Reweighting Foveal Visual Representations

Xuan-Thuy Vo, Duy-Linh Nguyen, Adri Priadana and Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan (44610), South Korea
Email: xthuy@islab.ulsan.ac.kr; {ndlinh301, priadana3202}@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

***Abstract*—Biological foveal vision consists of multiple contour regions, determined by the varying distances from the center of the gaze. Adopting foveal vision in deep neural networks can have the ability to capture various visual features in different regions. Long-range dependencies from the gaze are modeled by global operations (global self-attention and state-space model) and short-range dependencies are perceived by local operations (local self-attention and convolution). Existing works in visual backbones have improved the performance by modeling local and global features of the input images. However, fully perceiving foveal vision has not been well explored, which is crucial for modeling visual features. To address the above issue, this paper proposes a Reweighting Foveal (RF) mechanism for a visual representation to extract various features at different regions varied by the distance from the center of the query's position. Far regions from each query position are modeled by pooling self-attention on coarse input and nearest regions are perceived by local convolution on fine-grained input. The importance of each region to the model features is also emphasized by a reweighting module based on softmax attention to let the model learn to perceive the relationship among foveal regions. Based on this design, the RF Transformers are introduced by stacking RF blocks across stages. Extensive experiments are validated on image classification, object detection, and semantic segmentation. On image classification, RF-1 with 8.5M parameters and 0.7 GFLOPs achieves 78.2% Top-1 accuracy that surpasses recent ConvNets and Vision Transformer methods. When transferring trained RF Transformers to other tasks, the proposed methods obtain competitive performances compared to recent backbones while getting better efficiency.**

***Index Terms*—Foveal Vision, Image classification, Vision Transformers, Vision Tasks**

I. Introduction

Recent advanced ConvNets [1]–[4] and Vision Transformers [5]–[8] have attempted to improve performance by extracting both local and global features from the input images. Vision Transformers become dominant networks in solving vision and multimodal tasks as the self-attention layer directly captures long-range dependencies from the input sequences without inductive biases. This makes the model unify different input sequences and also stack more layers to achieve deeper and wider networks. Inspired by this line of research, advanced ConvNets expand the receptive fields of the model by enlarging kernel sizes [3], [4] and convolutional modulation [9], [10].

The main bottleneck of the Transformer is that self-attention has quadratic computational costs with the image length. Transferring the vision Transformers to dense prediction tasks results in extremely large computational costs due to the high-resolution inputs of these tasks. One possible solution is to use sparse attention where each query attends to smaller image regions. PVT [6] achieves this goal by reducing the size of key and value features. Swin Transformer [7] proposes window self-attention that has linear complexity with image resolutions. However, these methods still capture global and local features separately, and attentions are performed on irrelevant regions. DAT [12] introduces deformable attention that shifts key and value features to important regions. Another line of this research is to design hybrid methods [13]–[16] that combine the locality of convolution and long-range dependencies of self-attention into hierarchical networks. Although hybrid methods achieve better trade-offs between accuracy and computation costs, the interaction between local and global operations is not well explored in the literature. Modeling this interaction is crucial for biological foveal systems.

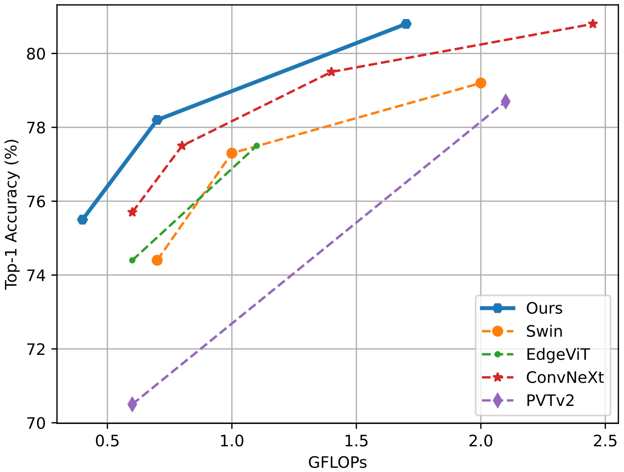
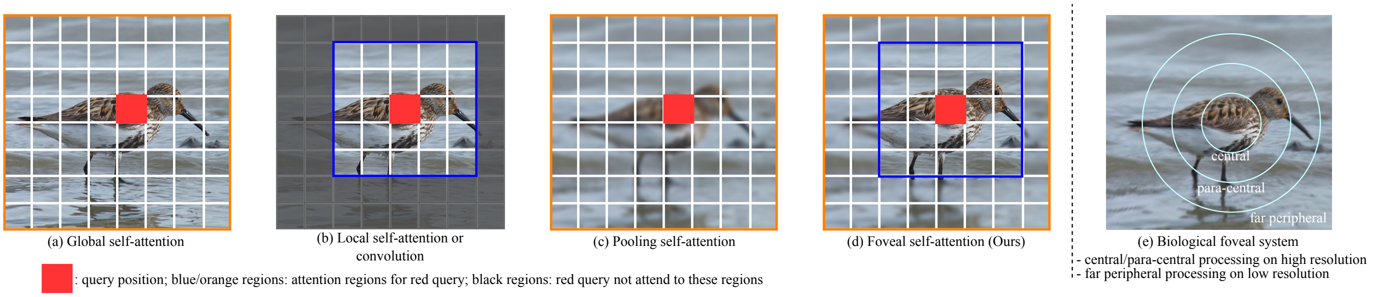


