IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 May 2024, accepted 3 June 2024, date of publication 5 June 2024, date of current version 14 June 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3410231*

RESEARCH ARTICLE

# DCT-ViT: High-Frequency Pruned Vision Transformer With Discrete Cosine Transform

**JONGHO LEE, (Member, IEEE), AND HYUN KIM, (Senior Member, IEEE)**
Department of Electrical and Information Engineering, Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Hyun Kim (hyunkim@seoultech.ac.kr)

This study was supported by the Research Program funded by the SeoulTech (Seoul National University of Science and Technology).

**ABSTRACT** Transformers have demonstrated notable efficacy in computer vision, extending beyond their initial success in natural language processing. The application of vision transformers (ViTs) to resource-constrained mobile and edge devices is hampered by their extensive computational demands and large parameter sets. To address this, research has explored pruning redundant components of ViTs. Given that the computational burden of ViTs scales quadratically with token count, previous efforts have aimed to decrease the number of tokens or to linearize the computational cost of self-attention. However, such methods often incur significant accuracy losses due to the disruption of critical information pathways within the ViT, which primarily focuses on shape rather than texture, potentially aligning its image interpretation more closely with human perception than convolutional neural network (CNN) models. This observation parallels the effectiveness of JPEG, a predominant image compression standard, which maintains high compression efficacy with minimal quality degradation by discarding high-frequency details that have less impact on human object recognition. In this work, we harness the discrete cosine transform (DCT), an integral component of JPEG, to enhance ViT performance. We considerably reduced computational demands by selectively eliminating high-frequency tokens via DCT while maintaining model accuracy. For instance, our DCT-enhanced ViT model exhibited a 25% reduction in computational costs relative to the DeiT-Small model on ImageNet, with an accuracy increase of 0.18% and only a 0.72% accuracy decrease at a 44% computational reduction. Compared to the DeiT-Tiny model, our approach improved accuracy by 0.17% despite a 47% decrease in computational costs. Furthermore, the proposed DCT-ViT model necessitates significantly fewer parameters than existing approaches, offering a more efficient alternative for deploying ViTs on edge devices.

**INDEX TERMS** Deep learning, discrete cosine transform, frequency domain, image classification, token pruning, vision transformer.

## I. INTRODUCTION

Recent advances in convolutional neural network (CNN)-based architectures [1], [2], [3] have led to substantial achievements across various computer vision applications, including image classification, object detection, and semantic segmentation [4], [5], [6], [7], [8]. Nonetheless, the introduction of vision transformers (ViTs) [9], employing the distinctive multi-head self-attention (MSA) mechanism and eschewing traditional convolutional frameworks, has heralded new standards of performance in these areas. This transition underscores the structural and operational divergence between CNNs and ViTs, wherein the latter interprets images through the lens of tokens, a method derived from natural language processing, enhancing global contextual understanding beyond the local scope typically emphasized by CNNs [10], [11], [12].

Despite this success, ViT has the disadvantage of quadratically increasing the number of computations in the number of input tokens; thus, considerable effort has been made to

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

solve this problem [11], [13], [14]. Studies by [15] and [16] proposed a method for reducing the amount of computation and inference time by reducing the number of unnecessary tokens as the depth increases. In [17], [18], and [19], methods were proposed to adaptively reduce the computational cost by comprehensively considering the number of ViT tokens, number of MSA channels, multilayer perceptrons (MLPs), and layer depth. Other studies [11], [20], [21] reduced the amount of self-attention computation by linearly increasing it in the sequence length. In these studies, when the compression rate was not high, the accuracy drop may have been significantly small or increased; however, serious performance degradation occurred when the compression rate was high (*i.e.*, reducing the calculation amount by more than half). Moreover, the existing methods are limited in that they cannot achieve excellent lightweight performance in ViT models that are sufficiently small to operate on mobile/edge devices.

This study proposes a high-frequency token-pruning method that can maintain the accuracy of ViT as much as possible and effectively reduce the computational cost by utilizing the discrete cosine transform (DCT) [22], which is traditionally and still widely used in image processing. When people look at images, they recognize objects using low-frequency rather than high-frequency information. Therefore, the JPEG algorithm [23], which is most commonly used for image compression, achieves a high compression rate by separating the high- and low-frequency information of the image through DCT and discarding the high-frequency information. DCT is a type of Fourier transform that allows input signals to be expressed as the sum of cosine functions. However, it is more suitable for compression tasks to operate on mobile/edge devices than the discrete Fourier transform (DFT) [24] because it effectively separates the high- and low-frequency elements of input sequences or images even with only the real part (i.e., without the imaginary part) by focusing the energy of the signal on a few coefficients. Therefore, we propose a lightweight ViT model that effectively removes high-frequency elements of ViT via DCT (DCT-ViT), inspired by the characteristics of ViT recognizing shapes like humans, rather than CNN models [25], [26], which recognize objects through texture. We used DCT to find unnecessary tokens in the frequency domain and remove them more effectively. As shown in Fig. 1, we train the proposed DCT-ViT from scratch on ImageNet-1k and improve accuracy by more than 0.4% with less computation than DeiT-S [10] without token pruning. Moreover, we reduced the computational cost while maintaining a lower accuracy drop than the state-of-the-art (SOTA) ViT token pruning methods [15], [16]. Consequently, the proposed DCT-ViT achieves a significantly better trade-off between accuracy and computation costs than other methods for pruning the MSA, MLP layer, and layer depth of ViT [17], [18], [27]. In particular, it is noteworthy that the performance improvement is excellent in areas lower than 1G multiply-accumulates (MACs), which is the level of model

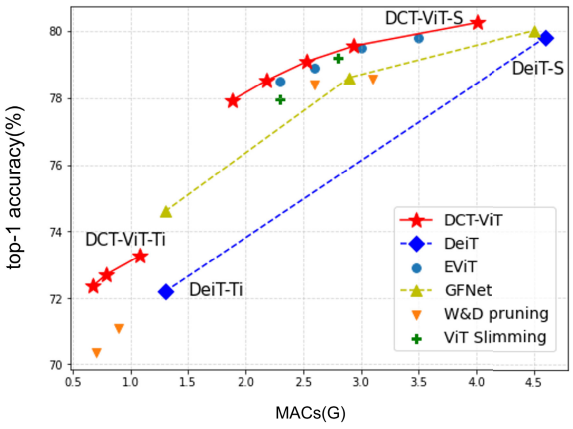


**FIGURE 1. Comparison of image classification performance with ours, DeiT, and other efficient vision transformer models on the ImageNet-1k dataset. 'MACs' in the x-axis stands for multiply-accumulates, a metric used to quantify the computational cost associated with a model operation.**

complexity that can be used in mobile/edge devices [3]. Our contributions can be summarized as follows:

- Our method is the first application of DCT, an algorithm commonly used for image compression, to ViT model compression, and shows outstanding performance with less computation than DeiT and the SOTA ViT pruning techniques.
- Even if the DCT-ViT applies a high compression (i.e., pruning) rate by removing high frequency, which is unnecessary information for human visual recognition, it maintains a small accuracy drop. Because the compression rate can be continuously selected in the DCT-ViT model, the size of the pruned model can be flexibly adjusted.
- As unnecessary frequency tokens are statically removed, additional modules or learning techniques are not required to determine the importance of the token. Specifically, unimportant tokens are effectively removed statically without additional computational cost.

