This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SpectFormer: Frequency and Attention is what you need in a Vision Transformer

Badri N. Patro
Microsoft
badripatro@microsoft.com

Vinay P. Namboodiri
University of Bath
vpn22@bath.ac.uk

Vijay S. Agneeswaran
Microsoft
vagneeswaran@microsoft.com

Abstract

Vision transformers have been applied successfully for image recognition tasks. There have been either multi-headed self-attention based (ViT [12], DeIT [54]) similar to the original work in textual models or more recently based on spectral layers (Fnet [29], GFNet [46], AFNO [15]). We hypothesize that spectral layers capture high-frequency information such as lines and edges, while attention layers capture token interactions. We investigate this hypothesis through this work and observe that indeed mixing spectral and multi-headed attention layers provides a better transformer architecture. We thus propose the novel Spectformer architecture for vision transformers that has initial spectral and deeper multi-headed attention layers. We believe that the resulting representation allows the transformer to capture the feature representation appropriately and it yields improved performance over other transformer representations. For instance, it improves the top-1 accuracy by 2% on ImageNet compared to both GFNet-H and LiT. SpectFormer-H-S reaches 84.25% top-1 accuracy on ImageNet-1K (state of the art for small version). Further, Spectformer-H-L achieves 85.7% which is the state of the art for the comparable base version of the transformers. We further validated the SpectFormer performance in other scenarios such as transfer learning on standard datasets such as CIFAR-10, CIFAR-100, Oxford-IIIT-flower, and Standford Car datasets. We then investigate its use in downstream tasks such as object detection and instance segmentation on the MS-COCO dataset and observe that Spectformer shows consistent performance that is comparable to the best backbones and can be further optimized and improved. The source code is available on this website https://github.com/badripatro/SpectFormers.

1. Introduction

Transformers originated in natural language processing with the seminal work by Vaswani *et al.* [59]. Transformers have gone on to revolutionize the language domain in the form of large language models such as GPT-4o [2], Gemini [51], Claude [52], Phi 3 [1] and Llama 3 [57]. Parallel work extended the concept of Transformer to computer vision and other domains. Interestingly, the main ideas explored retain much of the original transformer architecture. Clearly, the different domains could benefit from adapted transformers that are particular to the specific task. Through this work, we aim to specifically analyse the transformer for the image classification task using a vision transformer. We show that the proposed adaptation, Spectformer, can outperform the state-of-the-art for this task.

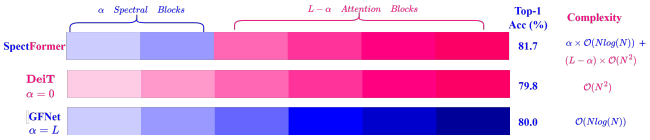

Figure 1. SpectFormer architecture consist of L Blocks, out of which α spectral blocks and $L-\alpha$ attention blocks. $\alpha = 0$ means all self-attention blocks such as DeiT (ViT) and $\alpha = L$ means all Spectral blocks such as GFNet

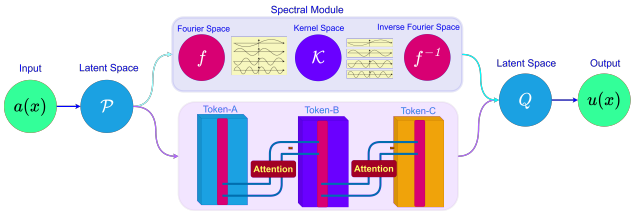

Figure 2. The SpectFormer architecture comprises a Spectral Block responsible for encoding the periodic characteristics of the input signal through Fourier kernel space. Simultaneously, the attention block facilitates the representation of the non-periodic aspects by enabling information flow between different tokens.

The adaptation of transformers for computer vision was first explored in the Vision Transformer (ViT) [12]. They made the important contribution of developing appropriate patch-based tokenization for images whereby the transformer architecture could be used for images. DeIT [55] further improved the training process. This study aims to address this gap by optimizing transformers for image classification using a vision transformer as shown in figure-1 and 2. The Fourier domain plays a major role in extracting frequency-based analysis of image information and

has been well-studied by the community. This is further supported by seminal work by Hubel and Weisel [25] that showed frequency-tuned simple cells in the visual cortex. In transformers, it has been shown that the Fourier transforms could replace the multi-headed attention layers and achieve similar performance by Rao *et al.* [46] where they presented GFNet. They suggested that this approach captures fine-grained properties of images. This approach was further extended by AFNO [15], where they treated token mixing as operator learning. We hypothesize that for the image domain, both spectral and multi-headed self-attention play an important role. We introduce SpectFormer, a novel architecture that combines spectral and multi-headed attention layers to enhance performance.

There have also been several hierarchical transformer architectures that have been explored in the literature [7, 38, 61]. One of the hierarchical approaches has been LiT [42] that uses less self-attention in the early layers, by using pure MLP (Multi-layer Perceptron) layers. They do use self-attention in deeper layers to capture longer dependencies. Motivated by the works related to spectral and also hierarchical transformers, we developed SpectFormer, a new transformer architecture that uses spectral layers implemented with Fourier Transform to capture relevant features in the initial layers of the architecture. Further, we use multi-headed self-attention in the deeper layers of the network. The SpectFormer architecture is simple and transforms the image tokens to the Fourier domain, then applies gating techniques using learnable weight parameters, and finally does the inverse Fourier transform to get the signal back.

We analyzed and observed that having all spectral layers in the transformer architecture results in reduced complexity to O(Nlog(N)). However, all spectral layered transformers such as GFNet [46] or AFNO [15] have a performance gap compared to state-of-art transformer models such as Volo [73], WavViT [70], Flatten Transformer [18], MixFormer [8], SVT [43], Biformer [78], STViT [24], and FDVIT [68] etc. In contrast, all attention-layered transformers such as DeIT have complexity as $O(N^2)$ and may not perform well for longer input token sizes. This motivates us to propose SpectFormer which combines the spectral layers which capture periodic information and multi-headed self-attention to capture aperiodic information as shown in the figure. Initial spectral layers will help SpectFormer to capture lines and edges, while the later attention layers capture long-range token dependencies. Our approach combines both spectral and multi-headed attention as shown in figure-2. As shown in the figure, SpectFormer can give rise to a spectrum of transformers, starting with sub-quadratic attention-free transformers (with all spectral layers) to regular quadratic attention-based transformers having complexity $O(N^2)$. We also validate the SpectFormer architecture by visualization of the learned filters and find that the filters for the spectral layers are more localized as compared to similar fully spectral GFNet [46]. The evidence suggests that adopting mixed spectral and later multi-headed attention results in improved results. It must be noted that during the training phase of SpectFormer, the deeper attention layers enable the initial spectral layers to learn the edges and lines of the image features, whereas in GFNet, since deeper attention layers are missing, the learned filters are not sharp enough in the initial layers.