Fig. 1: Comparison between the proposed method and recent ConvNets/Vision Transformers. The results are reported on ImageNet-1K [11] validation set

In global self-attention (Figure 2 (a)), all tokens closer or far to the query's location are treated evenly. This results in global features while local features are alleviated. Otherwise, local self-attention in Figure 2 (b) captures detailed information inside each window and the model requires further operations to exchange information across windows. Figure 2 (c) illustrates pooling self-attention where each query attends to coarse-grained features to obtain global context. Similar to global self-attention, pooling self-attention lacks fine-grained information from the query's position to its nearest regions. Unifying local and pooling self-attention into one layer can


(a) Global self-attention

(b) Local self-attention or convolution

(c) Pooling self-attention

(d) Foveal self-attention (Ours)

(e) Biological foveal system

central
para-central
far peripheral
- central/para-central processing on high resolution
- far peripheral processing on low resolution
🟥 : query position; blue/orange regions: attention regions for red query; black regions: red query not attend to these regions

Fig. 2: Comparison between self-attention: (a) global self-attention [5] (each query attends to all spatial locations); (b) local self-attention [7] or convolution (each query attends to window regions); (c) pooling self-attention [6] (each query attends to all pooled locations); (d) foveal self-attention (ours) (each query attends to both fine-grained regions for modeling local features and coarse-grained regions for modeling global context). Figure (e) shows the biological foveal system that is partitioned into multiple contour regions: central/para-central, and far peripheral.

get better visual representations.

Figure 2 (e) illustrates a biological foveal system, consisting of multiple contour regions: central/para-central and far peripheral regions. Fine-grained detail information is processed at the center regions, high-resolution regions. Coarse-grained information is processed at the far peripheral regions, low-resolution regions. Inspired by foveal processing, this paper proposes foveal self-attention with two attentions shown in Figure 2 (d). The first attention is to capture detailed information from the fine-grained input where each query attends to its nearest regions. The second attention is to model far peripheral regions, extracting global context information from coarse-grained input. In second attention, each query attends to coarse-grained regions (down-sampled input features). Both attentions can fully extract both local and global perceptions at low computational costs. Furthermore, the features of two attentions are aggregated via the reweighting module. The key intuition of this design is to highlight the importance of each captured feature.

Extensive experiments are conducted on image classification, object detection, and semantic segmentation tasks to validate the effectiveness of the proposed RF Transformers. Figure 1 shows the comparison among methods. As a result, RF Transformers achieve better trade-offs between Top-1 accuracy and computational costs (GFLOPs). For other visual tasks, RF Transformers attain consistent improvements in both efficiency and effectiveness.

II. Related Works

A. Vision Transformers

Transformer [17] was originally designed for language research, improving parallel computing of recurrent layers. The main advance of the Transformer is that self-attention layers can capture long-range dependencies from the sequence length and result in better next-token prediction. This motivates researchers to apply Transformers for vision tasks. The pioneering work for object detection is DETR [18] which adopts Transformer encoders and decoders to model the relation between image features and object queries. With this successful adaptation, DETR achieves competitive results with anchor-based detectors [19]–[21] while having high flexibility in capturing the object's locations. In visual extraction, ViT [5] explores Transformer encoders for image classification tasks and attains better performance and scalability compared to ConvNets [22]. From this milestone, many methods are introduced to significantly improve the performance of ViT by reducing model complexity [6], [7] and integrating inductive biases into self-attention layers [7], [13], [15], [16], [23].