## II. RELATED WORKS

### A. VISION TRANSFORMER

Recently, ViT [9], which is inspired by the transformer model that has been actively studied in the field of natural language processing (NLP), has received considerable attention and has shown good performance in various vision tasks, such as image classification, object detection, and semantic segmentation [11], [28], [29]. In ViT, MSA operations in the field of computer vision have become possible through patch embedding, which allows a single image token to act as a word in a sentence. DeiT [10] is a ViT-like architecture; however, several learning techniques have been proposed to enable us to achieve better performance with only ImageNet-1K [30] training. Self-attention, which is the basis of the ViT operation, differs significantly from CNN's convolution operation, including the amount of computation

being quadratic to the input resolution. ViT relies on self-attention operations that aggregate spatial information by modeling token interactions. Therefore, unlike a CNN that captures local information, ViT better captures global information and long-range interactions. In contrast to CNN, which prioritizes texture information, ViT is known to exhibit shape bias [25]. Park and Kim [31] proposed that the multi-head self-attention (MSA) block in ViT functions as a low-pass filter, unlike the convolution block in CNNs, which operates as a high-pass filter. Through these characteristics and the SOTA methodologies of the MSA, ViT models have demonstrated SOTA performance in various vision tasks [21], [32]. However, because ViT incurs a quadratic computational cost for input resolution, CNN models are still mainly used in mobile/edge device environments where power and hardware resources are limited [33], [34], [35].

### B. EFFICIENT VISION TRANSFORMERS

It is essential to improve computational efficiency to deploy transformer models on mobile/edge devices. ViT models perform well when sufficient computation and parameters are available; however, if the computing power is limited, they may perform worse than CNN models [2], [34], [36]. To solve this problem, various lightweight studies including MobileVit [37] and EfficientVit [38] have been conducted that consider the trade-off between accuracy and computation in ViT models. Zhu et al. [39] introduced a learnable coefficient to reduce embedding dimensions and remove neurons with small coefficient values. DynamicViT [15] reduced computation by using prediction modules to eliminate less important tokens, although further learning is required for the prediction modules. Another method of abandoning tokens proposed in [16] gradually eliminated unnecessary tokens as they changed to the latter layer using class-token information in the vanilla ViT without additional modules or learning to sort important tokens from unnecessary tokens. However, this method increases performance degradation by removing important areas if the area occupied by the object in the image is large. If 50% of the computation is reduced, only 13% of the token remains in the last layer. In addition, because a class token is used, applying this method to hierarchical ViT models [13], [21], [40] that employ global average pooling at the last layer without using a class token is difficult.

Some studies [17], [18], [19] proposed a method for aggressively reducing the computation by comprehensively considering the number of ViT tokens, number of channels/dimensions of MSA and MLP, and layer depth. However, [18] and [19] reduced the computation amount by considering several constraints but underperformed the case when the computation amount of the same ratio as that of [16] was reduced, and [17] increased the hyper-parameters to find the best option and required additional learning methods. In the case of studies using linear attention in NLP tasks, [41] and [42] approximated softmax, which makes the amount of attention quadratic to the input length and consequently changes the amount of operation linearly. However, in these studies, there was a significant performance gap for reasons such as the inability to grasp local features better than conventional attention [43].

### C. DEEP LEARNING IN FREQUENCY DOMAIN

The DCT has traditionally played an important role in signal processing [44]. In particular, in areas of image/video compression, such as JPEG [45] and MPEG [46], low-frequency and high-frequency information can be distinguished by expressing the input signals as the sum of the real cosine function, thus taking advantage of high- energy compression. With the recent remarkable advances in CNN and transformers in vision and NLP tasks, various attempts have been made to apply frequency-domain approaches to deep learning models [20], [47], [48], [49].

In the category of CNN models, [47] converted the input image, which usually enters an RGB of $224 \times 224 \times 3$, into a YUV image with higher resolution, and then DCTs were performed every $8 \times 8$ blocks to rearrange the same frequency information on the same channel. After learning unnecessary channels in the inference phase, the channels are removed to reduce the input image bandwidth of the CNN models, and the performance improves slightly. However, this yields no reduction in computation, and ultimately, unnecessary frequency-erasing channels show little difference in performance from statically erasing high frequencies. Liu et al. [49] performed iterative pruning after sending the weight matrix to the frequency domain through DCT when performing convolution operations in CNN models. This approach effectively reduces the parameters by removing relatively unnecessary high-frequency weights through DCT and restoring them close to the original model through IDCT.

In the NLP task category, Fnet [50] proposed a new method that combines tokens in NLP tasks using Fourier transforms. This suggests that instead of using the traditional attention mechanism to combine tokens, Fourier transforms can be used to transform each token vector from the time to the frequency domain. The resulting frequency vectors can be linearly mixed to obtain a new set of token vectors, which are then transformed back into the time domain using an inverse Fourier transform. FNet achieves competitive results in several NLP tasks, including text classification and language modeling, while requiring fewer computational resources than attention-based models. The DCT-former [48] is also based on an NLP model. It applies DCT to the query, key, and value in the MSA layer to remove high-frequency information and reduce the network size. Subsequently, IDCT is performed again to approximate the MSA layer. However, this method is quite different from our method in that it seeks to reduce the computations of the MSA itself and is accompanied by performance degradation of the MSA. In addition, if DCT is added to the MSA layer, the low-frequency (i.e., DC) value becomes too large, resulting in the divergence of losses in the training of vision tasks.