We outline our contributions below:

- **Spectral Layer**: SpectFormer is crafted with initial spectral layers and deeper layers employing multi-headed attention. The spectral layer incorporates a Spectral Gating Network utilizing a learnable Fourier transform. Empirical studies, including filter visualization, demonstrate the effectiveness of this design, showcasing more localized learned filters compared to fully spectral GFNeT.

- **Performance Comparison:** SpectFormer is systematically compared against various transformer architectures, including LiT, DeIT, GFNet, AFNO, PVT, Swin, FDViT, Flatten Transformer, MixFormer, STViT and Bi-Former on the ImageNet dataset. The results reveal that SpectFormer outperforms these models, establishing state-of-the-art results in ImageNet 1K classification.

- **Transfer Learning Capability:** SpectFormer demonstrates robust performance in transfer learning scenarios, particularly when trained on ImageNet and tested on CIFAR datasets (CIFAR-10 and CIFAR-100). The model achieves performance levels comparable to other transformers in this transfer learning mode.

- **Versatility in Tasks:** SpectFormer exhibits consistent and competitive performance beyond image classification. Evaluation on the MS COCO dataset demonstrates its effectiveness in tasks such as object detection and instance segmentation, showcasing its versatility across diverse computer vision tasks.

2. Related Work

The Vision Transformer (ViT) [12] was the first effort to adapt transformers to the vision domain. Touvron et al. [54] proposed an efficient transformer model based on distillation technique (DeiT). The vanilla transformer architecture which uses multi-headed self-attention (MSA) for efficient token mixing includes papers such as Tokens-to-token ViT [72], Transformer iN Transformer (TNT) [19], CrossViT [5], Class attention image Transformer(CaiT) [56],

Bidirectional Encoder [3] etc . The architectural complexity of most of the above transformers is O(N^2). Attempts at alleviating this include the Uniformer [30] which brings the best of convolutional nets and transformers by using multi-headed relation aggregation and RegionViT [4] as well as Token Pyramid Vision Transformer (TopFormer) [76]. The complexity has also been mitigated by using spectral transformers, which typically have O($N \log N$) complexity.

Spectral Transformers Inspired by an MLP-mixer-based token mixing technique, recent work uses a spectral mixing technique in which the self-attention layer of the transformer is replaced by a non-parameterized Fourier transformation (Fnet) [29], which is then followed by a non-linearity and feed-forward network. This was followed by the Global Filter network (GFNet) [46], which uses a depth-wise global convolution for token mixing. Guibias et al. [15] formulated the token mixing task as an operator-learning task that learns mapping among continuous functions in infinite dimensional space using Fourier Neural Operator (FNO) [33]. In Wave-ViT [70], the author has proposed a wavelet vision transformer to perform lossless down-sampling using wavelet transform over keys and values of the self-attention network. Recently, another work [39] proposes a Fourier integral theorem to characterize attention as non-parametric kernel regression and approximate key-query distributions. We compare the performance of the above methods with the proposed SpectFormer model.

3. Intuition: Conceptual basis for SpectFormer

SpectFormer integrates Fourier neural operators [33], specifically spectral layers, with multi-headed attention blocks to optimize the transformation of periodic and aperiodic signals. The spectral layer operates in the Fourier Spectrum domain, facilitating effective learning of linear transformations for periodic signals. It includes a bias term for aperiodic signal transformation. Multi-headed attention, acting as an independent transformation within a residual stream [14], enables efficient information processing. The value vector, which is derived from the previous residual block, combines with attention vectors, contributing to the output token for each head. This dual approach manages information through query-key and value-output vector subspace transformations. By alternating between spectral layers and multi-headed attention blocks, SpectFormer achieves enhanced performance by effectively learning optimal transformations for both periodic and aperiodic signals.

3.1. Empirical evidence for mixed Spectral Transformer:

To assess the impact on performance based on representation, we conducted an empirical study comparing all-attention, all-spectral, and mixed spectral-attention layers. The arrangement of spectral layers, whether they are the initial or later layers (referred to as inverse SpectFormer), was explored. The results in Table 4 indicate that the configuration with initial spectral layers followed by multi-headed attention layers is particularly advantageous. This empirical evidence highlights the adaptive nature of SpectFormer, leading us to propose an architecture that incorporates initial spectral layers followed by deeper multi-headed attention layers. Our multi-headed self-attention layer is similar to the original attention paper [59]. We show that SpectFormer achieves state-of-art performance compared to parallel architectures like LiT and outperforms complete spectral architectures like GFNet [46] and AFNO [15]. It also outperforms complete multi-headed attention-based transformers like DeIT on the ImageNet 1K dataset.

4. Method: SpectFormer

Our SpectFormer architecture combines the best features of the spectral block using Spectral Gating Network (SGN) and attention blocks as illustrated in figure- 3.

4.1. Spectral Block: Spectral Gating Network

For a given input Image $\mathbf{X}$ and its corresponding frequency domain conversion $\mathcal{X}$, the gating operations of frequency-domain on $\mathcal{X}$ can be equivalently represented as global convolutions on $\mathbf{X}$ in the spatial domain. This equivalence is expressed as follows:

$$\mathcal{X}\mathcal{W} + \mathcal{B} = \mathcal{F}(\mathbf{X} * W + B) \tag{1}$$

where $*$ denotes circular convolution, $\mathcal{W}$ and $\mathcal{B}$ represent the complex number weight and bias in the frequency domain, while W and B denote the weight and bias in the spatial domain, and $\mathcal{F}$ signifies the Discrete Fourier Transform (DFT). The output DFT operation is a complex number value $\mathcal{X} \in \mathbb{R}^{N \times D}$, and combined with a complex number weight matrix $\mathcal{W} \in \mathbb{R}^{D \times D}$ and a complex number bias $\mathcal{B} \in \mathbb{R}^{D}$ using the Spectral Gating Network (SGN). The SGN is expressed by the following formulation:

$$h^{\ell} = \sigma(h^{\ell-1}\mathcal{W}^{\ell} + \mathcal{B}^{\ell}), h^{0} = \mathcal{X} \tag{2}$$