PVT [6] builds hierarchical vision Transformers, and pools key and value features to mitigate the quadratic complexity of original self-attention layers. PVTv2 [15] improves PVT networks by inserting convolution to the MLP layer and attains great performances on both classification and dense prediction tasks. Swin Transformer [7] partitions images into window regions and applies self-attention layers for capturing the relationship between tokens inside each window. Even though Swin Transformer computes attention with linear complexity, the modeling ability and receptive fields are weak and the network requires additional designs to expand them. DAT [12] deforms key/value features to relevant regions based on pixel locations and learnable offsets. PerViT [23] augments relative inductive biases by introducing special designs of kernel weights.

B. Hybrid Networks

The goal of hybrid networks is to take the strengths of convolution and self-attention into account. Convolution has locality and strong inductive biases while self-attention results in global features without inductive biases. Combining convolution and self-attention layers can model both local and global features, and obtain better efficiency. EdgeViT [13] follows this line of research and proposes sequential local-to-global layers based on convolution and pooling self-attention. FAT [16] captures bidirectional interaction between local and global features based on gated aggregation. MixFormer [14] models bidirectional interaction between convolution and window attention in a parallel way to improve information exchange between channel and spatial directions. EMO [24] builds a

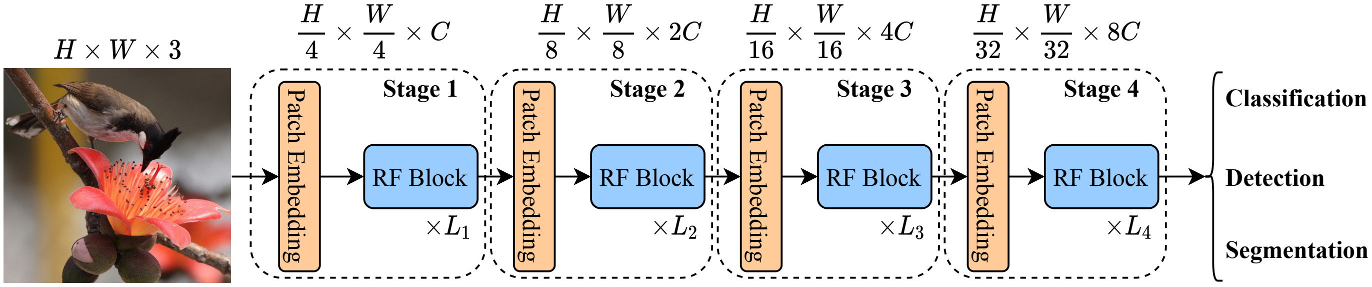

$H \times W \times 3$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Stage 1
Stage 2
Stage 3
Stage 4
Patch Embedding
RF Block
$\times L_1$
$\times L_2$
$\times L_3$
$\times L_4$
Classification
Detection
Segmentation

Fig. 3: Overall architecture of the proposed RF Vision Transformer for visual tasks: image classification, object detection, and segmentation. Following [6], [7], [15], the RF network is the hierarchical backbone, consisting of four stages. The spatial dimension is progressively down-sampled with ratio {1/4, 1/8, 1/16, 1/32} across stages through Patch Embedding layers. Along with the reduction of spatial sizes, channel dimension is progressively increased across stages to make the network deeper and wider. RF Block is the proposed Reweighting Foveal Block. $\{L_1, L_2, L_3, L_4\}$ is the number of stacked RF blocks across four stages. H, W, C are the height, width, and channel dimensions of the feature map.

sequential window self-attention and convolution layer for efficient networks.

Capturing local and global features in deep neural networks [8], [13], [14], [16], [24] is similar to biological foveal systems. However, biological foveal systems contain multiple regions based on the distance from the center of the gaze to the token's location. Different regions are treated unevenly, e.g., the nearest regions to their query location are modeled on fine-grained attention, and the far regions to their query location are captured on coarse-grained attention. Existing works have not fully explored the line of research and also the relationship between multiple regions.

III. The Proposed Method

An overview of the proposed method is shown in Figure 3. Similar to the existing methods [6], [7], [15], RF Transformers extract features in a hierarchical manner. Earlier stages tend to capture local features and later stages extract global features. The proposed network is divided into four stages and each stage consists of one Patch Embedding layer and stacked RF blocks. Patch Embedding separates the input feature map into a sequence of patches via convolution with kernel size p and stride p where p is patch size. Spatial information of each patch is embedded into channel direction with dimension $\frac{H}{p} \times \frac{W}{p} \times Cp^2$. Following the meta block [5], [25], the RF block includes two main layers: RF attention (spatial mixing) and MLP (channel mixing) as follows:

Before processing spatial and channel mixing, layer normalization is used to normalize the input features and stabilize training. Two residual connections for two main layers are applied to avoid vanishing and exploding gradient and stack the layers to be deeper. MLP is a multi-layer perceptron, consisting of two fully connected layers and one GELU() non-linear activation function inserted between two fully connected layers. The proposed RF Attention is discussed in the next section.

