Multi-scale Convolutions Meet Group Attention for Dense Prediction Tasks

Xuan-Thuy Vo, Duy-Linh Nguyen, Adri Priadana, Jehwan Choi, and Kang-Hyun Jo

Department of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan (4460), South Korea
Email: xthuy@islab.ulsan.ac.kr;
{ndlinh301,priadana3202}@mail.ulsan.ac.kr; cjh1897@ulsan.ac.kr;
acejo@ulsan.ac.kr

Abstract. Self-attention can capture long-range dependencies from input sequences without inductive biases, resulting in quadratic complexity. When transferring Vision Transformers to dense prediction tasks, the models suffer huge computational costs. Recent methods have drawn sparse attention to approximate attention regions and injected convolution into self-attention layers. Motivated by this line of research, this paper introduces group attention that has linear complexity with input resolution while modeling global context features. Group attention shares information across channels, and convolution is spatial sharing. Both operations are complementary, and multi-scale convolution can capture multiple views of the input. Merging multi-scale convolution into group attention layers can help improve feature representation and modeling abilities. To verify the effectiveness of the proposed method, extensive experiments are conducted on benchmark datasets for various vision tasks. On ImageNet-1K image classification, the proposed method achieves 77.6% Top-1 accuracy at 0.7 GFLOPs, surpassing other methods under similar computational costs. When transferring pre-trained model on ImageNet-1K to dense prediction tasks, the proposed method attains consistent improvements across visual tasks.

Keywords: Convolution · Self-Attention · Vision Transformer

1 Introduction

Convolution extracts local features and has static weights while self-attention captures long-range dependencies from the long sequences without considering the order of tokens and has dynamic weights. In other words, Vision Transformers (ViTs) generalize better than Convolution Neural Networks (CNNs), producing state-of-the-art performances across language, vision, and multimodal tasks. Another line of research is to design hybrid networks that combine the best of convolution and self-attention. Hybrid networks [8, 31, 18] can achieve better feature representation and trade-off between accuracy and computational costs. In the model cost aspect, the self-attention operation creates quadratic complexity while convolution results in linear complexity with input resolution. Many methods attempt to improve the efficiency of ViTs based on sparse attention such as spatial reduction attention [27, 28, 32, 33], window attention [15, 6, 2, 31], and linear self-attention [17, 22].

The goal of hybrid networks is to supplement strong inductive biases of convolution to self-attention. For example, this kind of network replaces absolute positional encoding with depth-wise convolution [4, 18, 3] or models local-to-global features in a hierarchical manner [20, 16, 17]. In the conventional methods, convolution layers are used in earlier stages and self-attention layers are adopted at later stages. In this way, earlier stages extract local features from high-resolution input and later stages capture global features from low-resolution input. Hence, high computational cost and memory access of self-attention are mitigated while the model can achieve better hierarchical representation.

Sparse attention alleviates the quadratic complexity of self-attention by limiting the query's attendance. Concretely, each query attends to down-sampled key/value tokens [27, 28, 32, 33], local windows [15, 6, 31, 2, 19], and relevant regions [29]. Spatial reduction attention still keeps less important tokens while relevant tokens are ignored. This leads to suboptimal selection of attention areas. Window attention performs attention inside each window and requires window shifting operation to exchange information across windows. Although window attention achieves linear complexity with input sequences, the model has weak receptive fields and modeling abilities.

Inspired by sparse attention research, this paper proposes group attention that can model global contextual information at a low cost. The image tokens are grouped into pre-defined token numbers. This is achieved by attending learnable group queries to image tokens. The number of group tokens is smaller than image tokens (8 *vs.* 3136 in the first stage). Then, MLPMixer [24] is used to exchange information across grouped tokens, resulting in global information. The global features are propagated back to local image features via cross-attention where global features are queried by local image features. The local image features are captured by multi-scale convolutions, interacting with global features to achieve better feature modeling. The role of multi-scale convolutions is to extract local features and encode the order of image tokens. Convolution shares information across the spatial dimension while group attention shares information across the channel dimension. Using both convolution and group attention in one layer can help each other and enhance modeling abilities.