Here, $h^{\ell} \in \mathbb{R}^{N \times D}$ represents the final output, ℓ denotes the ℓ-th layer, and σ is the activation function. Considering both $\mathcal{X}$ and $\mathcal{W}$ as complex numbers, we extend Equation (2) by employing the multiplication rule for complex numbers. The extended formulation is as follows:

$$\begin{aligned} Re(h)^{\ell} &= Re(h^{\ell-1})\mathcal{W}_r^{\ell} - Im(h^{\ell-1})\mathcal{W}_i^{\ell} + \mathcal{B}_r^{\ell} \\ Im(h)^{\ell} &= Re(h^{\ell-1})\mathcal{W}_i^{\ell}) + Im(h^{\ell-1}\mathcal{W}_r^{\ell} + \mathcal{B}_i^{\ell} \\ h^{\ell} &= \sigma(Re(h)^{\ell}) + j\sigma(Im(h)^{\ell}) \end{aligned} \tag{3}$$

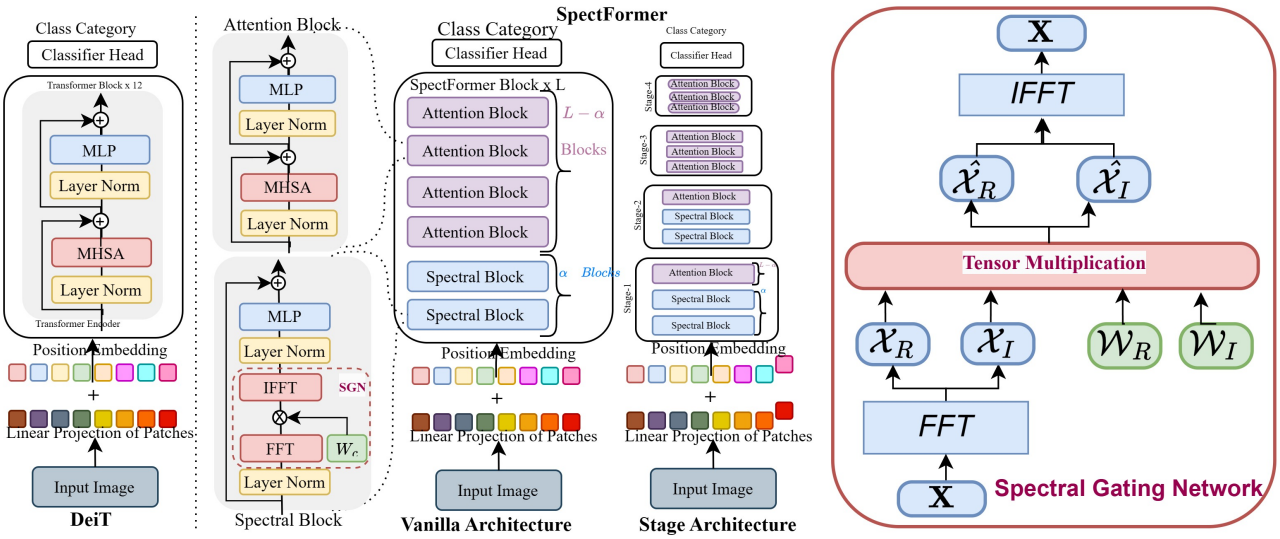

Figure 3. This figure shows Architectural details of SpectFormer. The first part shows the DeiT [54] architecture. The second part shows the vanilla and Stage architecture of the SpectFormer Model. This also shows the layer structure of Spectral and Attention Blocks.

Here, $\mathcal{W}^\ell = \mathcal{W}_r^\ell + j\mathcal{W}_i^\ell$ and $\mathcal{B}^\ell = \mathcal{B}_r^\ell + j\mathcal{B}_i^\ell$ represent the real and imaginary parts, respectively. The gating network in the frequency domain, denoted as SGN, is computing the real and imaginary parts of frequency components separately. Subsequently, these parts are stacked to form a complex number, with the final result obtained by concatenating the real and imaginary parts. Finally, we operate Inverse DFT transformation to bring back to the spatial domain from the frequency domain.

Hence, the operations of SGN, denoted as $\mathcal{X}\mathcal{W} + \mathcal{B}$, are tantamount to the operations $(\mathbf{X} * W + B)$ in the spatial domain. This signifies that the gating operations of the frequency domain can be interpreted as global convolutions in the spatial domain.

4.2. Spectral Block: SGN as Convolution Kernel

The objective is to learn a mapping between infinite-dimensional spaces from a finite set of observed input-output pairs $\{a_j, u_j\}_{j=1}^N$, where a function $\mathcal{F}$ maps from space $\mathcal{A}$ to space $\mathcal{U}$. This can be expressed as $\mathcal{F} : \mathcal{A} \times \Theta \to \mathcal{U}$, or equivalently $\mathcal{F}_\theta : \mathcal{A} \to \mathcal{U}$, where $\theta \in \Theta$.

The spectral layer of the SpectFormer, a specialized case of Fourier neural operators, transforms real inputs to real outputs using the Spectral Layer. The input $a \in \mathcal{A}$ undergoes projection to a higher-dimensional representation $v_0(x) = P(a(x))$ through a local transformation P, typically parameterized by a shallow fully-connected neural network. Multiple iterations of updates $v_t \mapsto v_{t+1}$ are then applied. The final output $u(x) = Q(v_T(x))$ results from the projection of v_T using the local transformation $Q : \mathcal{R}^{d_v} \to \mathcal{R}^{d_u}$. Each iteration's update $v_t \mapsto v_{t+1}$ involves a non-local integral operator $\mathcal{K}$ and a local, nonlinear activation function σ. The iterative update is defined as:

$$v_t(x) \mapsto v_{t+1}(x) = \sigma\Big(\big(\mathcal{K}(a;\phi)v_t\big)(x)\Big), \qquad \forall x \in D \tag{4}$$

The Kernel integral operator $\mathcal{K}$ is defined by

$$\big(\mathcal{K}(a;\phi)v_t\big)(x) := \int_D \kappa\big(x, y, a(x), a(y); \phi\big)v_t(y)\mathrm{d}y, \tag{5}$$