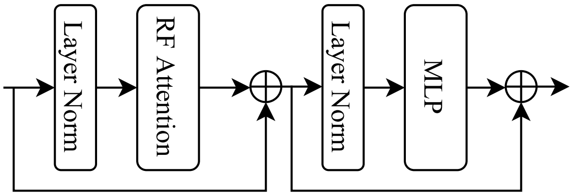


Fig. 4: The detailed structure of the RF (Reweighting Foveal) block

A. RF Attention

To model the biological foveal system, RF attention is proposed to perform attention on central/para-central and far peripheral regions. These two regions are treated unevenly based on the distance from the determined regions to the query's position. Similar to the biological foveal system, the center of the gaze is viewed as the query's position in the input feature map. For the nearest tokens to their query position, each query interacts with fine-grained regions to model local features. For the far tokens to their query position, each query attends to coarse-grained regions to capture global information. Both features are aggregated via a reweighting module that controls the importance of fine-grained and coarse-grained features based on softmax attention. Figure 5 shows the illustration of RF attention.

Given the input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and query location $\mathbf{q}_{i,j}$, pooled tokens and nearest tokens to the $\mathbf{q}_{i,j}$ are obtained by pooling operation and convolution window.

1) Pooled Tokens: The input feature map is pooled into a set of keys $\mathbf{k} \in \mathbb{R}^{N_p \times C}$ and values $\mathbf{v} \in \mathbb{R}^{\mathbf{N_p} \times \mathbf{C}}$ (N_p is the number of pooled tokens) as follows:

$$\mathbf{k} = \mathrm{AvgPool}(\mathbf{X})\mathbf{W}_k, \quad (1)$$

$$\mathbf{v} = \mathrm{AvgPool}(\mathbf{X})\mathbf{W}_v, \quad (2)$$

where $\mathrm{AvgPool}(.)$ is Average Pooling that down-samples the input feature map to coarse-grained features. $\mathbf{W}_k \in$

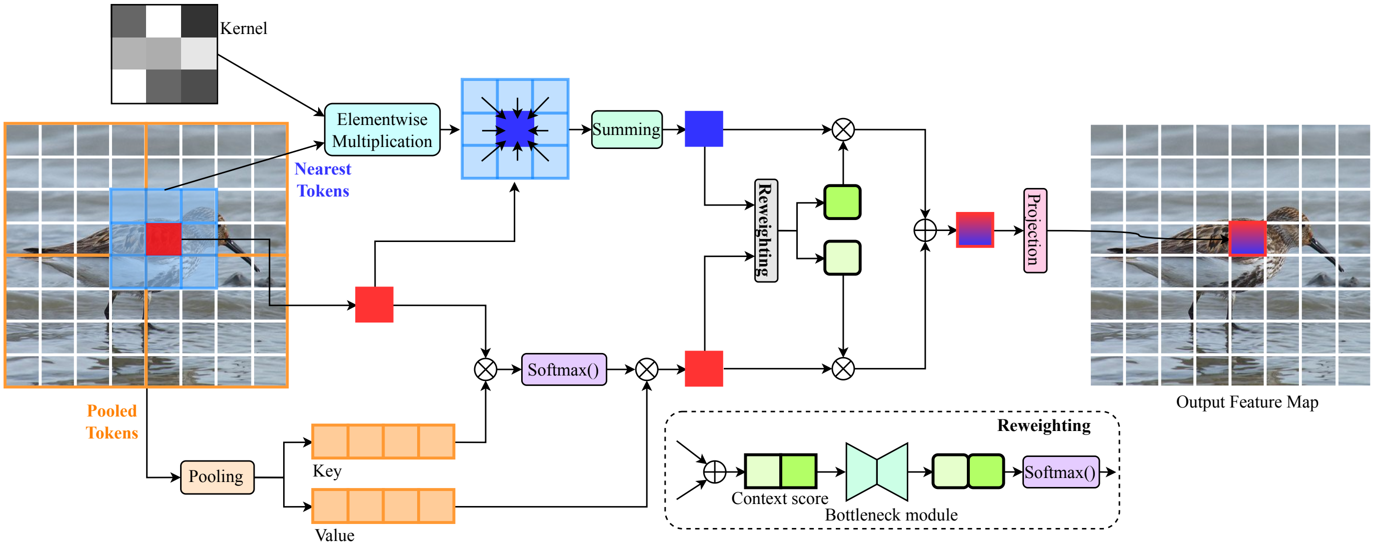


Fig. 5: Detailed structure of the RF attention. Central/para-central regions (nearest tokens to their red query location) are processed by local convolution to capture fine-grained features and far peripheral regions (far tokens from the red query location) are processed by pooling self-attention to model coarse-grained features. Both features are fused through reweighting module to emphasize the importance of each feature to the model learning. 8×8 feature map, 4×4 pooling size, and 3×3 convolution are shown for example.

