Unifying Local and Global Fourier Features for Image Classification

Xuan-Thuy Vo, Jehwan Choi, 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, jhchoi}@islab.ulsan.ac.kr; {ndlinh301, priadana3202}@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE) | 979-8-3503-9971-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIE51358.2023.10227936

***Abstract*—In the last decade, Convolutional Neural Networks (CNNs) have become a dominant algorithm in solving various domains such as computer vision, self-driving cars, medical imaging, and natural language processing. The core operation of the CNNs is convolution layer that can aggregate input features around local windows in a short-range manner and learn relative positions inside each window. For long-range modeling, common CNNs stack a bunch of convolutional layers that result in high computational costs to enlarge receptive field. Recently, Vision Transformers (ViTs) and its improvements have outperformed CNNs in the rankings of language, vision, and audio research. The main goal of the ViTs is that the model can extract short-range and long-range features in one layer. With this strategy, the network structure of the ViTs is simpler than CNNs. However, ViTs have quadratic complexity with the spatial length of the input feature. In the last year, many methods are proposed to relax the cost of ViTs and bring complicated designs of CNNs into ViT-based models. Inspired by the insightful properties of the ViTs and CNNs, this paper introduces a Local and Global Fourier Network (LGFNet) that jointly learns local and global receptive fields in the frequency domain rather than the spatial or time domain in conventional CNNs and ViTs. The input features, local, and global kernels are transformed to the frequency domain through Fast Fourier Transform. The local features are learned by a convolution between the input feature and local kernels. Concurrently, matrix multiplication between the input feature and global kernels is performed to extract low frequencies from the input Fourier feature. Since local and global Fourier features are complementary, the LGFNet efficiently fuses these information by summation operation based on the similarity degrees of the input signals. Therefore, our LGFNet performs unified representation from the input feature. To evaluate the effectiveness of the proposed method, we conduct experiments on the large-scale dataset ImageNet1k and the small dataset CIFAR100. As a result, the LGFNet surpasses the ViT-based models by a clear margin under similar parameters and GFLOPs.**

***Index Terms*—Fourier features, Fast Fourier Transform, Vision Transformer, Image classification**

I. Introduction

With the fast development of parallel computing, in 2012s, AlexNet [1] successfully trained CNNs-based huge models on large-scale datasets and achieved outstanding results compared to classical machine learning algorithms. Based on this milestone, many CNNs-based methods are introduced to improve feature extractions from different perspectives such as plain network VGG [2], residual connection [3], dense connection [4], and inception decomposition [5], and deformable convolution [6]. Based on the designs of the CNN-based model, many works have been introduced to face detection [7], instance segmentation [8], person re-identification [9] and achieved great performance in both accuracy and network efficiency.

As a spirit of the CNNs, convolution operation gathers features from the input around the local neighborhood windows and shares learnable weights across the spatial dimensions. Generally speaking, relative positions and content interactions inside local windows are filled by convolution layers. This creates local connectivity, relative geometry, and translation equivalent of CNN-based models. Because of the physical design of convolution kernels, the receptive field of the convolution operation is limited to the local window. In another word, in one layer, the model only captures short-range dependencies in the input data. To enlarge the view of kernels, a large amount of convolution layers is stacked up to 50 or 100 layers. This way produces high computational costs and optimization problems due to supervised signals far from the input image. For compensation with local connectivity of convolution operation, attention mechanisms are developed in a way that can integrate into CNN models and model long-range relations in one layer. The relation is interpreted through channel interactions [10], and spatial interactions [11]–[14].

Compared with CNNs, ViT-based models perform general relations, such as pixel-to-pixel relations, object-to-pixel relations, and object-to-object relations, through Transformer [14]. Originally, Transformer was developed for machine translation, processing word-to-word (token-to-token) relations in a global way. With the success of the Transformer, ViT [15] considers an image patch as a word and models interactions between patches via Transformer. This strategy results in long-range dependencies and reasons about the uniform representation of visual recognition tasks. There are two main drawbacks of the Transformer when solving visual data. Firstly, Transformer models require 2D flattened input from 3D input data. This requirement lost the order of pixels from the original input because there is no geometric modeling in Transformer. That means Transformer has a weak inductive bias. Secondly, at the heart of the Transformer, the self-attention operation computes similarity maps by the dot product between query and key matrices. As a result, this computation has quadratic complexity with the number of patches $O(N(logN))$, where N is the number of patches. To overcome these drawbacks, many lines of research are conducted such as modeling positional information, enlarging training data, improving training

mechanisms and settings, reducing the computational cost, and inheriting CNNs' designs.