Where κ_ϕ is a neural network parameterized by $\phi \in \Theta_\mathcal{K}$. Here κ_ϕ plays the role of a kernel function which we learn from data. Now the kernel integral operator is replaced by a convolution operator defined in Fourier space. Let $\mathcal{F}$ denote the Fourier transform of a function $f : D \to \mathcal{R}^{d_v}$ and $\mathcal{F}^{-1}$ its inverse then

$$\mathcal{F}_j(k) = \int_D f_j(x)e^{-2i\pi\langle x,k\rangle}\mathrm{d}x, \mathcal{F}_j^{-1}(x) = \int_D f_j(k)e^{2i\pi\langle x,k\rangle}\mathrm{d}k$$

For the special case of the Green's kernel $\mathcal{K}(s,t) = \mathcal{K}(s-t)$, the integral leads to global convolution. By applying $\kappa_\phi(x, y, a(x), a(y)) = \kappa_\phi(x - y)$ in (5) and applying the convolution theorem, we find that

$$\big(\mathcal{K}(a;\phi)v_t\big)(x) = \mathcal{F}^{-1}\big(\mathcal{F}(\kappa_\phi) \cdot \mathcal{F}(v_t)\big)(x), \qquad \forall x \in D.$$

The Fourier integral kernel $\mathcal{K}$ is defined as

$$\big(\mathcal{K}(\phi)v_t\big)(x) = \mathcal{F}^{-1}\Big(R_\phi \cdot (\mathcal{F}v_t)\Big)(x) \quad \forall x \in D \tag{6}$$

where R_ϕ is the Fourier transform of a periodic function $\kappa : \bar{D} \to \mathcal{R}^{d_v \times d_v}$ parameterized by $\phi \in \Theta_\mathcal{K}$.

4.3. Attention Block

The attention heads of the Transformer model consist of the Query-Key (QK) [14] circuit and the Output-Value (OV) circuit. The Query-Key (QK) circuit plays a crucial role in

computing attention scores between query and key tokens. The representation of each token in the attention head is derived by applying weight matrices $\mathbf{W}_Q$ and $\mathbf{W}_K$ to the input sequence $\mathbf{X}$, resulting in query vectors $q_i = \mathbf{W}_Q x_i$ and key vectors $k_i = \mathbf{W}_K x_i$. Attention scores are then calculated using the softmax operation on the dot product of these vectors, represented as $\mathbf{A} = \text{softmax}(\mathbf{q}^T \cdot \mathbf{k}) = \text{softmax}(\mathbf{x}^T \cdot \mathbf{W}_Q^T \cdot \mathbf{W}_K \cdot \mathbf{x})$. On the other hand, the Output-Value (OV) circuit determines how each token's information influences output logits. Value vectors $\mathbf{v}_i = \mathbf{x}_i \cdot \mathbf{W}_V$ are obtained by applying the weight matrix $\mathbf{W}_V$ to the input sequence. The final token representation, denoted as

$$\mathcal{R}_i = \sum_j \mathbf{A}_{i,j} \cdot \mathbf{v}_j \tag{7}$$

This is computed by linearly combining value vectors based on the attention pattern, as given by Equation 7. The essence of attention heads lies in extracting pertinent information from the residual stream of a designated token and transmitting it to the residual stream connected with another token. This involves two steps: extracting information in the first step and transmitting it in the second step, fostering information flow within the model.

4.4. SpectFormer Block

The SpectFormer block unifies two essential components: the attention representation, $\mathcal{R}$ (from Equation (7), and the spectral representation, $\mathcal{V}$ (from Equation (4). This integration is expressed as follows:

$$v_{t+1}(x) := \mathcal{R}_t + \sigma\Big(\big(\mathcal{K}(a;\phi)v_t\big)(x)\Big), \quad \forall x \in D$$

Here, $\mathcal{R}_t$ represents the attention accumulation from Equation (7), and the spectral transformation, facilitated by the kernel function $\mathcal{K}$ parameterized by a and ϕ, acts on the previous representation v_t. SpectFormer adeptly handles both periodic and aperiodic signals, surpassing limitations observed in GFNet. While GFNet focuses on transforming periodic signals in the Fourier spectrum domain and includes a bias term for aperiodic signals, our approach introduces a non-local transformation for aperiodic signals through multi-headed attention blocks. The combination of spectral layers and multi-headed attention blocks in SpectFormer is thus essential for effective learning of both periodic and aperiodic signal transformations.

Table 1. Ablation Analysis with different alpha value in SpectFormer architecture

Model	Params(M)	FLOPs(G)	Top-1(%)	Top-5(%)
SpectFormer_α_0	22.00	4.6	79.80	-
SpectFormer_α_2	21.03	4.3	79.87	94.69
SpectFormer_α_4	20.02	4.0	**80.21**	94.76
SpectFormer_α_6	19.01	3.7	80.14	**94.85**
SpectFormer_α_8	18.00	3.4	79.55.	94.59
SpectFormer_α_{10}	16.99	3.1	79.06	94.62
SpectFormer_α_{12}	16.00	2.9	78.60	94.20
iSpectFormer_α_4	20.02	4.0	79.03	94.30

5. Experiments and Results

Our proposed SpectFormer is evaluated through various empirical evidence on a range of mainstream computer vision tasks, including image recognition, object detection, and instance segmentation. To compare the quality of learned feature representations obtained from SpectFormer, we conduct the following evaluations: (a) SOTA comaprision on ImageNet1K for image recognition task ; (b) Conducting ablation studies that support each variant in our SpectFormer block and selecting the best α value for it; (c) Visualizing the learned visual representation by SpectFormer. (d) Transfer learning on CIFAR10, CIFAR-100, Oxford-IIIT flower, Standford Car dataset for Image recognition task using the SpectFormer (pre-trained on ImageNet1K) model; (e) Fine-tuning the SpectFormer (pre-trained on ImageNet1K) for downstream tasks such as object detection and instance segmentation on COCO; and We have included the DeiT-iii [55] comparison, model scaling results and details of training as well as fine-tuning experiments results, training setups and other details in the supplementary materials.

5.1. SOTA Comparision on ImageNet-1K

Table-2 presents a comparison of the performance of the state-of-the-art vision models and our SpectFormer variants. The ViT backbones with the best performance, VOLO-D2*, and VOLO-D3*, are trained using additional strategies such as Token Labeling objective with MixToken and convolutional stem for better patch encoding. We also use these strategies to train our SpectFormer variants in each size, which are denoted as SpectFormer-S*, SpectFormer-B*, and SpectFormer-L*.