$\mathbb{R}^{C\times C}, \mathbf{W}_v \in \mathbb{R}^{C\times C}$ are linear projections. The multi-head self-attention is performed on coarse-grained key and value regions to model global features, defined as:

$$\mathbf{y}^p_{i,j} = \underset{h\in[N_h]}{\text{concat}} \left[\text{SA}_h(\mathbf{q},\mathbf{k},\mathbf{v})\right] \mathbf{W}_o, \tag{3}$$

$$\text{SA}_h(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{C_h}}\right)\mathbf{v}, \tag{4}$$

where $\mathbf{y}^p_{i,j}$ is the coarse-grained output at location (i,j) for pooled tokens. N_h is the number of heads and $C_h = \frac{C}{N_h}$ is the head dimension. SA_h denotes self-attention operation for head h. $\mathbf{W}_o \in \mathbb{R}^{C\times C}$ indicates linear transformation to mix information across heads. $\text{concat}[.]$ stands for concatenation operation.

2) *Nearest Tokens:* The nearest regions around query $\mathbf{q}_{i,j}$ are extracted inside each window centered at location (i,j). Fine-grained features are captured by using local convolution to model geometric detail information. Technically, this process is defined as follows:

$$\mathbf{y}^n_l = \sum_{\mathbf{m}\in\mathcal{S}} \mathbf{w}\left(\mathbf{m}\right)\cdot\mathbf{x}\left(\mathbf{l}+\mathbf{m}\right), \tag{5}$$

where $\mathbf{y}^n_l$ is the fined-grained output at location $l=(i,j)$ for nearest regions centered at location (i,j). $\mathbf{w}(m)$ is the kernel weights (gray color in Figure 5) at location $\mathbf{m}\in\mathcal{S}$ and $\mathcal{S}$ is the sampling grid centered at location (i,j) to 8 neighbor tokens:

$$\mathcal{S} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}. \tag{6}$$

$\mathbf{x}(\mathbf{l}+\mathbf{m})$ is the input feature at location $(\mathbf{l}+\mathbf{m})$. Equation 5 is equivalent to elementwise multiplication between kernel weights and nearest regions centered at location (i,j), and information aggregation from neighbor tokens to center tokens (blue tokens in Figure 5).

3) *Reweighting Module:* After acquiring fine-grained and coarse-grained features, the reweighting module is proposed to control the importance of each feature based on their context information. This is achieved by softmax attention:

$$\alpha_{i,j} = \text{Sofmax}(\text{BM}(\text{GAP}(\mathrm{\mathbf{y}^p_{i,j}} + \mathrm{\mathbf{y}^n_{i,j}}))), \tag{7}$$

where $\mathbf{r}_{i,j} = [\alpha_{i,j},\beta_{i,j}]^\top \in \mathbb{R}^{2\times C}$ is reweighting coefficients conditioned on the content of the output fine-grained features $\mathbf{y}^n_{i,j}$ and output coarse-grained features $\mathbf{y}^p_{i,j}$. GAP() is global average pooling applied across the spatial dimension to obtain the context score with dimension $1\times C$. BM() indicates bottleneck module that including two fully-connected layers and GELU() inserted between them. The output of BM() is the context vector with dimension $2\times C$ and Softmax() is applied across the first dimension to create the context score, redistributing the shape of each context value. The context score is used to reweight the importance of fine-grained and coarse-grained features as follows:

$$\mathbf{y}_{i,j} = (\alpha_{i,j}\cdot\mathbf{y}^p_{i,j} + \beta_{i,j}\cdot\mathbf{y}^n_{i,j})\mathbf{W}_p, \tag{8}$$

where $\mathbf{y}_{i,j}$ is the final output at location (i,j). $\mathbf{W}_p$ is the linear projection matrix. The coefficients are constrained as $\alpha_{i,j}+\beta_{i,j}=1$.

B. Model Configuration

Based on the obtained RF attention and RF block, the RF Vision Transformer is introduced in Figure 3. By configuring the number of RF blocks and number of channels across

TABLE I: Model Configurations of RF Transformers

Variant	#Blocks	C	#heads	MLP exp.	#params	GFLOPs
RF-0	[2, 2, 6, 6]	24	[2, 4, 8, 16]	4	5.412	0.422
RF-1	[2, 2, 6, 6]	32	[2, 4, 8, 16]	4	8.492	0.717
RF-2	[2, 2, 8, 6]	48	[3, 6, 12, 24]	4	18.142	1.702

stages, RF variants are obtained in Table I.