One main line of the research is to build hybrid networks that combine the strengths of the CNNs and ViT models. Hierarchical representation of the CNNs is performed by inserting convolution layers in earlier stages and Transformer layers in later stages. Since low-level and high-level features are extracted in the earlier and later stages respectively, hybrid networks can inherit the hierarchical property of the CNNs and the general modeling capability of ViT. Another benefit of the combination is that the hybrid network can be deployed on edge devices since Transformer layers are applied to the later stages with a low spatial dimension. Another work is to embed convolution operation into self-attention operation because both operations can help each other in terms of local-and-global learning, and supplemented relative bias.

Inspired by the designs of hybrid networks, this paper learns Local-and-Global Fourier (LGF) features in one layer and reduces the cost of the self-attention operation from $O(N^2)$ to $O(Nlog(N))$. The LGF layer contains four main processes:

1) The input features in the spatial domain are transformed to Fourier feature in the frequency domain through Fast Fourier Transform. Local and global kernels are values in the complex domain.
2) Both local and global features are learned simultaneously. High frequencies are extracted by performing convolution between Fourier features and local complex features. Matrix multiplication between Fourier features and global complex kernels is performed to learn low frequencies.
3) The fusion operation is to gather similarity patterns between local and global Fourier features.
4) The fused local-and-global Fourier features are transformed back to the spatial domain by inverse Fast Fourier Transform.

Both convolution and matrix multiplication in the frequency domain are efficient since all the operations are separable. To verify the novel designs of the LGFNet, we replace all Transformer blocks in the ViT model with our LGFNet blocks and keep the training settings similar to the baseline. We train the proposed model on large-scale ImageNet1k and small dataset CIFAR-100. On ImageNet1k, the LGFNet achieves 75.4% Top-1 and 92.6% Top-5 accuracy with 7.5M parameters and 1.2 GFLOPs. On the small dataset CIFAR-100, the finetuned LGFNet gets 86.1% Top-1 and 96.5% Top-5 accuracy.

II. Related Works

In this section, we briefly introduce CNN-based models, Attention and ViT-based models, and Hybrid networks.

A. CNN-based models

The common networks of the CNN-based models have hierarchical representation that each stage extracts specific features from the input with a specific spatial dimension. VGG [2] introduces plain 3×3 convolution for image classification. ResNet [3] pointed out that the VGG network becomes saturated when stacking plain blocks up to 19 layers. From this analysis, ResNet proposes a residual block that can avoid vanishing gradient when stacking more layers. With the effective residual connection, ResNet is set as a baseline for many works related to visual analysis. DenseNet [4] enlarges residual to dense connections that can improve feature learning. Inception [5] split the input feature along channel dimension and apply various operations on each branch. Deformable ConvNets [6] samples and learning offsets to enlarge the receptive field of the kernels via bilinear operation.

B. Attention and ViT-based models

To complement with short-range dependencies of the CNN-based models, some methods [10]–[13] integrate the attention mechanism into stages of the CNN networks. SENet [10] proposes channel attention that can emphasize which channels are important during training through the simple network. Non-local network [11] aggregates features from all positions by classical non-local mean. The works in [12], [13] simplify the structure of the Non-local network and realize that the attention map at one query position that gathers features from all positions has a similar response.

ViT [15] was the first method that successfully apply Transformer for vision tasks. This method views the image as a set of patches and uses self-attention operations to model patch-to-patch relations. With the simple structure, in the last year, many researchers focus on the improvements of the ViT. PVT [16] develops a hierarchical network instead of the uniform network in ViT and reduces the cost of the Transformer through subsampling key and value matrices. Swin [17] splits the input feature into the windows and local self-attention operation is proposed to model relations within windows. TNT [18] learns token-to-tokens interactions with different scales.

C. Hybrid networks

Next-ViT [19] investigates efficient combinations of group convolution and multi-head self-attention to build lightweight models deployed on mobile devices. EfficientFormer [20] adopts MetaBlock [21] and neural architecture search the model can reach real-time speed on mobile devices. Based on the MobileNetv2 [22], a series of MobileViT [23]–[25] is presented by inserting Transformer blocks to later stages of the network. EdgeViT [26] adopts subsampling, self-attention, and upsampling operations into the MobileNetv2 network.