The table shows that our Wave-ViT variants consistently outperform existing vision models, including ResNet, SE-ResNet, Vanilla ViTs (TNT, CaiT, CrossViT), and hierarchical ViTs (Swin, Twins-SVT, PVTv2, VOLO), under similar GFLOPs for each group. In particular, under the Base size, the Top-1 accuracy score of SpectFormer-H-B* can reach 85.1%, which leads to the absolute improvement of 0.3% against the best competitive Wave-ViT-B* (Top-1 accuracy: 84.8%). Under the Large size, when compared to ResNet-152 and SE-ResNet-152, which solely rely on CNN architectures, vanilla ViTs (TNT-B, CaiT-S36, and CrossViT) capture long-range dependencies through Transformer structure and outperform them. However, the performances of CaiT-S36 and CrossViT are still lower than most hierarchical ViTs (PVTv2-B5, VOLO-D3*, CMT-L, MaxViT, DaViT and Wave-ViT-L) that aggregate multi-scale contexts. Moreover, unlike PVTv2-B5, which uses irreversible down-sampling for self-attention learning, Wave-ViT uses invertible down-sampling with wavelet transforms. In particular, under the large size, the Top-1 accuracy score of SpectFormer-H-L* can reach 85.7%, which

Table 2. The table shows the performance of various vision backbones on the ImageNet1K [9] dataset for image recognition tasks. ⋆ indicates additionally trained with the Token Labeling objective using MixToken and a convolutional stem [60] for patch encoding. We have grouped the vision models into three categories based on their GFLOPs (Small, Base, and Large). The GFLOP ranges: Small (GFLOPs<6), Base (6≤GFLOPs<10), Large (10≤GFLOPs<30).

Method	Params (M)	GFLOPs	Top-1(%)	Top-5(%)	Method	Params (M)	GFLOPs	Top-1(%)	Top-5 (%)
Small					Large				
ResNet-50 [21]	25.5	4.1	78.3	94.3	ResNet-152 [21]	60.2	11.6	81.3	95.5
BoTNet-S1-50 [48]	20.8	4.3	80.4	95.0	ResNeXt101 [66]	83.5	15.6	81.5	-
Cross-ViT-S [5]	26.7	5.6	81.0	-	gMLP-B [36]	73.0	15.8	81.6	-
Swin-T [38]	29.0	4.5	81.2	95.5	DeiT-B [54]	86.6	17.6	81.8	95.6
ConViT-S [13]	27.8	5.4	81.3	95.7	SE-ResNet-152 [23]	66.8	11.6	82.2	95.9
T2T-ViT-14 [72]	21.5	4.8	81.5	95.7	Cross-ViT-B [5]	104.7	21.2	82.2	-
RegionViT-Ti+ [4]	14.3	2.7	81.5	-	ResNeSt-101 [74]	48.3	10.2	82.3	-
SE-CoTNetD-50 [32]	23.1	4.1	81.6	95.8	ConViT-B [13]	86.5	16.8	82.4	95.9
Twins-SVT-S [7]	24.1	2.9	81.7	95.6	PoolFormer-M48 [71]	73.0	11.8	82.5	-
CoaT-Lite-S [67]	20.0	4.0	81.9	95.5	T2T-ViTt-24 [72]	64.1	15.0	82.6	95.9
FDViT-S [68]	21.5	2.8	81.5	-	FDViT-B [68]	67.8	11.9	82.4	-
PVTv2-B2 [62]	25.4	4.0	82.0	96.0	TNT-B [19]	65.6	14.1	82.9	96.3
LITv2-S [41]	28.0	3.7	82.0	-	CycleMLP-B4 [6]	52.0	10.1	83.0	-
MViTv2-T [31]	24.0	4.7	82.3	-	DeepViT-L [77]	58.9	12.8	83.1	-
Wave-ViT-S [70]	19.8	4.3	82.7	96.2	RegionViT-B [4]	72.7	13.0	83.2	96.1
CSwin-T [11]	23.0	4.3	82.7	-	CycleMLP-B5 [6]	76.0	12.3	83.2	-
DaViT-Ti [10]	28.3	4.5	82.8	-	ViP-Large/7 [22]	88.0	24.4	83.2	-
FLatten-CSwin-T [18]	21.0	4.3	83.1	-	CaiT-S36 [56]	68.4	13.9	83.3	-
iFormer-S [47]	20.0	4.8	83.4	96.6	AS-MLP-B [34]	88.0	15.2	83.3	-
CMT-S [16]	25.1	4.0	83.5	-	BoTNet-S1-128 [48]	75.1	19.3	83.5	96.5
MaxViT-T [58]	31.0	5.6	83.6	-	Swin-B [38]	88.0	15.4	83.5	96.5
Wave-ViT-S⋆ [70]	22.7	4.7	83.9	96.6	Wave-MLP-B [50]	63.0	10.2	83.6	-
BiFormer-S* [78]	26.0	4.5	84.3	-	LITv2-B [41]	87.0	13.2	83.6	-
SpectFormer-H-S⋆	**22.2**	**3.9**	**84.3**	**96.9**	PVTv2-B4 [62]	62.6	10.1	83.6	96.7
Base					ViL-Base [75]	55.7	13.4	83.7	-
ResNet-101 [21]	44.6	7.9	80.0	95.0	Twins-SVT-L [7]	99.3	15.1	83.7	96.5
BoTNet-S1-59 [48]	33.5	7.3	81.7	95.8	Hire-MLP-Large [17]	96.0	13.4	83.8	-
T2T-ViT-19 [72]	39.2	8.5	81.9	95.7	RegionViT-B+ [4]	73.8	13.6	83.8	-
CvT-21 [64]	32.0	7.1	82.5	-	Focal-Base [69]	89.8	16.0	83.8	96.5
GFNet-H-B [46]	54.0	8.6	82.9	96.2	PVTv2-B5 [62]	82.0	11.8	83.8	96.6
Swin-S [38]	50.0	8.7	83.2	96.2	SE-CoTNetD-152 [32]	55.8	17.0	84.0	97.0
Twins-SVT-B [7]	56.1	8.6	83.2	96.3	DAT-B [65]	88.0	15.8	84.0	-
SE-CoTNetD-101 [32]	40.9	8.5	83.2	96.5	LV-ViT-M⋆ [26]	55.8	16.0	84.1	96.7
PVTv2-B3 [62]	45.2	6.9	83.2	96.5	CSwin-B [11]	78.0	15.0	84.2	-
LITv2-M [41]	49.0	7.5	83.3	-	HorNet-B_{GF} [45]	88.0	15.5	84.3	-
RegionViT-M+ [4]	42.0	7.9	83.4	-	DynaMixer-L [63]	97.0	27.4	84.3	-
MViTv2-S [31]	35.0	7.0	83.6	-	MViTv2-B [31]	52.0	10.2	84.4	-
CSwin-S [11]	35.0	6.9	83.6	-	FLatten-CSwin-B [18]	75.0	15.0	84.5	-
DaViT-S [10]	49.7	8.8	84.2	-	DaViT-B [10]	87.9	15.5	84.6	-
VOLO-D1⋆ [73]	26.6	6.8	84.2	-	CMT-L [16]	74.7	19.5	84.8	-
CMT-B [16]	45.7	9.3	84.5	-	MaxViT-B [58]	120.0	23.4	85.0	-
FLatten-CSwin-S [18]	35.0	6.9	83.8	-	MixMAE [37]	88.0	16.3	85.1	-
STViT-S [24]	25.0	4.4	83.6	-	STViT-L [24]	95.0	15.6	85.3	-
SiMBA-L(EinFFT) [44]	36.6	9.6	84.4	-	VOLO-D2⋆ [73]	58.7	14.1	85.2	-
MaxViT-S [58]	69.0	11.7	84.5	-	BiFormer-B* [78]	58.0	9.8	85.4	-
iFormer-B [47]	48.0	9.4	84.6	97.0	VOLO-D3⋆ [73]	86.3	20.6	85.4	-
Wave-ViT-B⋆ [70]	33.5	7.2	84.8	97.1	Wave-ViT-L⋆ [70]	57.5	14.8	85.5	97.3
SVT-H-B⋆ [43]	32.8	6.3	85.2	97.3	SVT-H-L⋆ [43]	54.0	12.7	85.7	97.5
SpectFormer-H-B⋆	**33.1**	**6.3**	**85.1**	**97.3**	**SpectFormer-H-L⋆**	**54.7**	**12.7**	**85.7**	**97.5**