#Blocks is the number of stacked RF blocks $[L_1, L_2, L_3, L_4]$ across four stages. Following [7], [14], [16], putting more blocks in stage 3 and stage 4 achieves a better trade-off between accuracy and computational costs while the model can capture more global features. C is the base channel changed across stages with scales $\{1, 2, 4, 8\}$. #heads is the number of heads in pooling self-attention across four stages. MLP exp. indicates the MLP expansion ratio to expand channel dimension in MLP layers, unchanged across stages.

IV. Experiments and Results

To validate the effectiveness of the proposed method, the RF Transformers are trained and evaluated on the ImageNet-1K [11] image classification. After finishing the experiments on the ImageNet-1K dataset, the trained weights of the RF models are transferred to dense prediction tasks such as MS-COCO [26] object detection and instance segmentation, and ADE-20K [27] semantic segmentation. The goal of transferred models is to validate the versatile and general-purpose RF vision Transformers.

TABLE II: Results on ImageNet-1K image classification

Method	Input	#params(M)	GFLOPs	Top-1 Acc.
PVTv2-B0 [15]	224	3.7	0.6	70.5
EdgeViT-XXS [13]	256	4.1	0.6	74.4
Swin-0.7G [7]	224	4.4	0.7	74.4
MobileViT-XS [28]	256	2.3	1.1	74.8
LVT [29]	224	5.5	0.9	74.8
PVT-T [6]	224	13.1	1.6	75.1
EMO-2M [24]	224	2.3	0.5	75.1
RF-0 (Ours)	**224**	**5.4**	**0.4**	**75.5**
EfficientViT-M5 [30]	224	12.4	0.5	77.1
ResT-Lite [31]	224	10.5	1.4	77.2
Swin-1G [7]	224	7.3	1.0	77.3
EdgeViT-XS [13]	256	6.7	1.1	77.5
ConvNeXtV1-F [1]	224	5.2	0.8	77.5
tiny-MOAT-0 [32]	224	3.4	0.8	77.5
FAT-B0 [16]	224	4.5	0.7	77.6
RF-1 (Ours)	**224**	**8.5**	**0.7**	**78.2**
MobileViT-S [28]	256	5.6	2.0	78.4
PVTv2-B1 [15]	224	13.1	2.1	78.7
PerViT-T [23]	224	7.6	1.6	78.8
Swin-2G [7]	224	12.8	2.0	79.2
ConvNeXtV1-P [1]	224	9.1	1.4	79.5
ResT-S [31]	224	13.7	1.9	79.6
RF-2 (Ours)	**224**	**18.1**	**1.7**	**80.8**

A. *Image Classification*

Settings: The image classification experiments are conducted on the ImageNet-1K [11] dataset that includes 1.2M training and 50K validation images with 1,000 categories. The RF Transformers are trained for 300 epochs with a batch size of 1024. The optimizer is AdamW with a learning rate of 10^{-3}, and a weight decay of 0.05. Standard data augmentations are used to train the model such as Rand Augment, Cutmix, Mixup, and label smoothing, similar to existing methods [6], [7], [33]. The input images are resized to 224×224.

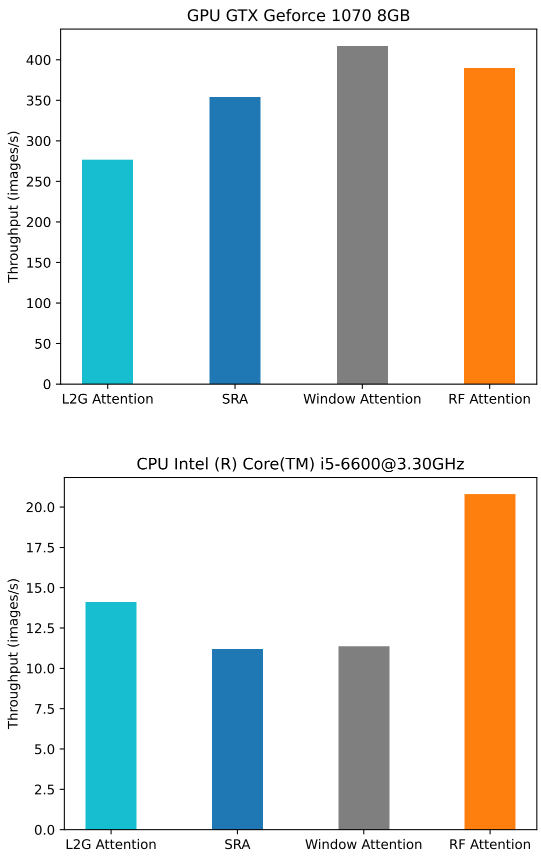


Fig. 6: Speed comparison between local-to-global attention (L2G Attention) [13], spatial reduction attention (SRA) [15], window attention [7], and our RF Attention. Throughput is measured on the same batch size.