In other research, several methods try to attempt Fast Fourier Transform [27] into CNNs and ViTs networks. FFC [27] separately performs conventional convolution layers on real and imaginary parts of the Fourier features. Inspired by the MLP mixer, AFNO [28] uses linear transformations to mix Fourier features along the channel axis. GFNet [29] replaces Transformer blocks with global filter blocks in which global Gaussian filters are learned by mixing with Fourier features. Differently, this paper separately learns low and high frequencies based on spatial mixings of the global and local complex filters with Fourier features.

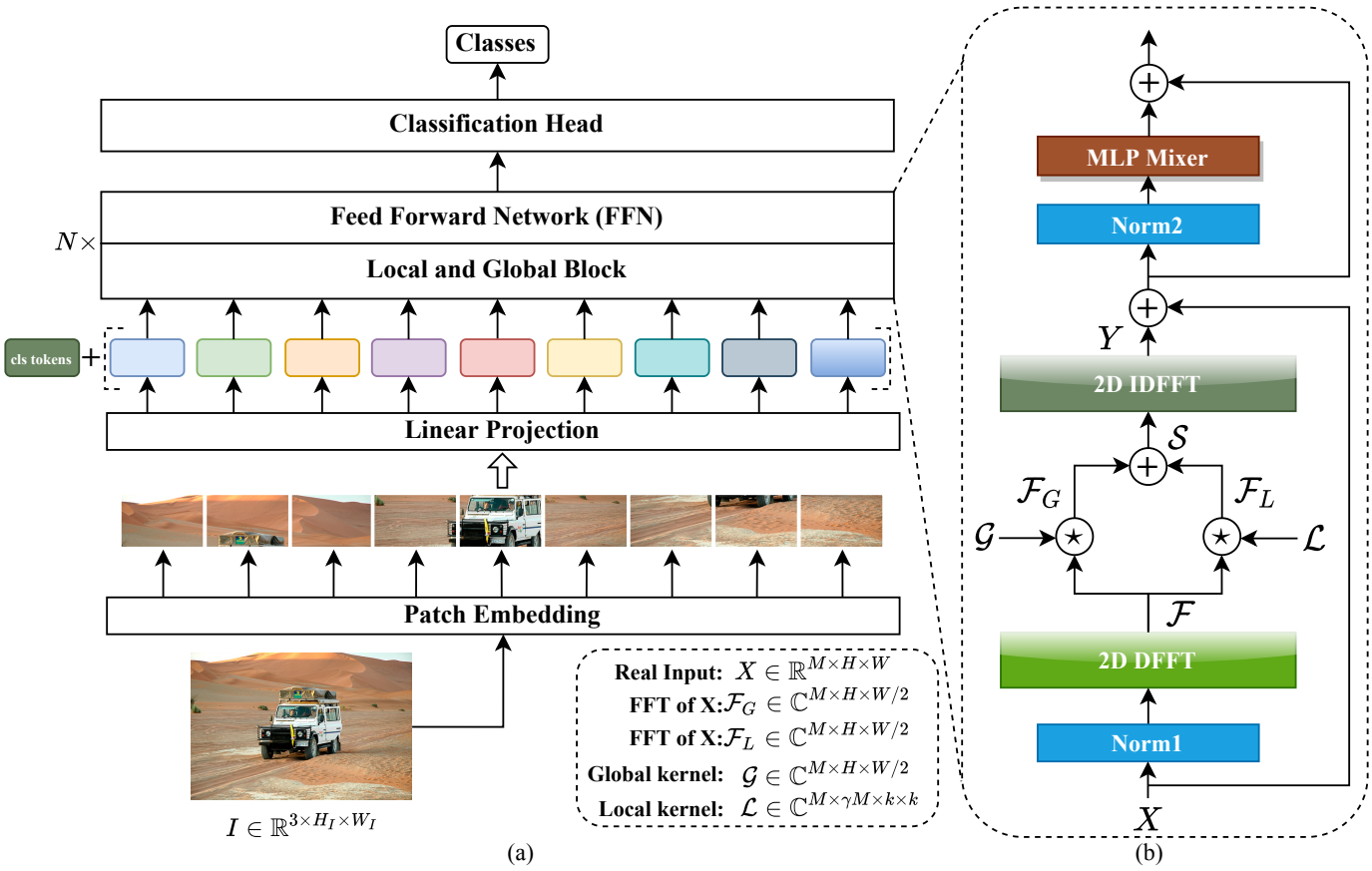

Fig. 1. The overall architecture of the LGFNet. Patch embedding separates the input image into a sequence of patches and linear projection projects the channels of the embedded features into channel model M. Local and Global Block (LGBlock) extracts local and global information via local kernels and global kernels. Classify Head includes one fully-connected layer to map the channel model to the number of the class. M, H, W indicate the number of channels, height and width of the Fourier feature. γ denotes the channel reduction ratio for local kernel $k \times k$. $\mathcal{S}$ is a similarity matrix computed by summation of global and local Fourier features. N is the number of the stacked LGBlocks.