leads to the absolute improvement of 0.2% against the best competitive Wave-ViT-L⋆ (Top-1 accuracy: 85.5%). Our SpectFormer-H-L⋆ achieves better accuracy and efficiency by enabling initial spectral blocks in the transformer encoder and attention blocks are at the top blocks.

5.2. Ablation analysis on spectral architectures

We conduct an experiment on the spectral network for the spectral layer in SpectFormer architecture as shown in figure-3. In the first study, we compare various spectral architectures to develop SpectFormer, such as the Fourier Network (FN), the discrete cosine transform (DCT), the Fourier Neural Operator (FNO), the Fourier Gating Network (FGN), and Wavelet Gating Network (WGN) as shown in table 3. The Fourier transform network indicates the spectral layer contains just a Fourier transform instead of a multi-headed self-attention network. Similarly, the Fourier gating network uses a Fourier transform and its contribution is controlled by learnable weight parameters, followed by the inverse Fourier transform. We use neural operator techniques for channel mixing and Fourier transform techniques for token mixing similar to FNO [15]. Wavelet gating network uses a wavelet transform followed by learnable weight parameters to control the wavelet decomposition. We observe that the Fourier gating network outperforms all other architectures as it uses a gating technique to control the Fourier features.

We conduct this ablation study on the small-size networks in hierarchical SpectFormer architecture on the ImageNet-1K dataset. We have also conducted a study to

Table 3. **Spectral Layer Variants:** This table shows the ablation analysis of various spectral layers in SpectFormer architecture such as the FN, the FNO, the DCT, the WGN, and the FGN. We conduct this ablation study on the small-size networks in stage architecture. This indicates that FGN performs better than other kinds of networks

Model	Params(M)	FLOPs(G)	Top-1(%)	Top-5(%)
FN	21.17	3.9	84.02	96.77
FNO	21.33	3.9	84.09	96.86
DCT	22.04	4.1	84.06	96.1
WGN	21.59	3.9	83.70	96.56
FGN	22.22	3.9	**84.25**	**96.93**

Table 4. **Initial Attention vs Spectral vs Convolutional Layer:** This table compares SpectFormer where initial Spectral layers and later attention layers, SpectFormer-Inverse where initial attention layers and later scatter layers, and SpectFormer with initial convolutional layers. Also, we show an alternative spectral layer and attention layer. This shows that the Initial spectral layer works better compared to the rest.

Model	Params(M)	FLOPs(G)	Top-1(%)	Top-5(%)
Init-Spectral	22.2	3.9	84.2	96.9
Init-CNN	21.7	4.1	84.0	95.7
Final-Spectral	21.8	3.9	83.1	94.6
Alt-Spectral	22.4	4.6	83.4	95.0

Table 5. This shows a performance comparison of SpectFormer with similar Transformer Architectures with different sizes of the networks on ImageNet-1K.

Network	Params	FLOPs	Top-1
Vanilla Transformer Comparison			
DeiT-Ti [54]	5M	1.2	72.2
Fourier [39]	-	-	73.3
GFNet-Ti [46]	7M	1.3	74.6
SpectFormer-T	9M	1.8	**76.9**
DeiT-S [54]	22M	4.6	79.8
Fnet-S [29]	15M	2.9	71.2
GFNet-XS [46]	16M	2.9	78.6
GFNet-S [46]	25M	4.5	80.0
SpectFormer-XS	20M	4.0	**80.2**
SpectFormer-S	32M	6.6	81.7
DeiT-B [54]	86M	17.5	81.8
GFNet-B [46]	43M	7.9	80.7
SpectFormer-B	57M	11.5	**82.1**

Network	Params	FLOPs	Top-1
Hierarchical Transformer Comparison			
PVT-S [61]	25M	3.8	79.8
Swin-T [38]	29M	4.5	81.3
GFNet-H-S [46]	32M	4.6	81.5
LIT-S [42]	27M	4.1	81.5
SpectFormer-H-S	22M	3.9	**84.2**
PVT-M [61]	44M	6.7	81.2
Swin-S [38]	50M	8.7	83.0
GFNet-H-B [46]	54M	8.6	82.9
LIT-M [42]	48M	8.6	83.0
SpectFormer-H-B	33M	6.3	**85.0**
PVT-L [61]	61M	9.8	82.3
Swin-B [38]	88M	15.4	83.3
LIT-B [42]	86M	15.0	83.4
SpectFormer-H-L	55M	12.7	**85.7**

illustrate the performance of SpectFormer using different kinds of networks as initial layer for instance, if we have initial convolutional layers or having only initial attention layers or having alternative spectral and attention layers or having initial spectral layers followed by deeper attention layers as in SpectFormer as reported in table-4. In this table, we compared SpectFormer with all variants and observed that the Initial spectral layer works better compared to the rest.