Results: Table II reports comparison among recent methods. RF-0 with 5.4M parameters and 0.4 GFLOPs achieves 75.5% Top-1 accuracy that outperforms PVTv2-B0 [15] by 5.0%, competitive method EdgeViT-XXS [13] by 1.1%, and state-of-the-art method EMO-2M [24] by 0.4%. RF-1 gets 78.2% Top-1 accuracy, higher than strong method ResT-Lite [31] by 1.0% with only a half of GFLOPs and lower parameters, ConvNeXtV1-F [1] by 0.7%, and recent method FAT-B0 by 0.6%. For larger settings, RF-2 achieves 80.8% Top-1 accuracy that surpasses MobileViT-S [28] by 2.4%, PerViT-T [23] by 2.0%, and ResT-S [31] by 1.2%. The results verify the efficiency and effectiveness of the proposed methods.

The speed comparisons between recent attentions are provided in Figure 6. As a result, the RF Transformer has a similar speed with SRA [15], window attention [7], and faster than L2G attention [13] on the GPU device. On the CPU device, the proposed RF attention runs faster than other methods while achieving better accuracy shown in Table II.

B. Object Detection

Settings: MS-COCO [26] is used to validate the efficient and general-purpose RF Transformers. MS-COCO has 118K training and 5K validation images with 80 categories. Following recent methods [6], [7], [15], ResNet-50 [22] is replaced with the proposed RF Transformer backbone, and other model settings are kept the same as in RetinaNet [20] for the object detection task. The integrated models are trained for 12 epochs with a batch size of 16. Similar to training receipts [6], [7], AdamW is used as the optimizer with a learning rate of 10^{-3}, and a weight decay of 0.05. The input images are resized to 1333×800. The final results are reported on the MS-COCO validation set.

TABLE III: Object detection results on MS-COCO dataset

Method	#params(M)	GFLOPs	AP	AP^{50}	AP^{75}
ResNet-18 [22]	21	189	31.8	49.6	33.6
ResNet-50 [22]	38	239	36.3	55.3	38.6
ResNet-101 [22]	57	315	38.5	57.8	41.2
PVT-T [6]	23	183	36.7	56.9	38.9
PVT-S [6]	34	226	40.4	61.3	43.0
PVTv2-B0 [15]	13	160	37.2	57.2	39.5
PVTv2-B1 [15]	24	187	41.2	61.9	43.9
Swin-T [7]	38	245	41.5	62.1	44.2
RF-0 (Ours)	**13**	**158**	**38.4**	**59.3**	**40.3**
RF-1 (Ours)	**16**	**164**	**40.6**	**61.5**	**43.2**
RF-2 (Ours)	**26**	**183**	**43.1**	**64.6**	**45.9**

Results: Table III shows the comparison between methods. The RF Transformer achieves consistent improvements compared to other methods. Typically, RF-0 surpasses the baseline ResNet-50 [22] by 2.1% AP while saving 34% GFLOPs and 65% parameters. RF-1 outperforms PVT-T [6] by 3.9% AP with lower GFLOPs and parameters. RF-2 achieves 43.1% AP greater than recent methods, such as PVTv2-B1 [15] by 1.9% AP with similar costs, and competitive method Swin-T [7] by 1.6% AP with only 74% GFLOPs. The results clarify the general and scalable ability of the RF Transformers.

C. Semantic Segmentation

Settings: The proposed RF Transformers are trained and evaluated on ADE-20K [27] for semantic segmentation task using Semantic FPN [34]. For fair comparisons, the training receipts in [6], [7], [15] are adopted to evaluate the performance. The model is trained for 80K iterations with a batch size of 16. The optimizer is AdamW with a learning rate of 10^{-3} and a weight decay of 0.05. The input images are resized to 512×512.