Extensive experiments are conducted and evaluated on ImageNet-1K [5] image classification, MS-COCO [14] object detection and instance segmentation, and ADE-20K [34] semantic segmentation. As a result, the proposed method achieves competitive performances across dense prediction tasks. Typically, the proposed method gets 77.% Top-1 accuracy at 0.7 GFLOPs that outperforms other methods with similar computational costs. On MS-COCO dataset, the proposed method surpasses the baseline RetinaNet with ResNet-50 by 1.2% while reducing GFLOPs by 34.8%. The performances across instance segmentation and semantic segmentation tasks attain consistent improvements. It verifies the effectiveness and generalization of the proposed method.

2 Related Works

2.1 Vision Transformers

In 2021s, ViT [7] successfully applies Transformer encoder [26] for the image classification task, achieving promising performance compared to CNNs. Due to the high resolution of the input image, ViT separates the images into a sequence of 16×16 patches and considers a patch as a token. Self-attention blocks are performed on all tokens and capture long-range dependencies. With this scheme, ViT has high flexibility in extracting image features. When transferring ViT to dense prediction tasks, the model creates high computational costs. To deal with this issue, PVT [27] designs hierarchical ViT that performs attention on multi-scale tokens. For fine-grained tokens, spatial reduction attention is used to reduce the computational cost of self-attention. ResTv2 [33] recovers information loss in spatial reduction attention via the upsampling layer. QuadTree [23] constructs a token pyramid and allows each query to attend to relevant regions via Topk selection. Swin Transformer [15] limits attention inside each square window and window shifting is proposed to communicate information across non-overlapped windows. CSWin [6] expands square windows to cross-shaped windows and performs attention on these expanded windows. Slide-Transformer [19] proposed window sliding that efficiently connects information across windows via depthwise convolution.

2.2 Hybrid Networks

Hybrid networks combine strengths from convolution and self-attention operation to achieve the balance between accuracy and computational costs. PVTv2 [28] incorporates depthwise convolution between two fully connected layers of the MLP module. This incorporation effectively extracts local features on high-dimension input and also works as adaptive positional encoding. CPVT [4] proposes conditional position embedding using depthwise convolution. CMT [8] adopts 3×3 depthwise convolution that downsamples key/value features and encodes local information. EdgeViT [18] captures local and global features sequentially, achieving state-of-the-art performances on mobile devices. LVT [20] uses convolution layers at stages 1-2, and self-attention layers at stages 3-4. MobileViT [16] improves MobileNetv2 [21] by inserting self-attention into the latter stages of this network. MobileViTv2 [17] introduces separable self-attention that has linear complexity with image resolution and has inserted positions similar to MobileViT. MixFormer [2] combines depthwise convolution and window self-attention in a parallel way. Furthermore, MixFormer also captures bidirectional interactions across channel and spatial dimensions based on conventional attention. EMO [31] unifies the window self-attention layer and Convolution MLP

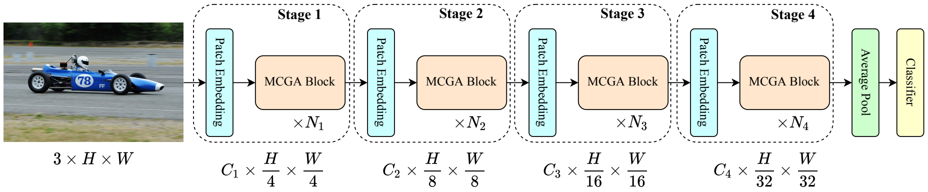
Stage 1
Stage 2
Stage 3
Stage 4
Patch Embedding
MCGA Block
$\times N_1$
$\times N_2$
$\times N_3$
$\times N_4$
Average Pool
Classifier
$3 \times H \times W$
$C_1 \times \frac{H}{4} \times \frac{W}{4}$
$C_2 \times \frac{H}{8} \times \frac{W}{8}$
$C_3 \times \frac{H}{16} \times \frac{W}{16}$
$C_4 \times \frac{H}{32} \times \frac{W}{32}$