In the ViT category, GFNet [20], which replaces the MSA layer with a layer composed of a 2D discrete Fourier transform (2D DFT) and global filters (GFs), exhibits competitive performance over DeiT models [10] without MSA layer, which is thought to be a large part of the transformer. GFNet uses a GF layer as an alternative to the self-attention layer. The GF layer converts the spatial dimensions into the frequency domain, which in turn can mix tokens representing different spatial locations using 2D DFT along the spatial dimensions. GFNet argues that these GFs have the same meaning as a global circular convolution with a size of $H \times W$, and the computation is smaller. The computational cost of the GF layer is insignificant compared to that of the MSA layer of the ViT model because the complexity of self-attention is $O(HWD^2 + H^2W^2D)$ whereas the computational complexity of the GF is $O(HWD[log_2(HW) + HWD)$. Here, 'H', 'W', and 'D' represent the height, width, and depth of the input tensor, respectively. 'HW' denotes the product of height and width, indicating the spatial dimension of the input. However, the performance degradation is severe because GFNet removes all self-attention operations from the ViT structure. To address this issue, GFNet enhances the depth of the layers and expands the hidden dimensions. For example, the depth of GFNet-S increases from 12 to 19, and the dimension of GFNet-Ti increases from 192 to 256. This is because the decrease in accuracy is quite large when GFNet is trained under the same conditions as DeiT. Finally, GFNet overcomes this problem by selecting the optimal hyperparameters after increasing the depth and dimensions without any rules, thus increasing the parameters when the amount of computation is similar to that of the baseline model.

## III. PROPOSED METHOD

### A. DISCRETE COSINE TRANSFORM

The DCT belongs to the Fourier transform family and is important in digital signal processing. This is similar to the DFT used in the discrete domain; however, there is a significant difference. First, the DFT is Fourier transformed for the discrete input signal, and it is possible to express input signals as a linear combination with an exponential function; however, the DCT can express cosine signals as a result of real numbers. The DCT is also known to achieve better energy compaction than the DFT because the coefficients are not correlated with each other and are excellent for separating low and high frequencies, which are considered important in image processing. Owing to this characteristic, the DCT can be well restored when an IDCT is performed, even if a large amount of high-frequency information is lost. The IDCT can be completely restored to the original signal unless the coefficient is impaired by the inverse transformation of the DCT. Using these characteristics, JPEG [23], a representative image-compression algorithm, reduces the amount of information by performing a DCT and quantizing a given signal to remove high-frequency

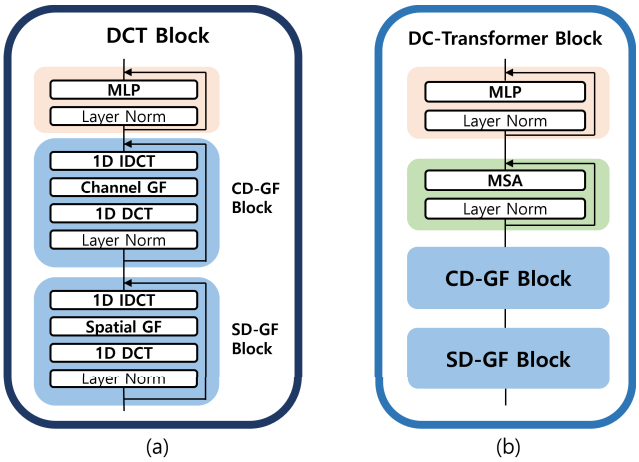


**FIGURE 2.** DCT and DC-Transformer blocks used in DCT-ViT. (a) shows the overall DCT block with spatial, channel-wise, and MLP transforms. (b) depicts the DC-Transformer incorporating DCT blocks within the standard transformer layer. In the spatial/channel GF layer, the learnable filter 'K' is applied after the DCT to capture the frequency components before IDCT.

components that are not well recognized by the human eye. The 1D DCT with sequence $x$ and finite length N can be calculated as follows:

$$X_k = \alpha(k) \sum_{n=0}^{N-1} x_n \cos\left[\frac{(2n+1)k\pi}{2N}\right] \quad \text{for } k = 0, \ldots, N-1 \tag{1}$$

where $\alpha$ is a scale factor for ensuring that the transform is orthonormal, and is given as follows:

$$\alpha(k) = \begin{cases} \frac{1}{\sqrt{N}} & \text{if } k = 0 \\ \frac{1}{\sqrt{2N}} & \text{otherwise} \end{cases} \tag{2}$$

It should be noted that DCT operations can be easily implemented through the Pytorch library and can be implemented with simple matrix multiplication without using Pytorch.

### B. DCT BLOCK

GFNet [20], which replaces MSA operations with the DFT in ViT, performs 2D DFT on normalized tokens in the GF layer, element-wise multiplication of learnable parameters, and 2D IDFT. In contrast, we adopted the DCT, which is commonly used for loss-image compression, to cut high-frequency components because it can collect low frequencies at fewer coefficients than the DFT. The DCT is more suitable for the model compression of mobile/edge devices because it has better energy compaction than DFT and operates in real numbers without imaginary parts. We propose a new DCT-based GF block in spatial tokens (SD-GF) and DCT-based GF blocks in channel dimensions (CD-GF). These layers have significantly low computations because, similar to the GF block of GFNet, the computational complexity is $O(HWD[log_2(HW)] + HWD)$. The SD-GF and CD-GF