III. The Proposed Method

In this section, the overall architecture of the LGFNet is shown in Fig. 1(a) and analyzed in subsection III-A. The detailed LGBlock is illustrated in Fig. 1(b) and discusses in subsection III-B.

A. Overall architecture

We adopt the common architecture of the ViT [15] as the baseline and replace self-attention blocks with Local and Global Blocks (LGBlock). All other operations are kept the same as the baseline, as follows:

1) **Patch Embedding:** Given the input image $I \in \mathbb{R}^{3\times H_I\times W_I}$, Patch Embedding splits I into embedded feature X with spatial dimension $\{H = H_I/P_H, W = W_I/P_W\}$ and channel dimension $C = 3 * P_H * P_W$. P_H, P_W indicate patch size along height and width axes. Following common implementation, we use patch size 16×16.
2) **Linear Projection:** The role of the linear projection is to map X to latent feature with dimension $\mathbb{R}^{M\times H\times W}$, where M is the number of the channel model.
3) **Feed Forward Network (FFN):** The FFN contains two fully-connected layers that mix tokens across the channel dimension. The first fully-connected layer transforms the input feature into a higher space so that the model can effectively extract information. The second fully-connected layer maps the feature in higher space to low space. The design of the FFN is similar to the inverted bottleneck in MobileNetv2 [22].
4) **Classify Head:** The classifier head includes global average pooling and one fully-connected layer that outputs classification scores.

In the ViT, the cls tokens are supplemented with embedded features and served as the final representation for the classification task. However, in this work, LGBlock still preserves the 2D structure of the input images. Hence, the cls tokens are not important to the proposed model.

B. LGBlock

Like the Transformer block, our LGBlock showed in Fig. 1(b) includes Norm1, spatial mixing, Norm2, and channel mixing. Two residual connections are inserted between two mixings. The main components of the spatial mixing layer

contain four operators: 2D Discrete Fast Fourier Transform (2D DFFT), Local-and-Global learning, similarity map computation $\mathcal{S}$, and 2D Inverse Discrete Fast Fourier Transform (2D IDFFT).

1) 2D DFFT: Instead of learning features in the spatial domain, this work tries to extract the spectrum of the helpful frequencies through Fast Fourier Transform and learnable complex filters. Given the real input $X \in \mathbb{R}^{M\times H\times W}$, the LGBlock transforms this feature to the Fourier feature as follows:

$$\mathcal{F}[:,u,v] = \sum_{m}^{H-1}\sum_{n}^{W-1} X[:,m,n]e^{-j2\pi(\frac{um}{H}+\frac{vn}{W})}, \quad (1)$$

where u, v are the index of the Fourier feature $\mathcal{F}$ and m, n are the index of the real feature X. This operation is efficient since the DFFT is separable. The content and geometry of the real feature when transforming are still preserved. As seen in the equation 1, the Fourier feature contains a spectrum of the frequencies. The amplitude of the Fourier feature is the main information of this transform and a function of the frequencies. The goal is to extract helpful frequencies from the Fourier feature that can increase representation ability.

One of the main properties of the FFT is that the conjugate symmetry of the Fourier features is revealed because $\mathcal{F}[:,u,v] = \mathcal{F}^*[:,H-u,W-v]$. Therefore, we only calculate a haft of Fourier feature $\mathcal{F}[:,:,W/2+1]$, and the cost of the next computations is reduced.

2) Local-and-Global learning: The aim of this process is to extract low and high frequencies separately. In the Fourier theory, the low frequencies contain high-level features and the high frequencies contain low-level features. To properly get full information from the Fourier feature, local and global complex kernels are defined as $\mathcal{L} \in \mathbb{C}^{M\times\gamma M\times k\times k}$ and $\mathcal{G} \in \mathbb{C}^{M\times H\times W/2}$. The high frequencies are learned by the convolution between local kernel $\mathcal{L}$ and Fourier feature $\mathcal{F}$, stored in local Fourier feature $\mathcal{F}_L$,

$$\mathcal{F}_L = \mathcal{F} \star \mathcal{L}, \quad (2)$$

where $\star$ indicate convolution operation. γ denotes channel reduction ratios and we set $\gamma = 1/M$ same as in depth-wise separable convolution. $k \times k$ is the kernel size of the convolution operation. In the FFC [27] and AFNO [28], linear transformations are performed in real and imaginary parts individually. Otherwise, in this design, the convolution operation is computed in the frequency domain. It means that there is interaction learning between real and imaginary parts of the Fourier feature.