We also illustrate the performance differences in using the number of spectral layers(α) to make variant of the SpectFormer architecture as shown in figure-3. We select a vanilla transformer architecture similar to DeIT and we replace the number of attention layers with spectral layers. We choose Deit-Small [54] that has 12 layer architecture with a hidden dimension of D (dim=384) and a similar architecture in GFNet [46] is GFNet-XS. We characterize the study using a hyper-parameter α. We select the α value zero for the DeiT-S network and a value of twelve for the GFNet-XS transformer. We fine-tune the α value on ImageNet-1K dataset and find that the ideal α value is four. This result is captured in the table 1. We started with different α values such as 2, 4, 6, 8, and 10 for validating the DeIT small network where α_2 indicates two layers of spectral and ten (12-4) layers of attention in the architecture, while α_4 indicates four layers of spectral and eight (12-4) layers of attention network.

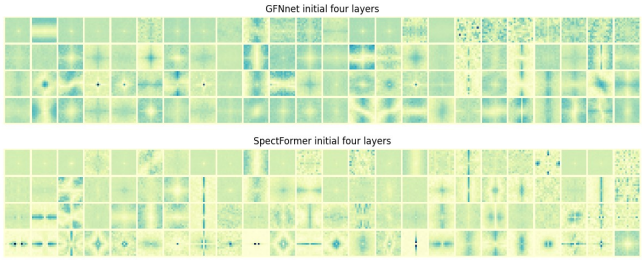

Figure 4. This figure shows the Filter characterization of the initial four layers of the DeiT [54], GFNet [46] and **SpectFormer** model. It clearly shows that SpectFormer captures local filter information, such as lines and edges, more sharply compared to other transformers.

5.3. Comparison with Similar Architectures

We observed that all the hierarchical models (SpectFormer-H-S, SpectFormer-H-B, and SpectFormer-H-L) performed better than the vanilla architecture and are state-of-the-art, as shown in table-2. We compared the performance of SpectFormer to similar architectures on the ImageNet-1k dataset as shown in table-5. Compared to attention-based models like DeiT, SpectFormer performed better than DeiT in all size models (T, XS, S, and B). Compared to spectral-based models like Fnet, FourierFormer, and GFNet, SpectFormer performed better than all of them in all sizes (T, XS, S, and B). We then compared SpectFormer to hierarchical attention architectures such as PVT, Swin, LiT, and LiTv2 as well as spectral architectures like GFNet-H-S/B. We have observed that SpectFormer outperforms vanilla transformers, hybrid transformers, other spectral transformers, and even other weighted attention transformers. SpectFormer performs 2% better than the latest similar model LiTv1 for small architecture and 3% better than the best spectral architecture GFNet-H-S. When compared to DeiT and DeiT III, SpectFormer performs better. Thus, we have shown in the SoTA studies, that SpectFormer is the first spectral transformer to bridge the performance gap with state-of-art attention-based transformers such as Volo, MaxViT and WavViT. Other spectral transformers such as GFNet reduce complexity to O(Nlog(N), however, have a significant performance gap with state of art attention-based transformers.

5.4. Filter Visualisation analysis

Our primary focus is to analyze the filter components to emphasize Spectformer's exceptional sharpness in filter

Table 6. The performances of various vision models on the COCO val2017 dataset for the downstream tasks of object detection and instance segmentation. RetinaNet is used as the object detector for the object detection task, and the Average Precision (AP) at different IoU thresholds or two different object sizes (*i.e.*, small and base) are reported for evaluation. For instance segmentation task, we adopt Mask R-CNN as the base model, and the bounding box and mask Average Precision (*i.e.*, AP^b and AP^m) are reported for evaluation.

Backbone	Mask R-CNN 1x [20]						RetinaNet 1x [35]					
	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50 [21]	38.0	58.6	41.4	34.4	55.1	36.7	36.3	55.3	38.6	19.3	40.0	48.8
Swin-T [38]	42.2	64.6	46.2	39.1	61.6	42.0	41.5	62.1	44.2	25.1	44.9	55.5
Twins-SVT-S ([7])	43.4	66.0	47.3	40.3	63.2	43.4	43.0	64.2	46.3	28.0	46.4	57.5
LITv2-S [41]	44.9	-	-	40.8	-	-	44.0	-	-	-	-	-
RegionViT-S [4]	44.2	-	-	40.8	-	-	43.9	-	-	-	-	-
PVTv2-B2 [62]	45.3	67.1	49.6	41.2	64.2	44.4	**44.6**	**65.6**	**47.6**	**27.4**	**48.8**	58.6
SpectFormer-S	**46.2**	**68.1**	**50.8**	**42.0**	**65.2**	**45.4**	44.2	64.8	47.3	27.3	48.1	**59.5**
ResNet101 [21]	40.4	61.1	44.2	40.4	61.1	44.2	38.5	57.8	41.2	21.4	42.6	51.1
Swin-S [38]	44.8	66.6	48.9	40.9	63.4	44.2	44.5	65.7	47.5	27.4	48.0	59.9
Twins-SVT-B ([7])	45.2	67.6	49.3	41.5	64.5	44.8	45.3	66.7	48.1	28.5	48.9	60.6
RegionViT-B [4]	45.4	-	-	41.6	-	-	44.6	-	-	-	-	-
LITv2-M [41]	46.8	-	-	42.3	-	-	46.0	-	-	-	-	-
PVTv2-B3 [62]	47.0	68.1	51.7	42.5	65.7	45.7	45.9	**66.8**	49.3	28.6	**49.8**	**61.4**
SpectFormer-B	46.9	**68.8**	**51.8**	**42.7**	**65.9**	**45.7**	**46.0**	66.4	**49.7**	**29.5**	49.7	61.1

Table 7. **Results on transfer learning datasets**. We report the top-1 accuracy on the four datasets as well as the number of parameters and FLOPs.