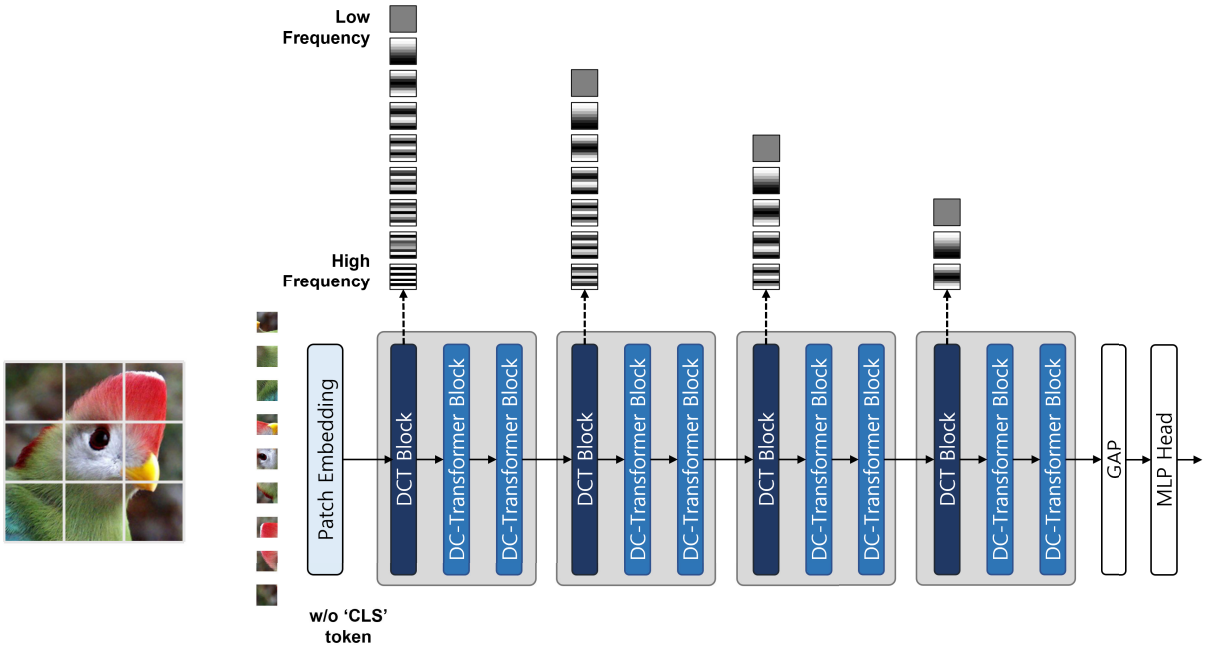


**FIGURE 3.** Overall architecture of the proposed DCT-ViT. The model has four stages with progressive token pruning layers. GAP stands for global average pooling, a layer employed to reduce the spatial dimensions of the feature map to a single vector, facilitating subsequent classification tasks.

blocks are designed to process the input data in the frequency domain. The SD-GF block focuses on spatial relations, allowing the network to interpret better and compress image information based on frequency components, while the CD-GF block operates along the channel dimension, enhancing channel-wise interactions.

In the structure of the DCT layer, the layer normalization, SD-GF, CD-GF, and MLP blocks are serially connected, as shown in Fig. 2(a). The residual connection is conducted at the end of each block. It should be noted that except for the red-shaded box from the original ViT block, all blue-shaded areas are our innovative proposals. 1D DCT block in SD-GF block is converted to the frequency domain by performing 1D DCT on a given input $x \in \mathbf{R}^{N\times D}$ in the spatial direction with N tokens. In addition, element-wise multiplication is conducted using the learnable filter K. Following this, 1D IDCT is applied again in the direction of the spatial tokens, facilitating information mixing among different tokens. Therefore, the model can capture the frequency-domain features of the tokens. The CD-GF block executes operations similar to those in the SD-GF block. However, in the CD-GF block, 1D-DCT and 1D-IDCT are applied along the channel dimension($1 \times D$) instead of the token direction($N \times 1$) in the SD-GF block. Because of the MLP block at the end of the DCT layer, it is possible to mix the information between different channels. However, by adding CD-GF blocks, the model enjoys the $O(NlogN)$ complexity of the GF block and captures the frequency-domain feature of the channel. By adding a CD-GF block, the performance improved slightly, with a slight increase in the computation of the model.

We used 1D DCT rather than 2D DCT in the SD-GF and CD-GF blocks for a more flexible token pruning ratio. If 2D DCT is used, the dimensions of the feature maps are $W \times H \times D$; therefore, it is necessary to maintain the size of $W' \times H' \times D$ to perform IDCT. Therefore, it is impossible to remove high-frequency tokens individually as desired. Even if the smallest unit is to be removed, $W + H - 1$ tokens must be removed individually to maintain the dimensions of $(W - 1) \times (H - 1) \times D$. However, when using 1D DCT, it is possible to prune the high-frequency tokens more flexibly because it is only necessary to maintain the dimensions of $N' \times D$ as in the existing transformer models, and the difference in accuracy between adopting 1D DCT and 2D DCT is minimal. Therefore, we adopted 1D DCT instead of the other 2D transforms.

Because we aim to propose a new pruning method for high-frequency tokens comparable to unnecessary spatial token pruning, high-frequency token pruning is performed only after 1D DCT on the SD-GF blocks. When 1D DCT is performed, the latter high-frequency token can be selected and removed without any effort because the low-frequency to high-frequency tokens are sorted without the need for a prediction module or class token to distinguish between the important and unimportant tokens. Therefore, if high-frequency token pruning is performed according to the keeping ratio ($\rho$) in the SD-GF blocks, a feature map of $N \times D$ size can easily be compressed to $(N \times \rho) \times D$. $(0 < \rho <= 1)$

### C. DC-TRANSFORMER BLOCK

The existing transformer block of ViT was subjected to MLP operations after MSA operations, but the DC-transformer layer of DCT-ViT has SD-GF and CD-GF blocks added in front of the transformer block, as shown in Fig. 2(b). It should be noted that except for the red-shaded and green-shaded boxes from the original ViT block, all blue-shaded areas are our innovative proposals. If the existing transformer block is used instead of the DC-transformer block in DCT-ViT, the SD-GF and CD-GF blocks of the 2nd, 3rd, and 4th stages may face challenges in delivering the frequency information interpreted to the rear stage. The DCT block aims to transform spatial image data into the frequency domain, facilitating efficient information compression and feature extraction. Conversely, the DC-Transformer block integrates this frequency-domain data back into the transformer architecture, enabling enhanced representation learning while leveraging the reduced computational complexity. Thus, DCT-ViT can better capture frequency information while achieving a small computational cost of SD-GF and CD-GF blocks.

### D. OVERALL ARCHITECTURE

The overall architecture of the proposed DCT-ViT is shown in Fig. 3. The architecture is segmented into four stages, drawing inspiration from the hierarchical nature of models such as ResNet [1], PVT [13], HVT [14], and token-pruned vision transformer models [15], [16]. Each stage is designed to progressively refine features at different scales, similar to hierarchical ViT models. We constructed one DCT layer and two DC-transformer layers per stage, which were based on 12 layers, and each stage was constructed using the same blocks. Although DCT blocks exist in all four stages, high-frequency token pruning is not applied in the first stage but only in the second, third, and fourth stages. This is because, in the case of the token-pruned ViT model, the number of tokens is reduced from the initial layer to a higher rate, resulting in a larger accuracy drop than in the case with the same computation [16].

## IV. EXPERIMENTAL RESULTS

In this section, we present various experiments that demonstrate the excellence of DCT-ViT. First, we present the results of the ImageNet-1k [30] classification task experiment by applying our method to the DeiT-small and DeiT-tiny models, which have a similar amount of computation as CNN models [1], which have been widely used recently. It was then compared with various architectures that allow ViT to perform efficiently. We also compared the 1D DCT used to interpret the tokens in the frequency domain with other transforms. Finally, each layer provided a visualization to intuitively understand the characteristics of SD-GF and CD-GF. We followed the training strategy in [10]. We trained our models for 300 epochs using AdamW [51] and did not use add-on techniques, such as knowledge distillation. However, we used a batch size of 2048, as in [16] for our single machine with four Nvidia V100 GPUs. The detailed hyper-parameters are listed in Table 1. It should be noted that all computational costs were measured, including the number of calculations required for additional DCT operations. All codes and real-time demo videos related to this study are available at: https://github.com/JHLEE17/DCT-ViT.