The complementary with the local Fourier feature is the global Fourier feature the model can extract low frequencies from this process. The computation of the global complex kernel $\mathcal{G} \in \mathbb{C}^{M\times H\times W/2}$ and Fourier feature $\mathcal{F} \in \mathbb{C}^{M\times H\times W/2}$ is shown as,

$$\mathcal{F}_G = \mathcal{F} \star \mathcal{G}, \quad (3)$$

Both local and global Fourier features are learned simultaneously. There is a way the model can efficiently fuse two features together.

3) Similarity map: Following the superposition property of the complex number, both local and global Fourier features are aggregated through summation operation. The superposition of the two features is computed as follows:

$$\mathcal{S} = \mathcal{F}_L + \mathcal{F}_G. \quad (4)$$

Because local and global information is complementary, the summation operation can measure which tokens are similar. In another word, the fused features that can be enhanced or weakened depend on their similarity. If a position in $\mathcal{F}_L$ has the same phase as a position in $\mathcal{F}_G$, the output feature is enhanced by each other and otherwise.

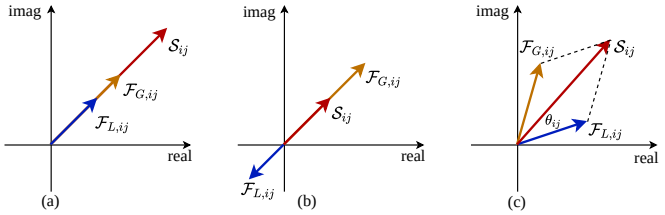

Fig. 2. Superposition of the local and global Fourier features in complex domain. (a) two tokens have the same phase, (b) two tokens have the opposite phase, and (c) general case. $\mathcal{S}_{ij}$ is the fused Fourier token from a pair of local-and-global tokens. θ_{ij} is the phase between a local and global token. imag, real indicates imaginary and real axes, respectively.

Fig. 2 illustrates different cases of the token superposition. Case (a) shows two feature tokens have the same phase and the fused feature is enhanced. Case (b) describes two tokens that have the opposite phase and the fused feature is weakened. The general case is shown in Fig. 2(c) where the amplitude of the fused feature relies on the phase of two tokens. Mathematically, the amplitude of the fused result $|\mathcal{S}_{ij}|$ is calculated as,

$$|\mathcal{S}_{ij}| = \sqrt{|\mathcal{F}_{L,ij}|^2 + |\mathcal{F}_{G,ij}|^2 + 2|\mathcal{F}_{L,ij}||\mathcal{F}_{G,ij}|cos(\theta_{ij})}, \quad (5)$$

As shown in Equation 5, the phase θ_{ij} directly affects to amplitude of the fused tokens.

4) 2D IDFFT: The fused local-and-global Fourier feature $\mathcal{S}$ is inverted back from complex domain to spatial domain through 2D inverse discrete fast Fourier transform defined as follows,

$$Y[:,m,n] = \frac{1}{\sqrt{H*W}}\sum_{u}^{H-1}\sum_{v}^{W-1} \mathcal{S}[:,u,v]e^{j2\pi(\frac{um}{H}+\frac{vn}{W})}. \quad (6)$$

Similarly, the 2D IDFFT operation still preserves the information from the Fourier domain.

Since, in the 2D DFFT and 2D IDFFT, the accumulation phenomenon is revealed, the normalization term $1/\sqrt{H*W}$ is added after the forward or inverse process. This normalization makes Fourier features orthonormal.

IV. Experiments and Results

A. Experiments

The proposed LGFNet is conducted and evaluated on the large-scale dataset ImageNet1k [30] and small dataset CIFAR-100 [31]. The ImageNet1k contains 1.2M training images and 50k validation images. The CIFAR-100 includes 50k training and 10k testing images. The number of classes of the ImageNet and CIFAR-100 is 1000 and 100, respectively.