Fig. 1: Overview architecture of the proposed method. H, W are the height and width of the input image. C_1, C_2, C_3, C_4 are the number of channels across four stages and N_1, N_2, N_3, N_4 are the number of stacked blocks across four stages. MCGA block indicates Multi-scale Convolution meets Group Attention.

layer into a meta mobile block. Different from MixFormer, EMO implements self-attention and convolution in a sequential manner. SwiftFormer [22] performs local-to-global features in sequential implementation and global features are captured by efficient additive attention.

3 Methodology

The overview architecture of the proposed method is shown in Figure 1. Following [27, 28, 15], the hierarchical network is obtained, including four stages. Spatial dimensions are progressively down-sampled across four stages with stride {4, 8, 16, 32}. Channel dimensions are doubled at every stage and the model becomes deeper and wider. Each stage consists of one Patch Embedding and stacked MCGA blocks. Patch Embedding splits the input image into a sequence of patches, achieved by depthwise convolution with stride p (p is patch size). Then, a sequence of patches (tokens) is fed into MCGA blocks to model short-range and long-range dependencies across tokens. Finally, average pooling and classifier implemented by linear layers are used to produce digit scores and final output embedding. In the following, the MCGA block and the main part of MCGA are analyzed in detail.

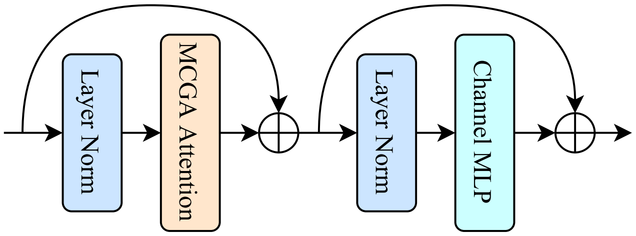

Fig. 2: The detailed structure of the proposed MCGA block.

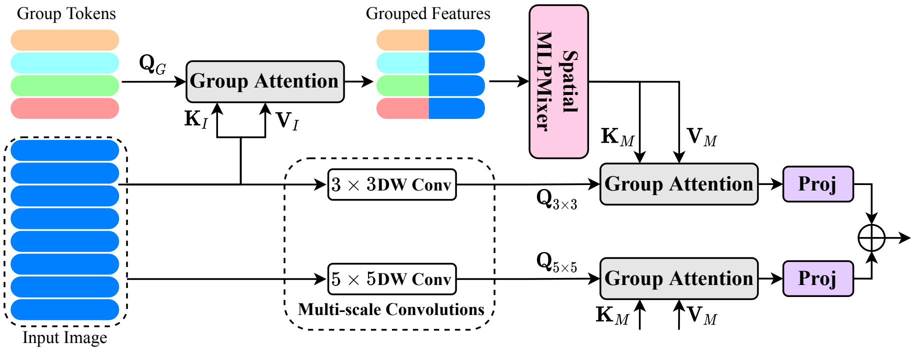

Fig. 3: Detailed structure of the proposed MCGA attention. DW Conv denotes depthwise convolution. Proj is 1×1 linear projection.

3.1 MCGA Block

Figure 2 shows the detailed structure of the proposed MCGA block. Similar to [25, 27, 15], the MCGA block consists of Layer Normalization, MCGA attention (spatial attention), Layer Normalization, Channel MLP (Multi-Layer Perceptron), and two shortcut connections between spatial and channel layers. Channel MLP includes two fully connected (FC) layers expanding channel dimensions and one GELU activation function inserted between two FC layers. The role of the activation function is to learn non-linear mapping among tokens due to the linearity of spatial and channel layers. A pair of layer normalization and spatial attention is used to model similarity across spatial locations, and layer normalization followed by channels MLP is to mix information across the channel dimension. Both spatial and channel operations are complementary, resulting in better token relationships.