**TABLE 1. Hyper-parameters required for training in the proposed method. This provides comprehensive details on training configurations.**

| Method | DCT-ViT |
|---|---|
| **Epochs** | 300 |
| **Batch size** | 2048 |
| **Optimizer** | AdamW |
| **Learning rate** | 1e-3 |
| **Learning rate decay** | cosine |
| **Weight decay** | 0.05 |
| **Warmup epochs** | 5 |
| **Warmup learning rate** | 1e-6 |
| **Label smoothing $\epsilon$** | 0.1 |
| **Dropout** | × |
| **Stoch. Depth** | 0.1 |
| **Repeated Aug.** | ✓ |
| **Gradient Clip** | × |
| **Rand Augment** | 9/0.5 |
| **Mixup prob.** | 0.8 |
| **Cutmix prob.** | 1.0 |
| **Erasing prob.** | 0.25 |

### A. MAIN RESULTS

We present comprehensive results applying DCT token pruning to DeiT-Small and DeiT-Tiny baseline models on ImageNet-1K image classification. The proposed DCT-ViT models have four stages, with progressive token pruning ratios of $[1, \rho, \rho^2, \rho^3]$ in stages 2-4 based on the keeping ratio hyperparameter $\rho$. Table 2 shows Top-1 and Top-5 accuracy, multiply-accumulate operations (MACs), and number of parameters for DCT-ViT-Small and DCT-ViT-Tiny with various keeping ratios from 50% to 100%. With no pruning ($\rho = 100\%$), DCT-ViT-Small attained a slightly higher accuracy than DeiT-Small, demonstrating the benefits of the proposed DCT blocks. Pruning to $\rho = 70\%$ reduces MACs by 44% in DCT-ViT-Small, with only a 0.72% drop in Top-1 accuracy. This demonstrates very strong efficiency gains with a minimal impact on accuracy. For DCT-ViT-Tiny, $\rho = 70\%$ decreased MACs by 47.8%, whereas the Top-1 accuracy increased by 0.17%. This demonstrates the effectiveness of the method in tiny models. Higher pruning ratios, such as $\rho = 50\%$, also induced small drops in accuracy, indicating a trade-off between efficiency and accuracy. The parameter counts also decreased with more aggressive pruning.

In summary, the extensive results on ImageNet substantiate that DCT-based token pruning provides superior accuracy/efficiency trade-offs compared to DeiT baselines. This

**TABLE 2. Main results on ImageNet-1k classification task with Top-1/Top-5 accuracy, MACs, and parameters of the models according to the range of keeping ratios.**

| Model | Keeping Ratio(%) | Top-1 Acc.(%) | Top-5 Acc.(%) | MACs(G) | Param.(M) |
|---|---|---|---|---|---|
| DeiT-Tiny | | 72.20 | 91.10 | 1.3 | 5.0 |
| DCT-ViT-Tiny | 100 | 73.27 (+1.07) | 91.59 (+0.49) | 1.1 (-16%) | 6.0 |
| | 90 | 72.91 (+0.71) | 91.49 (+0.39) | 0.94 (-28%) | 5.9 |
| | 80 | 72.71 (+0.51) | 91.22 (+0.12) | 0.8 (-38%) | 5.8 |
| | 70 | 72.37 (+0.17) | 90.91 (-0.19) | 0.69 (-47%) | 5.7 |
| | 60 | 71.86 (-0.14) | 90.61 (-0.49) | 0.59 (-54%) | 5.6 |
| | 50 | **71.53 (-0.67)** | **90.43 (-0.67)** | **0.51 (-60%)** | 5.5 |
| DeiT-Small | | 79.80 | 95.00 | 4.6 | 22.0 |
| DCT-ViT-Small | 100 | 80.24 (+0.44) | 94.94 (-0.06) | 4.04 (-12%) | 21.5 |
| | 90 | 79.96 (+0.18) | 94.69 (-0.31) | 3.47 (-25%) | 21.3 |
| | 80 | 79.55 (-0.25) | 94.61 (-0.39) | 2.97 (-36%) | 21.0 |
| | 70 | **79.08 (-0.72)** | **94.30 (-0.7)** | **2.55 (-44%)** | 20.8 |
| | 60 | 78.51 (-1.29) | 94.04 (-0.96) | 2.2 (-52%) | 20.7 |
| | 50 | 77.92 (-1.88) | 93.50 (-1.5) | 1.9 (-59%) | 20.5 |

approach maintains accuracy even with high pruning rates, demonstrating its viability for compressing ViT models. Moreover, because DeiT has the same structure as the original vision transformer (*i.e.*, vanilla ViT [9]), the demonstrated efficacy of DeiT shows that the approach can be generalized well to many other ViT variants.

### B. COMPARISON WITH OTHER METHODS

We compared our method with SOTA-level efficient vision transformer methods, such as ViT pruning, network architecture search (NAS), and hierarchical architecture design in the ImageNet-1k classification task. The comparison targets include DynamicViT [15] and EViT [16], which use token pruning methods, such as DCT-ViT models, but only in the spatial domain. In addition, the comparison also included NAS techniques, such as Width&Depth pruning [18], ViT-slimming [17], and $S^2$ViTE [27], which prune some of the MSA, MLP, and layer depths, and ViT models, such as PVT [13] and HVT [14], designed hierarchically. The results are listed in Table 3, in which models with similar computational costs are grouped. Here, we can see that DCT-ViT has the best accuracy/computational cost trade-off relationship at all levels. In particular, DCT-ViT can prune significantly small models, such as DeiT-Tiny, without a drop in accuracy. Moreover, because DCT-ViT statically eliminates frequency tokens, it can be expected to produce better results if combined with additional learning methods that consider MSA, MLP, and depth comprehensively to find a better model or eliminate unimportant spatial tokens.

### C. ABLATION STUDIES

#### 1) EXPLORING THE VERSATILITY AND ROBUSTNESS OF DCT-ViT ACROSS DIVERSE DATASETS

In the field of image classification tasks, the ImageNet dataset is predominantly used as the benchmark. However, we have conducted experiments on both the ImageNet-A [52] and ImageNet-v2 [53] datasets to confirm the compatibility of our proposed method on various datasets. ImageNet-A and ImageNet-v2 are datasets developed with specific purposes relative to the original ImageNet dataset. ImageNet-A is designed to evaluate the robustness and error tendencies of AI systems, consisting of images that are typically misclassified by existing ImageNet models (*e.g.*, ResNet-50 [1]), thereby highlighting the limitations of AI in handling challenging or confusing cases. On the other hand, ImageNet-v2 aims to explore the reproducibility issues of the original ImageNet, containing images collected in a similar but slightly varied manner, focusing on assessing the overfitting of models to specific patterns in the original dataset and understanding their generalization capabilities. As detailed in Table 4, the experimental results underscore the superior top-1 accuracy of our proposed DCT-ViT model across both datasets, outperforming the competing models in terms of efficiency and precision. Despite operating at a lower computational cost and utilizing fewer parameters, the DCT-ViT model demonstrates an enhanced ability to maintain, and in several instances exceed, the benchmark accuracy standards set by previous models. This high accuracy balance with reduced resource demands signifies a substantial advancement in the application of vision transformers, particularly in resource-constrained environments. It should be noted that when testing ResNet-50 with ImageNet-A, which is a collection of images that ResNet-50 misclassified, it is reasonable for ResNet-50 to have an accuracy of 0.0 on ImageNet-A.