The Pytorch framework is used for all implementations. Numerical computation of the 2D DFFT, 2D IDFFT, and local-and-global complex values is well supported by the *cuFFT* library. Flowing common methods, the codebase *Timm* [32] is utilized. All the implementation settings are kept the same as the baseline ViT [15] for fair comparisons. Specifically, the LGFNet is trained for 300 epochs, on two GPU Tesla V100-32GB. We set a batch size of 512 per GPU and the total batch size is 1024 for one-time updating network parameters. For computing loss, the Cross-Entropy algorithm is adopted. The AdamW optimizer with a momentum of 0.9 and a weight decay of 0.05 is used along with the learning rate of $5 \times e^{-4}$. The cosine learning schedule is to manage the learning rate with the warmup epochs of 5. The input image of the model is resized to 224×224.

B. Results

1) Results on the large-scale dataset ImageNet: The comparison between the proposed LGFNet and state-of-the-art ViT-based models on the ImageNet validation set is shown in Table I. *Size* indicates the input image size resized to 224×224. *P (M)* and *G* are the number of parameters with the unit Millions and GFLOPs. *Top-1* and *Top-5* denote Top-1 Accuracy (%) and Top-5 Accuracy (%).

TABLE I
Results of the LGFNet and recent ViT-based models on the ImageNet dataset

Method	Size	P (M)	G	Top-1 (%)	Top-5 (%)
T2T-ViT [33]	224	4.3	1.2	71.7	-
DeiT-T [34]	224	5.7	1.3	72.2	91.1
PiT-Ti [35]	224	4.9	0.7	72.9	-
ConViT-Ti [36]	224	5.7	1.4	73.1	91.7
CrossViT-Ti [37]	224	6.9	1.6	73.4	-
GFNet-Ti [29]	224	7.5	1.3	74.6	92.2
LocalViT-T [38]	224	5.9	1.3	74.8	92.6
LGFNet (ours)	224	**7.5**	**1.2**	**75.4**	**92.6**

As described in Table I, our LGFNet achieves 75.4% Top-1 and 92.6% Top-5 accuracy with the tiny model setting around 7.5 M parameters and 1.2 GFLOPs. This performance surpasses the recent ViT-based models such as T2T-ViT [33] with 71.7% Top-1 accuracy, DeiT-T [34] with 72.2% Top-1 accuracy, PiT-Ti [35] with 72.9% Top-1 accuracy, ConViT-Ti [36] with 73.1% Top-1 accuracy, CrossViT-Ti [37] [37] with 73.4% Top-1 accuracy, GFNet-Ti [29] with 74.6% Top-1 accuracy, and LocalViT-T [38] with 74.8% Top-1 accuracy. This comparison clarifies the effectiveness of the local-and-global Fourier feature learning.

TABLE II
Results of the LGFNet and recent ViT-based models on the CIFAR-100 dataset

Method	#Params (M)	GFLOPs	Top-1 Acc (%)
DeiT-T [34]	5.4	0.4	67.59
DeiT-S [34]	21.4	1.4	66.55
PVT-T [16]	15.8	0.6	69.62
PVT-S [16]	27	1.4	69.79
Swin-T [17]	27.5	1.4	78.07
CvT-13 [39]	19.6	4.5	81.81
DHVT-T [40]	5.8	1.4	83.57
DHVT-S [40]	22.8	5.6	85.68
LGFNet (ours)	**7.5**	**1.2**	**86.10**

2) Results on the small dataset CIFAR-100: Table II shows the performance of the LGFNet and ViT-based models on the small dataset CIFAR-100.

As a result, the LGFNet with only 7.5M parameters and 1.2 GFLOPs outperforms other methods by a clear margin. While scaling the ViT-based models from the tiny to large version, the improvement is minor such as 0.2% gain in PVT [16], 2% gain in DHVT [40]. It demonstrates the generalization capability of the proposed method on both small and large datasets.

TABLE III
The investigation on the local Fourier kernel $\mathcal{L}$

Kernel size k	#Params (M)	GFLOPs	Top-1 (%)	Top-5 (%)
1×1	7.5	1.2	75.1	92.2
3×3	**7.5**	**1.2**	75.4	92.6
5×5	7.6	1.3	**75.5**	**92.7**

3) Ablation study: We investigate the effect of the kernel size k in local Fourier learning on the performance of the model shown in Table III.

When changing the kernel size from $1 \rightarrow 5$, the performance is improved from 75.1% to 75.5% Top-1 accuracy and the increases in the cost are acceptable. For balancing accuracy and computational cost, we set $k = 3$ in all experiments and comparisons.

4) Kernel visualization: We visualize the amplitude spectrum of the global and local complex kernel illustrated in Fig. 3 and 4. In this experiment, the learnable weights over 12 stacked LGF layers are shown, and in each layer, 24 channels of global and local Fourier kernels are analyzed.