3.2 MCGA Attention

Figure 3 illustrates the detailed structure of the proposed MCGA attention. The main goal of the MCGA is to capture both local and global features at low computational costs. For local features, multi-scale convolutions are utilized to enlarge receptive fields, learn multiple scales of objects, and also encode the order of image tokens. To extract global features, group attention is proposed to group the image tokens into a small number of predefined tokens (8 tokens). Self-attention in ViT [7] has quadratic complexity with the image resolution while the proposed group attention only creates linear computational costs.

Multi-scale Convolutions. This paper applies depthwise convolutions that capture the locality and translation-invariant of the image feature. This is achieved by 3×3 and 5×5 depthwise convolutions with a small increase in parameters and

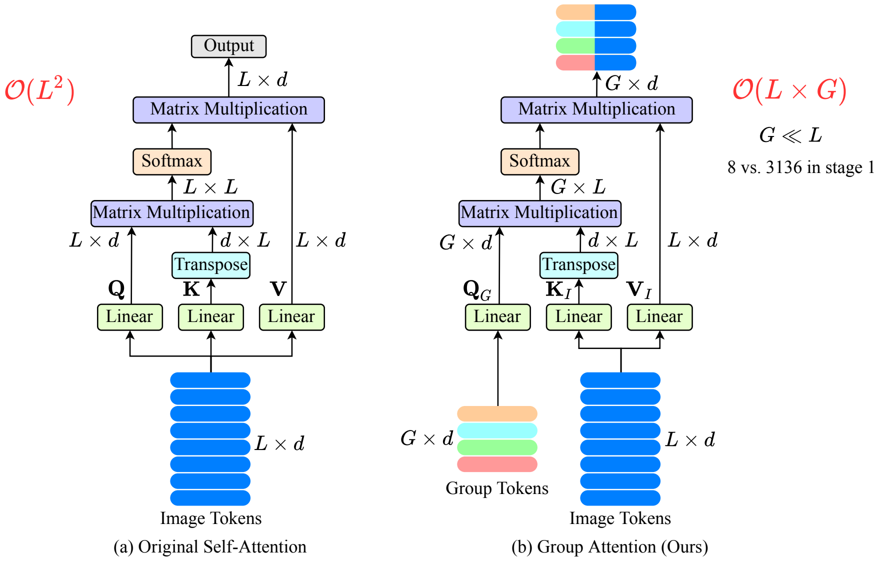

Fig. 4: Comparision between vanilla self-attention [26, 25] and the proposed group attention. $\mathcal{O}()$ denotes the model complexity. The proposed group attention achieves better efficiency than original self-attention while still capturing global information from the input tokens. G, L is the number of group tokens and image tokens, respectively.

computational costs. The output of the two branches is denoted as $\mathbf{Q}_{3\times3}$ and $\mathbf{Q}_{5\times5}$ for the input query of the group attention, as follows:

$$\mathbf{Q}_{3\times3} = \text{DWConv}_{3\times3}(\mathbf{X}^r), \tag{1}$$

$$\mathbf{Q}_{5\times5} = \text{DWConv}_{5\times5}(\mathbf{X}^r), \tag{2}$$

where $\mathbf{X}^r \in \mathbb{R}^{H\times W\times d}$ is the input tokens after reshaping the input token $\mathbf{X} \in \mathbb{R}^{L\times d}$. Token length is denoted by $L = H \times W$ (H, W is the height and width of the input feature) and channel d. DWConv$_{3\times3}$ indicates 3×3 depthwise convolution. DWConv$_{5\times5}$ denotes 5×5 depthwise convolution.

Group Attention. Figure 4 shows the detailed structure of the group attention. Compared to original sef-attention [26, 25], the proposed group attention results in much lower computational and memory access. The image tokens are queried by group tokens to produce grouped features. It means that each group query attends to all image tokens. Therefore, the proposed group attention still models long-range dependencies from the image tokens.