#### 2) EVALUATION ON SPATIAL TO FREQUENCY DOMAIN TRANSFORMS

In this study, we propose SD-GF and CD-GF blocks to reduce computations while minimizing the loss of accuracy by interpreting frequency information well and flexibly removing high-frequency tokens. Here, 1D DCT and 1D IDCT were used to interpret information from the spatial

**TABLE 3.** Comparison with other efficient vision transformer models in ImageNet-1k classification. The naming schemes such as '0.8' and '0.7' in DCT-ViTs denote the keeping ratio, indicating different levels of model compression. 'Small' and 'Tiny' indicate the type of baseline DeiT models.

| Model | Top-1 Acc.(%) | Top-1 ↓(%) | MACs(G) | MACs ↓ | Method |
|---|---|---|---|---|---|
| DeiT-Small | 79.8 | | 4.60 | | - |
| W&D pruning ViT-S-0.3-12 | 78.55 | -1.25 | 3.10 | -33% | MSA, MLP, depth pruning |
| GFNet-XS | 78.6 | -1.2 | 2.90 | -37% | Learning in Frequency Domain without MSA |
| $S^2ViTE-S$ | 79.2 | -0.6 | 3.1 | -33% | MSA, MLP pruning |
| ViT-slim_ps-0.6 | 79.2 | -0.6 | 2.80 | -39% | MSA, MLP pruning |
| DyViT-S-0.7 | 79.3 | -0.5 | 3.00 | -35% | Token pruning |
| EViT-S-0.7 | 79.5 | -0.3 | 3.00 | -35% | Token pruning |
| **DCT-ViT-Small-0.8 (ours)** | **79.55** | **-0.25** | **2.97** | **-36%** | **Token pruning in Frequency Domain** |
| $S^2ViTE+-S$ | 78.2 | -1.6 | 2.70 | -41% | MSA, MLP pruning |
| W&D pruning ViT-S-0.3-11 | 78.38 | -1.42 | 2.60 | -43% | MSA, MLP, depth pruning |
| EViT-S-0.6 | 78.9 | -0.9 | 2.60 | -43% | Token pruning |
| **DCT-ViT-Small-0.7 (ours)** | **79.08** | **-0.72** | **2.55** | **-44%** | **Token pruning in Frequency Domain** |
| DyViT-S-0.5 | 77.5 | -2.3 | 2.30 | -50% | Token pruning |
| ViT-slim_ps-0.5 | 77.94 | -1.86 | 2.30 | -50% | MSA, MLP pruning |
| HVT-S-1 | 78 | -1.8 | 2.40 | -48% | Hierarchical Arch. |
| EViT-S-0.5 | 78.5 | -1.3 | 2.30 | -50% | Token pruning |
| **DCT-ViT-Small-0.6 (ours)** | **78.51** | **-1.29** | **2.2** | **-52%** | **Token pruning in Frequency Domain** |
| PVT-Tiny | 75.1 | -4.7 | 1.90 | -59% | Hierarchical Arch. |
| PoolFormer-S12 | 77.2 | -2.6 | 2.00 | -57% | Hierarchical Arch. without MSA |
| **DCT-ViT-Small-0.5 (ours)** | **77.92** | **-1.88** | **1.9** | **-59%** | **Token pruning in Frequency Domain** |
| DeiT-Tiny | 72.2 | | 1.30 | | - |
| W&D pruning ViT-Tiny-0.3-11 | 70.34 | -1.86 | 0.70 | -46% | MSA, MLP, depth pruning |
| W&D pruning ViT-Tiny-0.3-12 | 71.1 | -1.1 | 0.90 | -31% | MSA, MLP, depth pruning |
| EViT-Tiny-0.7 | 71.82 | -0.38 | 0.82 | -37% | Token pruning |
| **DCT-ViT-Tiny-0.7 (ours)** | **72.37** | **0.17** | **0.69** | **-47%** | **Token pruning in Frequency Domain** |
| EViT-Tiny-1.0 | 72.57 | 0.37 | 1.256 | -3% | Token pruning |
| **DCT-ViT-Tiny-1.0 (ours)** | **73.27** | **1.07** | **1.1** | **-16%** | **Token pruning in Frequency Domain** |

**TABLE 4.** Top-1 accuracy comparison on ImageNet-A and ImageNet-v2 datasets.

| Model | Top-1 Acc.(%) | | MACs | Params |
|---|---|---|---|---|
| | ImageNet-A | ImageNet-v2 | (G) | (M) |
| ResNet-50 | 0.0 | 67.4 | 4.1 | 26 |
| ResNeXt50-32x4d | 10.7 | 68.2 | 4.3 | 25 |
| DeiT-S | 18.9 | 68.1 | 4.6 | 22 |
| GFNet-S | 14.3 | 68.5 | 4.5 | 25 |
| DCT-ViT-S | **19.7** | **68.5** | 4.045 | 21.5 |
| DCT-ViT-S-90 | 17.4 | 67.8 | 3.47 | 21.3 |
| DCT-ViT-S-80 | 16.6 | 67.7 | 2.97 | 21.0 |
| DCT-ViT-S-70 | 14.5 | 66.6 | 2.55 | 20.8 |