As seen in Fig. 3, over 12 layers, the amplitude spectrum of the global kernel is diverse and has symmetry demonstrated in subsection III-B. This information means that, in different layers, the model captures the different frequencies, a wide spectrum of the frequencies.

Otherwise, in Fig. 4, local kernel values are changing along axes. It means that the local Fourier kernels focus on low-level information (edge information) and only extract high frequencies.

From two visualizations, we verify that local and global kernels are complementary, and leveraging these cues into the model learning can capture full information from the input image.

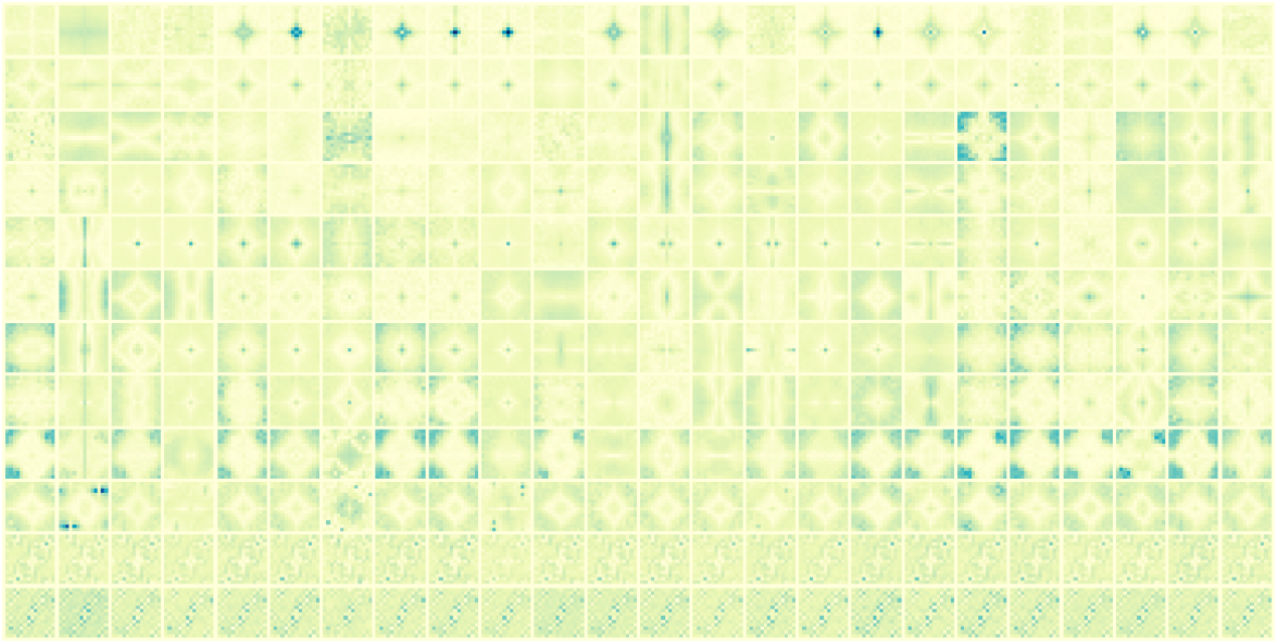

Fig. 3. The visualization of the global complex kernel $\mathcal{G}$ across 12 layers of the LGFNet. 24 channels in the global complex kernel are used to compute the amplitude spectrum of the learnable kernels.