Given the image token $\mathbf{X} \in \mathbb{R}^{L\times d}$, two linear projections are used to map $\mathbf{X}$ to $\mathbf{K}, \mathbf{V}$ matrices. The group tokens $\mathbf{T} \in \mathbb{R}^{G\times d}$ are defined as learnable parameters optimized with the network's parameters. The tokens $\mathbf{T}$ are projected to query $\mathbf{Q}_G$ via linear transformation. Attention are performed between $\mathbf{Q_G}$ and $\mathbf{K}_I$ to create attention matrix with dimension $G \times L$. Intuitively, each row of the attention matrix reveals the similarity between each query and all image tokens. It results in long-range dependencies. Then, softmax() are applied on each row of the attention matrix, generating attention weights. The input and attention weights are aggregated via matrix multiplication between value $\mathbf{V}$ and the attention weights. Technically, these processes are summarized as follows,

$$\mathrm{GA}(\mathbf{T}, \mathbf{X}) = \mathrm{softmax}(\frac{\mathbf{Q}_\mathrm{G}\mathbf{K}_\mathrm{I}^\mathrm{T}}{\sqrt{\mathrm{d}}})\,\mathbf{V}_I, \tag{3}$$

where $\mathrm{GA}(\mathbf{T}, \mathbf{X})$ is group attention with the two inputs $\mathbf{T}$ and $\mathbf{X}$. $\mathbf{Q}_G = \mathbf{T}\mathbf{W}_Q^G, \mathbf{K}_I = \mathbf{X}\mathbf{W}_K^I, \mathbf{V}_I = \mathbf{X}\mathbf{W}_V^I$ are group query, image key, and image value. $\mathbf{W}_Q^G, \mathbf{W}_K^I, \mathbf{W}_V^I \in \mathbb{R}^{d \times d}$ are linear projections from group tokens and image tokens. d is the number of channels. The output of group attention is the grouped features with dimension $G \times d$.

Feature Propagation. After producing the grouped features, Spatial MLPMixer [24] is applied to mix the grouped features spatially. This means that each grouped feature is fully connected with all grouped features. Hence, Spatial MLPMixer exchanges global information across grouped features. In this paper, as the number of group tokens is fixed, using Spatial MLPMixer is suitable for mixing information. In conventional methods [24], when training the input with different sizes, Spatial MLPMixer requires further interpolation layers to up/down-sample the input token.

After mixing grouped features, two group attentions are used to propagate global information to local features modeled by multi-scale convolutions. For the first group attention branch, the global features are queried by local features. In other words, the global features are set as a pair of key $\mathbf{K}_M$ and value $\mathbf{V}_M$, and the output of 3×3 depthwise convolution is set as query $\mathbf{Q}_{3\times3}$. Intuitively, local features are updated by attending each local query to global features. This process is similar for 5×5 depthwise convolution branch. The two outputs from two branches are fused together by simple linear projections and summation.

Model Configuration. After obtaining the MCGA block, the number of channels and stacked blocks across stages are configured. Typically, the number of channels across four stages is set to {32, 64, 128, 256} and the number of stacked blocks is configured to {2, 2, 6, 6}. As the channel numbers are increasing, the number of heads is {2, 2, 4, 8} for each group attention. Similar to [27, 28, 15], the expansion ratio in MLP is set to 4 and kept unchanged across stages.

4 Experiments

Settings. The proposed method is trained and evaluated on ImageNet-1K for the image classification task. After pretraining on ImageNet, the model is trans-

Table 1: Performance on ImageNet-1K image classification

Method	Input Size	#params (M)	GFLOPs	Top-1 Acc (%)
PVTv2-B0 [28]	224	3.7	0.6	70.5
QuadTree-B-b0 [23]	224	3.5	0.7	72.0
DeiT-T [25]	224	6.0	1.3	72.2
EdgeViT-XXS [18]	256	4.1	0.6	74.4
LVT [20]	224	3.4	0.9	74.8
PVT-T [27]	224	13.0	1.8	75.1
VAN-B0 [9]	224	4.1	0.9	75.4
ResT-Lite [32]	224	10.5	1.4	77.2
Ours	**224**	**10.3**	**0.7**	**77.6**

ferred to dense prediction tasks, object detection, instance segmentation, and semantic segmentation.