**TABLE 5.** Various spatial to frequency domain transforms in DCT-ViT-Small with same token pruning ratio.

| Transform method | Top-1(%) | MACs(G) | Param.(M) |
|---|---|---|---|
| **1D DCT** | **79.13** | **2.56** | **20.86** |
| 2D DCT | 79.04 | 2.57 | 20.86 |
| 2D DFT | 78.99 | 2.56 | 20.90 |

domain to the frequency domain. The purpose of the CD-GF block is to mix information between channels; thus, 1D transform is suitable. However, the 2D transform is also structurally available in the SD-GF block. In addition, in GFNet, the tokens were mixed using 2D DFT and 2D IDFT. Therefore, Table 5 presents the results of an experiment conducted by changing the transforms. In the case of experiments in which 1D DCT was replaced with 2D DCT, 2D DCT was used in SD-GF blocks. However, in the CD-GF block, 1D DCT was used because of the problem of matching dimensions. In another experiment, where 1D DCT was replaced with 2D DFT, 2D DFT substituted for 1D DCT in the SD-GF block, but 1D DFT was used in CF-GF blocks for the same reason as in the 2D DCT case. When 2D DCT and 2D DFT operations are used, because the dimensions of the input $x$ are $H \times W \times D$, the keeping ratio cannot be obtained continuously. Therefore, for each block in which token pruning occurs, the tokens in the last $r$ rows and columns of each channel are pruned to reduce $H \times W$ to $(H - r) \times (W - r)$. At the same time, the experiments using 1D DCT maintained the same token number as the other experiments. We set $r$ to 2 for these experiments. Table 5 shows that there is no significant difference between the accuracy and MACs for all three cases. However, in the case of using 1D DCT, unlike the 2D transform, the performance is the best, and the keeping ratio can be set continuously; thus, 1D DCT was selected as the spatial-to-frequency-domain transform in this study.

### 3) SCALABILITY: DCT BASED PRUNING ON MLP LAYERS

Reducing the parameter size is crucial for the practical deployment of the ViT model on mobile or edge devices. Pruning the MLP layer proved to be effective in reducing the parameter size. The proposed pruning method leveraging

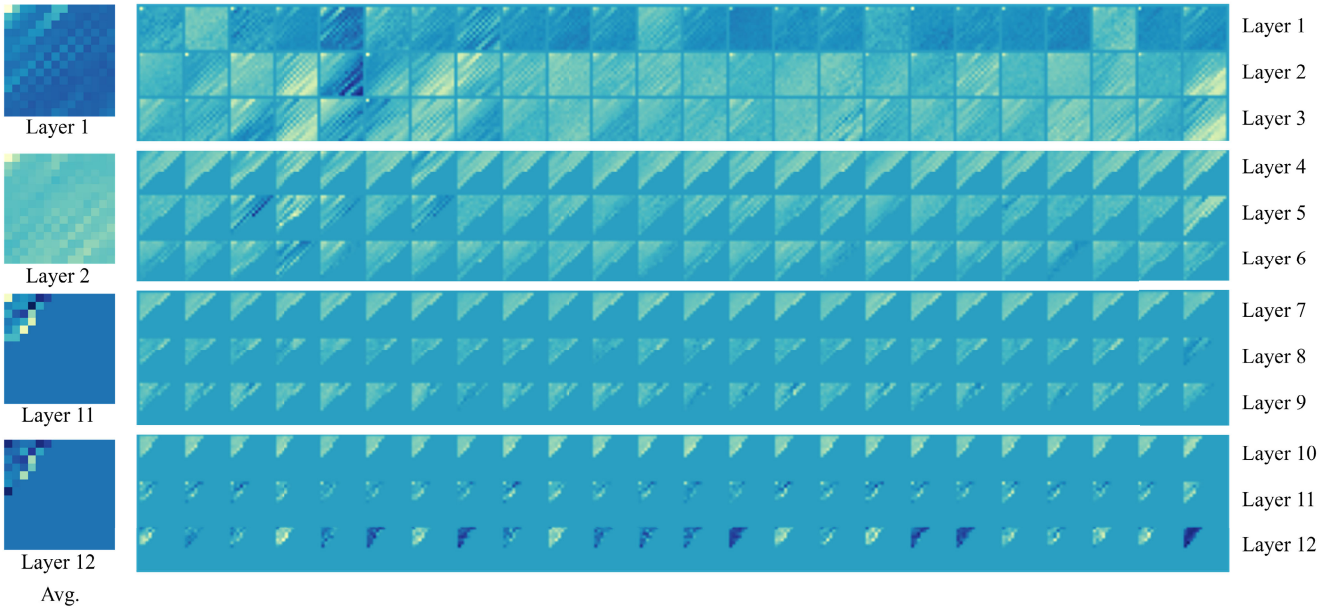


**FIGURE 4.** Visualization of the spatial global filters in DCT-ViT-Small. The filters for each layer are listed horizontally and the average of the filters for layers 1, 2, 11, and 12 is shown on the left. This provides intuition about frequency separation.

DCT can be readily applied to MLP layers as well as ViT tokens owing to its high scalability. Like the CD-GF block shown in Fig. 2, we perform 1D DCT in the channel direction before the MLP block to prune the high-frequency channels. After the MLP block, we add zero padding to the pruned channel positions to maintain the feature map size and then perform a 1D IDCT. Using this method, we can reduce the computational cost and number of parameters in the MLP layers while maintaining the final channel size. Table 6 presents the results of the MLP channel pruning on the DC-transformer blocks based on the DCT-ViT-small model. Unlike the token pruning method, which has no significant effect on parameter reduction, the additional application of MLP layer pruning can effectively reduce the number of parameters as well as MACs. Considering that the EViT-S-0.5 [16] presented in Table 3 has the same parameter size of 22M as DeiT-Small [10], the proposed method with token and MLP keeping ratios of 80% and 60%, respectively, can additionally reduce the parameters by 32% with similar accuracy and MACs.

**TABLE 6.** Additional DCT based MLP pruning results on DCT-ViT-small with various token/MLP pruning ratios. This table demonstrates DCT pruning can be applied to tokens and MLP channels.

| Model | Token Keeping Ratio(%) | MLP Keeping Ratio(%) | Top-1 Acc.(%) | MACs(G) | Param.(M) |
|---|---|---|---|---|---|
| DCT -ViT -small | 100 | - | 80.24 | 4.04 (-12%) | 21.5 (-2%) |
| | 100 | 80 | 79.92 | 3.76 (-18%) | **18.1 (-18%)** |
| | 100 | 60 | 79.27 | 3.14 (-32%) | **15.5 (-30%)** |
| | 80 | - | 79.55 | 2.97 (-36%) | 21.0 (-5%) |
| | 80 | 80 | 79.12 | 2.76 (-40%) | **17.7 (-20%)** |
| | 80 | 60 | 78.56 | 2.3 (-50%) | **15.0 (-32%)** |
| | 60 | - | 78.51 | 2.2 (-52%) | 20.7 (-6%) |
| | 60 | 80 | 78.44 | 2.04 (-56%) | **17.3 (-21%)** |
| | 60 | 60 | 77.50 | 1.7 (-63%) | **14.7 (-33%)** |

### D. VISUALIZATION

DCT-ViT operates on the premise that similar to human vision, the ViT model would still function effectively, even with the removal of some high-frequency information. We visualized the global filter parameters learned in each block to validate our hypothesis and ascertain whether the SD-GF and CD-GF blocks can adequately distinguish frequency features.