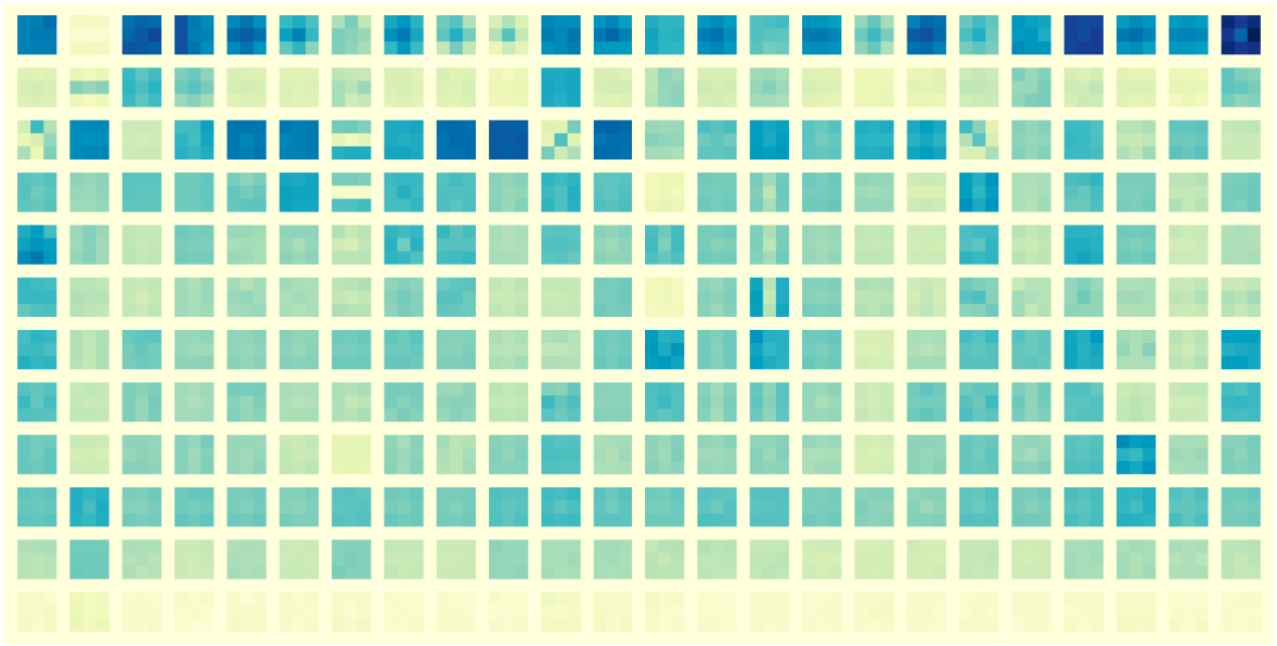

Fig. 4. The visualization of the local complex kernel $\mathcal{L}$ across 12 layers of the LGFNet.

V. Conclusion

This paper inspects local and global extraction of the ViT-based model in the frequency domain via the Fast Fourier Transformer algorithm. Local and global complex kernels capture different information from the input. The local features are presented by high-frequency representation and global features bring information about the low frequencies. Both signals are complementary and fusing these terms into one layer can enhance model learning. Numerical and visualized results demonstrate the effectiveness of the LGFNet and its theoretical analysis. On two kinds of datasets, our LGFNet outperforms recent ViT-based methods under the same settings.

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] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] 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.

[4] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[7] M. D. Putro and K.-H. Jo, "Fast face-cpu: a real-time fast face detector on cpu using deep learning," in *2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2020, pp. 55–60.

[8] X.-T. Vo, T.-D. Tran, D.-L. Nguyen, and K.-H. Jo, "Multi-level feature reweighting and fusion for instance segmentation," in *2022 IEEE 20th International Conference on Industrial Informatics (INDIN)*. IEEE, 2022, pp. 317–322.

[9] Q. Tang and K.-H. Jo, "Fully unsupervised person re-identification via centroids and neighborhoods joint learning," in *2022 IEEE 31st International Symposium on Industrial Electronics (ISIE)*. IEEE, 2022, pp. 1127–1132.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[11] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[12] X.-T. Vo, L. Wen, T.-D. Tran, and K.-H. Jo, "Bidirectional non-local networks for object detection," in *Computational Collective Intelligence: 12th International Conference, ICCCI 2020, Da Nang, Vietnam, November 30–December 3, 2020, Proceedings 12*. Springer, 2020, pp. 491–501.

[13] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[14] 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.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] 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.

[17] 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.

[18] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[19] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, and X. Pan, "Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios," *arXiv preprint arXiv:2207.05501*, 2022.

[20] Y. Li, G. Yuan, Y. Wen, E. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at mobilenet speed," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.

[21] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 819–10 829.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[23] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[24] ——, "Separable self-attention for mobile vision transformers," *arXiv preprint arXiv:2206.02680*, 2022.

[25] S. N. Wadekar and A. Chaurasia, "Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," *arXiv preprint arXiv:2209.15159*, 2022.

[26] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*. Springer, 2022, pp. 294–311.

[27] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[28] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, "Efficient token mixing for transformers via adaptive fourier neural operators," in *International Conference on Learning Representations*, 2021.

[29] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[30] 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.

[31] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[32] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.

[33] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.

[34] 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.

[35] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 936–11 945.

[36] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2286–2296.

[37] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[38] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," *arXiv preprint arXiv:2104.05707*, 2021.

[39] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[40] Z. Lu, H. Xie, C. Liu, and Y. Zhang, "Bridging the gap between vision transformers and convolutional neural networks on small datasets," *arXiv preprint arXiv:2210.05958*, 2022.