For the image classification task, the proposed method is trained for 300 epochs with a batch size of 1024, following [27, 15, 6, 18]. The optimizer is AdamW with a learning rate of 10^{-3} and a weight decay of 0.05. The input image is resized to 224×224. Standard data augmentations such as RandAug, Cutmix, Mixup, and label smoothing are adopted, defined by [25, 27, 15].

For the object detection task, the proposed backbone is integrated with RetinaNet [13] using the codebase [1]. All the hyperparameters are similar to PVT [27], Swin [15]. Specifically, the model is trained on the MS-COCO [14] dataset for 12 epochs with a batch size of 16. The optimizer AdamW is used with a learning rate of 10^{-4}. The input image is resized to 1333×800.

For the instance segmentation task, the baseline model Mask R-CNN [10] is used with the proposed backbone. Similar to the detection task, the MS-COCO dataset is utilized to train and evaluate the integrated model.

For the semantic segmentation task, the backbone ResNet-50 [11] in segmentor Semantic FPN [12] is replaced with the proposed method. The integrated model is trained and evaluated on the ADE-20K [34] dataset. The model is trained for 80K iterations with a batch size of 16. AdamW is used as an optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-4}. The input image is resized to 512×512.

Image Classification Results. Table 1 reports the performance on Image-1K image classification. The proposed method achieves 77.6% Top-1 accuracy with 0.7 GFLOPs, surpassing other methods under similar computational costs. For example, the proposed method outperforms the baseline PVT-T [27] by 2.5% with 38% GFLOPs, hybrid model PVTv2-B0 [28] by 7.1% with similar GFLOPs, EdgeViT-XXS [18] by 3.2% with the similar computational cost, and the recent method ResT-Lite by 0.4% with half of GFLOPs. It verifies the effectiveness of the proposed method.

Table 2: Performance on MS-COCO object detection using the RetinaNet

Method	#params (M)	GFLOPs	AP	AP^{50}	AP^{75}
ResNet-18 [11]	21	189	31.8	49.6	33.6
PoolFormer-S12 [30]	22	207	36.2	56.2	38.2
ResNet-50 [11]	38	250	36.3	55.3	38.6
PVT-T [27]	23	183	36.7	56.9	38.9
Ours	**18**	**163**	**37.5**	**58.3**	**39.4**

Table 3: MS-COCO instance segmentation results using Mask R-CNN [10]

Methods	#params (M)	GFLOPs	AP^{box}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
ResNet-18 [11]	31	207	34.0	31.2	51.0	32.7
PVT-T [27]	33	208	36.7	35.1	56.7	37.3
ResNet-50 [11]	44	260	38.0	34.4	55.1	36.7
Ours	**28**	**181**	**38.1**	**36.0**	**57.8**	**38.5**

Object Detection Results. Table 2 shows performance on MS-COCO validation set using detector RetinaNet [13]. As a result, the proposed method outperforms the baseline RetinaNet with backbone ResNet-50 [11] by 1.2% AP while saving 34.8% GFLOPs, PVT-T [27] by 0.8% AP with smaller GFLOPs. It clarifies the versatility of the proposed method.

Instance Segmentation Results. Table 3 reports performance on the MS-COCO [14] validation set using the model Mask R-CNN [10]. The proposed method achieves consistent improvements similar to the object detection task. For example, the proposed method gets 38.1% AP^{box}, 36.0% AP^{mask} with 181 GFLOPs better than ResNet-18 by 4.1% AP^{box}, 1.4% AP^{box}, 6.8% AP^{mask} with smaller GFLOPs, PVT-T by 0.9% AP^{mask}, and the baseline ResNet-50 by 1.6% AP^{mask} while saving 30% GFLOPs.