In Fig. 4, we visualize the filters of the SD-GF block, trained in the frequency domain, using the DCT-ViT-0.5 model. These filters are of dimension $N \times D$; however, for an intuitive understanding, they are reshaped to $H \times W \times D'(D' = 24 \ll D)$. Here, the zigzag algorithm is employed to represent the low-frequency components in the upper left and the high-frequency components in the lower right. Each row comprises 24 filters of the same layer, each with a size of $W \times H$. As we progress downward, each row represents a filter from a deeper layer of the model. In this instance, because the token count was halved for each of the three layers, excluding the first stage, the token count is sequentially reduced to 196, 98, 49, and finally, 25. The average filter values for each layer are shown on the left. As per our assumptions, Fig. 4 shows that the last layer successfully separates features in the low-frequency components.

## V. CONCLUSION

The ViT model has limitations in its applicability to resource-constrained mobile/edge devices due to its enormous computational cost and parameters. In this study, we propose a high-frequency token pruning technique that has not been

focused on ViT model compression, inspired by traditional image compression. DCT-ViT achieves a better trade-off between accuracy and computational costs than SOTA ViT methods by statically reducing only the token in the frequency domain while maintaining the MSA and MLP modules. The proposed method has a technical limitation in that it is difficult to realize the full latency reduction effect of ViT computations without the help of a dedicated accelerator because DCT computations are not efficiently accelerated on the GPU. However, these limitations are expected to be solved by implementing the DCT operation library on CUDA or designing various application-specific integrated circuits for ViT. We anticipate that DCT-ViT will contribute to the practical application of ViTs in mobile/edge devices. Our future work will focus on applying DCT-ViT to a broader range of ViT architectures and assessing its impact on various models. We will also explore applying this technique to object detection and segmentation tasks, considering the potential loss of spatial information.

## REFERENCES


[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[4] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1440–1448.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[7] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 502–511.

[8] S. I. Lee and H. Kim, "GaussianMask: Uncertainty-aware instance segmentation based on Gaussian modeling," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3851–3857.

[9] 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 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[10] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[11] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021, *arXiv:2111.06377*.

[13] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[14] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable vision transformers with hierarchical pooling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 367–376.

[15] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C. J. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 13937–13949.

[16] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie, "Not all patches are what you need: Expediting vision transformers via token reorganizations," 2022, *arXiv:2202.07800*.

[17] A. Chavan, Z. Shen, Z. Liu, Z. Liu, K.-T. Cheng, and E. Xing, "Vision transformer slimming: Multi-dimension searching in continuous optimization space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4931–4941.

[18] F. Yu, K. Huang, M. Wang, Y. Cheng, W. Chu, and L. Cui, "Width & depth pruning for vision transformers," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022.

[19] X. Su, S. You, J. Xie, M. Zheng, F. Wang, C. Qian, C. Zhang, X. Wang, and C. Xu, "ViTAS: Vision transformer architecture search," 2021, *arXiv:2106.13700*.

[20] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 980–993.

[21] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

[22] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. COM-100, no. 1, pp. 90–93, Jan. 1974.

[23] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991.

[24] S. Winograd, "On computing the discrete Fourier transform," *Math. Comput.*, vol. 32, no. 141, pp. 175–199, 1978.

[25] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 23296–23308.

[26] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12116–12128.

[27] T. Chen, Y. Cheng, Z. Gan, L. Yuan, L. Zhang, and Z. Wang, "Chasing sparsity in vision transformers: An end-to-end exploration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 19974–19988.

[28] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.

[29] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] N. Park and S. Kim, "How do vision transformers work?" 2022, *arXiv:2202.06709*.

[32] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12124–12134.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[34] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[35] H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware," 2018, *arXiv:1812.00332*.

[36] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," 2019, *arXiv:1908.09791*.

[37] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[38] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14420–14430.

[39] M. Zhu, Y. Tang, and K. Han, "Vision transformer pruning," 2021, *arXiv:2104.08500*.
[40] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10819–10829.
[41] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. A. Smith, and L. Kong, "Random feature attention," 2021, *arXiv:2103.02143*.
[42] X. Ma, X. Kong, S. Wang, C. Zhou, J. May, H. Ma, and L. Zettlemoyer, "Luna: Linear unified nested attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2441–2453.
[43] H. Cai, J. Li, M. Hu, C. Gan, and S. Han, "EfficientViT: Multi-scale linear attention for high-resolution dense prediction," 2022, *arXiv:2205.14756*.
[44] I. Pitas, *Digital Image Processing Algorithms and Applications*. Hoboken, NJ, USA: Wiley, 2000.
[45] A. M. Raid, W. M. Khedr, M. A. El-Dosuky, and W. Ahmed, "JPEG image compression using discrete cosine transform—A survey," 2014, *arXiv:1405.6147*.
[46] C. Fogg, D. J. LeGall, J. L. Mitchell, and W. B. Pennebaker, *MPEG Video Compression Standard*. Norwell, MA, USA: Kluwer, 2007.
[47] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, and F. Ren, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1740–1749.
[48] C. Scribano, G. Franchini, M. Prato, and M. Bertogna, "DCT-former: Efficient self-attention with discrete cosine transform," 2022, *arXiv:2203.01178*.
[49] Z. Liu, J. Xu, X. Peng, and R. Xiong, "Frequency-domain dynamic pruning for convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.
[50] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," 2021, *arXiv:2105.03824*.
[51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[52] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15262–15271.
[53] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5389–5400.

**JONGHO LEE** (Member, IEEE) received the B.S. and M.S. degrees in electrical and information engineering from Seoul National University of Science and Technology, Seoul, South Korea, in 2021 and 2023, respectively. In 2020, he was a Research Student with Korea Institute of Science and Technology (KIST), Seoul. In 2022, he was a Visiting Student with the University of Wisconsin–Madison, Madison, WI, USA. In 2023, he was a Visiting Student with the German Research Centre for Artificial Intelligence (DFKI), Kaiserslautern, Germany. His research interests include deep learning and machine-learning algorithms for computer vision tasks.

**HYUN KIM** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2009, 2011, and 2015, respectively. From 2015 to 2018, he was with the BK21 Creative Research Engineer Development for IT, Seoul National University, Seoul, as a BK Assistant Professor. In 2018, he joined with the Department of Electrical and Information Engineering, Seoul National University of Science and Technology, Seoul, where he is currently an Associate Professor. His research interests include algorithms, computer architecture, memory, and SoC design for low-complexity multimedia applications and deep neural networks.