Model	CIFAR-10	CIFAR-100	Flowers-102	Cars-196
ResNet50 [21]	-	-	96.2	90.0
EfficientNet-B7 [49]	98.9	91.7	98.8	92.7
ViT-B/16 [12]	98.1	87.1	89.5	-
ViT-L/16 [12]	97.9	86.4	89.7	-
Deit-B/16 [54]	**99.1**	**90.8**	98.4	92.1
ResMLP-24 [53]	98.7	89.5	97.9	89.5
GFNet-XS [46]	98.6	89.1	98.1	92.8
GFNet-H-B [46]	99.0	90.3	98.8	93.2
Spectformer-B	98.9	90.3	**98.9**	**93.7**

visualization. It clearly shows that the filter coefficient captures local information such as lines, edges, and different orientations of an Image, as illustrated in Figure 4. Providing compelling evidence that SpectFormer's spectral layers are more localized and sharpened compared to fully spectral transformers such as GFNet [46]. It must be noted that during the training phase of SpectFormer, the deeper attention layers enable the initial spectral layers to learn the edges and lines of the image features, whereas in GFNet, since deeper attention layers are missing, the learned filters are not sharp enough in the initial layers.

5.5. Task Learning: Object Detection

We conducted experiments on MS COCO 2017, which is a widely used benchmark for object detection and instance segmentation, comprising around 118K images for the training set and approximately 5K images for the validation set. Our approach involved experimenting with two detection frameworks, namely RetinaNet [35] and Mask R-CNN [20], and we measured model performance using Average Precision (AP). We use the pre-trained model SpectFormer trained on the ImageNet-1K dataset to initialize the backbone architecture and Xavier initialization for additional layers of the network. These results are shown in table- 6. The experimental results, as presented in table- 6, indicate that SpectFormer has comparative results on both the RetinaNet [35] and Mask R-CNN [20] models. We have compared with the latest work including LITv2 [41], RegionViT, and PVT [61] transformer models. Further, our SpectFormer model demonstrated significantly better performance than ResNet in terms of AP. More importantly, SpectFormer outperformed all compared vanilla ViT models and hierarchical transformer models, achieving the best AP performance.

5.6. Transfer Learning Comparison

In order to assess the effectiveness of SpectFormer's architecture and learned representation, we conducted evaluations on multiple transfer learning benchmark datasets, which included CIFAR-10 [28], CIFAR-100 [28], Stanford Cars [27], and Flowers-102 [40]. Here we compare the performance of SpectFormer pre-trained on ImageNet-1K and fine-tuned on the new datasets for the image classification task. Both the basic and best models were evaluated for their transfer learning performance, and the comparison is captured in table- 7. The results show that the proposed models performed well on downstream datasets, surpassing ResMLP models by a significant margin and achieving highly competitive performance comparable to state-of-the-art spectral network, GFNet [46]. Our models also exhibited competitive performance when compared to state-of-the-art CNNs and vision transformers.

6. Conclusion

Through this work, we analyzed the core architecture of transformers by using a mixed approach that includes spectral and multi-headed attention. Previously, transformers used either all-attention layers or more recently spectral layers have been used. Spectformer combines both these aspects and shows consistently better performance than either all-attention or all-spectral layers. The work achieves state-of-the-art (85.7%) top-1 accuracy on the ImageNet-1K. We use a parameterized approach that suggests further scope for adaptation of this work for specific tasks.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 3

[4] Chun-Fu Chen, Rameswar Panda, and Quanfu Fan. Regionvit: Regional-to-local attention for vision transformers. In *International Conference on Learning Representations*, 2022. 3, 6, 8

[5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2, 6

[6] Shoufa Chen, Enze Xie, GE Chongjian, Runjian Chen, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. In *International Conference on Learning Representations*, 2022. 6

[7] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021. 2, 6, 8

[8] Y. Cui, C. Jiang, G. Wu, and L. Wang. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(06):4129–4146, jun 2024. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022. 6

[11] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 8

[13] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 6

[14] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021. 3, 4

[15] John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 6

[16] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022. 6

[17] Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, and Yunhe Wang. Hire-mlp: Vision mlp via hierarchical rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–836, June 2022. 6

[18] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5961–5971, 2023. 2, 6

[19] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021. 2, 6

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 8

[22] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2022. 6

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[24] Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, and Tieniu Tan. Vision transformer with super token sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22690–22699, 2023. 2, 6

[25] David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959. 2

[26] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *Advances in Neural Information Processing Systems*, 34:18590–18602, 2021. 6

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 8

[28] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 8

[29] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. 1, 3, 7

[30] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 3

[31] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 6

[32] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6

[33] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar, et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020. 3

[34] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. In *International Conference on Learning Representations*, 2022. 6

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 8

[36] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021. 6

[37] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6252–6261, 2023. 6

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 6, 7, 8

[39] Tan Minh Nguyen, Minh Pham, Tam Minh Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. In *Advances in Neural Information Processing Systems*, 2022. 3, 7

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 8

[41] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. In *Advances in Neural Information Processing Systems*, 2022. 6, 8

[42] Zizheng Pan, Bohan Zhuang, Haoyu He, Jing Liu, and Jianfei Cai. Less is more: Pay less attention in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2035–2043, 2022. 2, 7

[43] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Scattering vision transformer: Spectral mixing matters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 6

[44] Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024. 6

[45] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems*, 35:10353–10366, 2022. 6

[46] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 1, 2, 3, 6, 7, 8

[47] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng YAN. Inception transformer. In *Advances in Neural Information Processing Systems*, 2022. 6

[48] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021. 6

[49] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 8

[50] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang. An image patch is a wave: Phase-aware vision mlp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10935–10944, 2022. 6

[51] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[52] Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024. 1

[53] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 8

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1, 2, 4, 6, 7, 8

[55] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 1, 5

[56] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021. 2, 6

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[58] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 459–479. Springer, 2022. 6

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 3

[60] Pichao Wang, Xue Wang, Hao Luo, Jingkai Zhou, Zhipeng Zhou, Fan Wang, Hao Li, and Rong Jin. Scaled relu matters for training vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2495–2503, 2022. 6

[61] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2, 7, 8

[62] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6, 8

[63] Ziyu Wang, Wenhao Jiang, Yiming M Zhu, Li Yuan, Yibing Song, and Wei Liu. Dynamixer: a vision mlp architecture with dynamic mixing. In *International Conference on Machine Learning*, pages 22691–22701. PMLR, 2022. 6

[64] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 6

[65] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022. 6

[66] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[67] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021. 6

[68] Yixing Xu, Chao Li, Dong Li, Xiao Sheng, Fan Jiang, Lu Tian, and Ashish Sirasao. Fdvit: Improve the hierarchical architecture of vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5950–5960, 2023. 2, 6

[69] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 6

[70] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 328–345. Springer, 2022. 2, 3, 6

[71] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 6

[72] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 2, 6

[73] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 6

[74] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022. 6

[75] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3008, 2021. 6

[76] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12083–12093, 2022. 3

[77] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 6

[78] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10323–10333, 2023. 2, 6