Table 4: ADE-20K semantic segmentation using Semantic FPN [12]

Method	#params (M)	GFLOPs	mIoU
ResNet-18 [11]	15.5	32.2	32.9
PVT-T [27]	17.0	33.2	35.7
ResNet-50 [11]	28.5	45.6	36.7
Ours	**12.8**	**24.4**	**37.0**

Sematic Segmentation Results. Table 4 shows the performance of the proposed method on ADE-20K [34] validation set using the baseline segmentor Semantic FPN [12] with original backbone ResNet-50 [11]. The proposed method achieves 37.0% mIoU with 24.4 GFLOPs that surpasses ResNet-18 by 4.1%

mIoU while saving 24.2% GFLOPs, the PVT-T [27] by 1.3% mIoU with 73.5% GFLOPs, and the baseline ResNet-50 [11] by 0.3% mIoU while saving 46.5% GFLOPs.

The results on image classification, object detection, instance segmentation, and semantic segmentation verify the efficiency, effectiveness, and generalization of the proposed methods.

5 Conclusion

This paper combines the strengths of multi-scale convolutions and group attention (MCGA). The MCGA attention can capture both local and global features at low computational costs. This is achieved by the proposed group attention. Attending each group query to all spatial image tokens can still capture global information while reducing a lot of computational costs. Extensive experiments are conducted on benchmark datasets for image classification and dense prediction tasks. As a result, the proposed method achieves better performances on the ImageNet-1K dataset compared to previous methods. On dense prediction tasks, the proposed method produces consistent improvements across detection, instance segmentation, and semantic segmentation. In the future, the proposed method will be scaled to bigger models and applied to other tasks such as keypoint detection, and video classification.

Acknowledgement

This result was supported by “Region Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).

References

1. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
2. Chen, Q., Wu, Q., Wang, J., Hu, Q., Hu, T., Ding, E., Cheng, J., Wang, J.: Mixformer: Mixing features across windows and dimensions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5259 (2022)
3. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems **34**, 9355–9366 (2021)
4. Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C.: Conditional positional encodings for vision transformers. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=3KWnuT-R1bh

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
8. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12175–12185 (2022)
9. Guo, M.H., Lu, C.Z., Liu, Z.N., Cheng, M.M., Hu, S.M.: Visual attention network. Computational Visual Media **9**(4), 733–752 (2023)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6399–6408 (2019)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
15. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
16. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=vh-0sUt8HlG
17. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=tBl4yBEjKi
18. Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., Martinez, B.: Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In: European Conference on Computer Vision. pp. 294–311. Springer (2022)
19. Pan, X., Ye, T., Xia, Z., Song, S., Huang, G.: Slide-transformer: Hierarchical vision transformer with local self-attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2082–2091 (2023)

20. Pan, Z., Zhuang, B., He, H., Liu, J., Cai, J.: Less is more: Pay less attention in vision transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2035–2043 (2022)
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
22. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
23. Tang, S., Zhang, J., Zhu, S., Tan, P.: Quadtree attention for vision transformers. ICLR (2022)
24. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021)
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
27. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
29. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022)
30. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10819–10829 (2022)
31. Zhang, J., Li, X., Li, J., Liu, L., Xue, Z., Zhang, B., Jiang, Z., Huang, T., Wang, Y., Wang, C.: Rethinking mobile block for efficient attention-based models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1389–1400 (2023)
32. Zhang, Q., Yang, Y.B.: Rest: An efficient transformer for visual recognition. Advances in neural information processing systems **34**, 15475–15485 (2021)
33. Zhang, Q., Yang, Y.B.: Rest v2: simpler, faster and stronger. Advances in Neural Information Processing Systems **35**, 36440–36452 (2022)
34. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)