Results: Table IV reports the performance on ADE-20K [27] dataset using Semantic FPN [34]. The RF Transformers gain performance better than the improvement in the object detection task. For instance, RF-0 achieves 40.8 mIoU which outperforms the baseline ResNet-50 [22] by 4.1 mIoU with only a half of GFLOPs and much lower parameters. RF-1 with 11M parameters and 24 GFLOPs gets 42.0 mIoU greater than PVT-S [6] by 2.2% while saving 54% GFLOPs, and PVTv2-B0 by 4.8% with similar costs. RF-2 achieves 44.2 mIoU which surpasses other methods by clear margins, such as the recent method Swin-T by 2.7% with only 63% GFLOPs, and PVTv2-B1 by 1.7% with lower GFLOPs. The performance on visual tasks verifies the efficient and general-purpose RF Transformers.

TABLE IV: Results on ADE-20K semantic segmentation

Method	Crop size	#params(M)	GFLOPs	mIoU
ResNet-18 [22]	512^2	16	32	32.9
ResNet-50 [22]	512^2	29	45	36.7
ResNet-101 [22]	512^2	48	65	38.8
PVT-T [6]	512^2	17	33	35.7
PVT-S [6]	512^2	28	44	39.8
PVTv2-B0 [15]	512^2	8	25	37.2
PVTv2-B1 [15]	512^2	18	34	42.5
Swin-T [7]	512^2	32	46	41.5
RF-0 (Ours)	$\mathbf{512^2}$	**8**	**23**	**40.8**
RF-1 (Ours)	$\mathbf{512^2}$	**11**	**24**	**42.0**
RF-2 (Ours)	$\mathbf{512^2}$	**20**	**29**	**44.2**

V. Conclusion

This paper introduces efficient and versatile RF vision Transformers that leverage biological foveal processing into deep neural networks. The proposed method partitions the image feature map into multiple regions: coarse-grained and fine-grained regions based on the distance from the query location to their contour. In RF attention, each query attends to both fine-grained regions for capturing local features and coarse-grained regions for extracting global features. With this design, the RF Transformer has better efficiency as global self-attention is only performed on low-resolution input while modeling global context. Furthermore, a reweighting module is proposed to capture the relationship between multiple regions based on their features. The proposed RF Transformers are trained and evaluated on various visual tasks: image classification, object detection, and semantic segmentation. In the future, the investigation of the biological foveal system will be further explored in other visual tasks such as video understanding and vision-language models.

Acknowledgment

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] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[2] Y. Xiong, Z. Li, Y. Chen, F. Wang, X. Zhu, J. Luo, W. Wang, T. Lu, H. Li, Y. Qiao *et al.*, "Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications," *arXiv preprint arXiv:2401.06197*, 2024.

[3] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.

[4] H. Chen, X. Chu, Y. Ren, X. Zhao, and K. Huang, "Pelk: Parameter-efficient large kernel convnets with peripheral convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[6] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[8] D. Shi, "Transnext: Robust foveal visual perception for vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[9] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4203–4217, 2022.

[10] W. Lin, Z. Wu, J. Chen, J. Huang, and L. Jin, "Scale-aware modulation meet transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6015–6026.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[12] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4794–4803.

[13] J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, and B. Martinez, "Edgevits: Competing light-weight cnns on mobile devices with vision transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 294–311.

[14] Q. Chen, Q. Wu, J. Wang, Q. Hu, T. Hu, E. Ding, J. Cheng, and J. Wang, "Mixformer: Mixing features across windows and dimensions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5249–5259.

[15] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[16] Q. Fan, H. Huang, X. Zhou, and R. He, "Lightweight vision transformer with bidirectional interaction," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1922–1933, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] J. Min, Y. Zhao, C. Luo, and M. Cho, "Peripheral vision transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 097–32 111, 2022.

[24] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, Y. Wang, and C. Wang, "Rethinking mobile block for efficient attention-based models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2023, pp. 1389–1400.

[25] W. Yu, C. Si, P. Zhou, M. Luo, Y. Zhou, J. Feng, S. Yan, and X. Wang, "Metaformer baselines for vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[27] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[28] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=vh-0sUt8HlG

[29] C. Yang, Y. Wang, J. Zhang, H. Zhang, Z. Wei, Z. Lin, and A. Yuille, "Lite vision transformer with enhanced self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 998–12 008.

[30] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 420–14 430.

[31] Q. Zhang and Y.-B. Yang, "Rest: An efficient transformer for visual recognition," *Advances in neural information processing systems*, vol. 34, pp. 15 475–15 485, 2021.

[32] C. Yang, S. Qiao, Q. Yu, X. Yuan, Y. Zhu, A. Yuille, H. Adam, and L.-C. Chen, "Moat: Alternating mobile convolution and attention brings strong vision models," in *The Eleventh International Conference on Learning Representations*, 2022.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[34] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6399–6408.

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.