↑0.8%), with a similar number of parameters and computa-

8

tional complexity. Although our model achieves inferior performance compared with state-of-the-art segmentation models, *i.e.*, DeepLab V3 + ResNeSt-200 and SETR, these models require a larger model size and more computation. From this perspective, our model achieves a good trade-off between performance and efficiency.

5   Conclusion

In this paper, we propose to incorporate local and global attention into a Vision Transformer. By building a multi-path structure in the hierarchical Vision Transformer framework, our model takes advantage of both local the feature learning mechanisms in CNNs and global feature learning mechanisms in Transformers. We conduct thorough studies on two computer vision tasks, and the results demonstrate that our framework yields improved performance with limited sacrifice in model parameters and computational overhead. As a multi-path framework, one of the limitations of our model is that the inference speed is lower than its single-path Transformer counterparts. We will try to improve the efficiency in our future work.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2, 3

[2] 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. *CoRR*, abs/2103.14030, 2021. 2, 3, 5, 6, 8

[3] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *CoRR*, abs/2103.15808, 2021. 2, 3

[4] 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. *CoRR*, abs/2102.12122, 2021. 2, 3, 5

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 2, 3

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 2, 3

[7] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017. 2, 3

[8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 2, 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 6

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 2

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 2

[15] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics, 2019. 2

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 3

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. *CoRR*, abs/2010.04159, 2020. 3

[18] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5790–5799. IEEE, 2020. 3

[19] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE, 2019. 3

[20] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *CoRR*, abs/2012.00364, 2020. 3

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3, 5

[22] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020. 3

[23] Irwan Bello, Barret Zoph, Quoc Le, Ashish Vaswani, and Jonathon Shlens. Attention augmented convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3285–3294. IEEE, 2019. 3

[24] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *CoRR*, abs/2104.05707, 2021. 3

[25] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *CoRR*, abs/2012.00759, 2020. 3

[26] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. 3

[27] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 886–893. IEEE Computer Society, 2005. 3

[28] Pedro F. Felzenszwalb, Ross B. Girshick, David A. McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 3

[29] Piotr Dollár, Ron Appel, Serge J. Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, 2014. 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[31] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 2261–2269, 2017. 3, 5

[32] Shreyas Saxena and Jakob Verbeek. Convolutional neural fabrics. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4053–4061, 2016. 3

[33] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 354–370. Springer, 2016. 3

[34] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 3

[35] Damien Fourure, Rémi Emonet, Élisa Fromont, Damien Muselet, Alain Trémeau, and Christian Wolf. Residual conv-deconv grid network for semantic segmentation. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017. 3

[36] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 173–190. Springer, 2020. 3, 8

[37] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017. 3

[38] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 122–137. Springer, 2018. 3

[39] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. NAS-FPN: learning scalable feature pyramid architecture for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7036–7045. Computer Vision Foundation / IEEE, 2019. 3

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2017, Honolulu, HI, USA, July 21-26, 2017, pages 6230–6239. IEEE Computer Society, 2017. 3

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 483–499. Springer, 2016. 3

[42] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017. 5

[43] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520. IEEE Computer Society, 2018. 5

[44] Ilija Radosavovic, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10425–10433. IEEE, 2020. 5

[45] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 5

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. 5

[47] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *CoRR*, abs/2103.14899, 2021. 5

[48] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *CoRR*, abs/2101.11986, 2021. 5

[49] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021. 5

[50] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. *CoRR*, abs/2104.06399, 2021. 5

[51] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers, 2021. 5

[52] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3146–3154. Computer Vision Foundation / IEEE, 2019. 8

[53] Jun Fu, Jing Liu, Yuhang Wang, Yong Li, Yongjun Bao, Jinhui Tang, and Hanqing Lu. Adaptive context network for scene parsing. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6747–6756. IEEE, 2019. 8

[54] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pages 191–207. Springer, 2020. 8

[55] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018,*

Proceedings, Part VII, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018. 8

[56] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, 2021. 8

[57] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 432–448. Springer, 2018. 8

[58] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 8