AFIDAF: Alternating Fourier and Image Domain Adaptive Filters as an Efficient Alternative to Attention in ViTs

Yunling Zheng[1](✉), Zeyi Xu[1], Fanghui Xue[1], Biao Yang[1], Jiancheng Lyu[2], Shuai Zhang[2], Yingyong Qi[2], and Jack Xin[1]

[1] University of California, Irvine, Irvine, USA
{yunliz1,zeyix1,fanghuix,biaoy1,jack.xin}@uci.edu
[2] Qualcomm AI Research, San Diego, USA
{jianlyu,shuazhan,yingyong}@qti.qualcomm.com

Abstract. We propose and demonstrate an alternating Fourier and image domain filtering approach for feature extraction as an efficient alternative to build a vision backbone without using the computationally intensive attention. The performance among the lightweight models reaches the state-of-the-art level on ImageNet-1K classification, and improves downstream tasks on object detection and segmentation consistently as well. Our approach also serves as a new tool to compress vision transformers (ViTs).

Keywords: Fourier domain filtering · Group shuffled large kernel convolution · Dual domain feature extraction

1 Introduction

Two mainstream computer vision (CV) networks are convolutional neural network (CNN, [15]) and vision transformer (ViT, [14]). ViTs have surpassed the performance of CNNs in recent years however at the expense of large model size and flops even though efficient attention is utilized [24]. To achieve high performance lightweight (LW) backbone models with parameter size around 5 million, attention free networks with low cost global mechanism to upgrade standard convolution has been a successful line of inquiry. For example, Fourier transform is a global convolution and can facilitate such a possibility as demonstrated in AFF network [12] lately. On the other hand, large convolution kernel vision networks [9] approach this goal from the image domain, while hybrid LW models combine mobile convolution and attention [44].

The main contributions of our paper include:

- Identify the lack of spatial mixing in AFFNet [12] and propose an *alternating Fourier and image domain adaptive filtering* (AFIDAF) proxy to attention in ViTs. The spatial filtering equips the large kernel convolution [9] with group shuffling operations for added efficiency.

G. Bebis et al. (Eds.): ISVC 2024, LNCS 15046, pp. 17–30, 2025.
https://doi.org/10.1007/978-3-031-77392-1_2

- Show that AFIDAF improves AFF consistently on CV (ImageNet-1K classification and downstream) tasks while remaining in the LW category.
- Develop a hierarchical AFIDAF framework based on Swin [24] for ViT compression while maintaining performance on CV tasks.

The rest of the paper contains sections on related work, method, experiments and conclusions.

2 Related Work

2.1 CNNs

Convolution has been the basic operation of image feature extraction for over two decades [15], due its flexibility in adopting various kernel sizes for various receptive field of views to cover the image domain under translation invariance as well as its natural interpretation as filtering. However, the convolution operation uses static weights and so lacks adaptability across pixels in different parts of an image. It is also spatially local due to limitation of the kernel size. As a result, ViTs ([14] and its variants), based on global attention originally designed for natural language processing (NLP) tasks [37], outperformed well-known CNNs on computer vision (CV) tasks, see [24,42] among others.

To improve CNNs to and over the level of benchmark ViTs [14,24], additional functionalities have been introduced in recent years. One is large kernel approximation (LKA, [9]) that leverages the strengths of both convolution and self-attention by including local structure (contextual) information, long-range dependence, and spatial-channel adaptability. Another line of inquiry is ConvNext where large kernel sizes and layer norm [23], and global response normalization layer (see [39] for inter-channel feature competition) are utilized for enhancement. These developments are motivated by Swin transformers [24] yet at similar or larger capacities.

2.2 ViTs

Due to quadratic complexity of attention in ViT [14], efficient token mixing and global attention approximations have been actively studied with various ideas stemming from shifted window of Swin [24]. In lieu of window shifting, competitive performances have been reported on ImageNet-1K and downstream tasks by techniques such as pooling (Poolformer [45]), shuffling (Shuffleformer [13]), mixing across windows and dimensions (Mixformer [3]), high/low frequency global attention decomposition (Hiloformer [31]), pale shaped window attention (Paleformer, [41]), cross shaped windown attention (CSwinformer [4]) among others. See also hybrid and unified CNN–ViT models [6,7,11,18,19,40,43].

2.3 Fourier Transform Based Vision Networks

Fourier transform has been proposed first for NLP tasks [16] and then found effective in promoting token mixing in CV for frequency domain filtering and feature extractions [32,34]. FFT is also a form of convolution, though with a global kernel size and circular padding. Injecting adaptivity in the Fourier domain has been found useful for mimicing self-attention in ViTs, see [8,12].

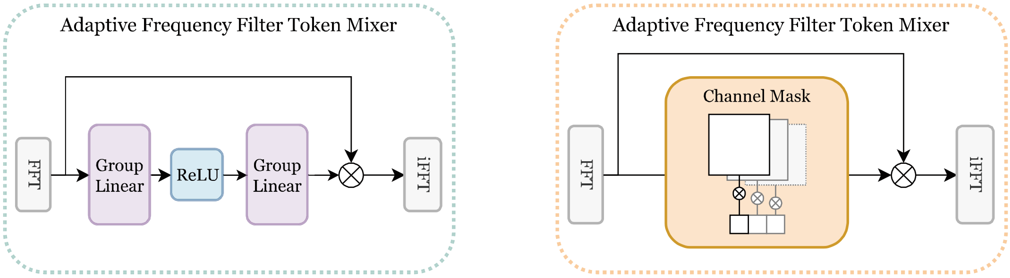
2.4 Lightweight Vision Networks

Lightweight networks are desirable for mobile deployment and resource constrained applications. Separable (group) convolutions and shuffle operations are effective techniques for designing CNNs in the lightweight category, see MobileNets [33], ShuffleNets [25,26,47] and references therein among others. Lightweight ViTs have been proposed combining MobileNet and efficient attention blocks in [20,29,44], see also [38] for a ViT motivated mobile CNN. A lightweight Fourier transform based attention-free vision network is AFF [12] which forms the baseline of our work here.

3 Method

We first review the adaptive Fourier filters for efficient token mixing proposed in AFF [12], point out its limitation (or lack of action in the frequency/image domain) and present our method as an alternating dual domain adaptive filter to enhance performance on visual tasks while keeping the model size in the lightweight range.

3.1 AFF Block and Limitation


Adaptive Frequency Filter Token Mixer
FFT
Group Linear
ReLU
Group Linear
iFFT
Adaptive Frequency Filter Token Mixer
FFT
Channel Mask
iFFT

(a) General (theoretical) adaptive Fourier filter proposed in AFFNet [12] for mixing features in the Fourier domain.

(b) Implemented AFFNet, applying the Fourier domain filter channel-wise as a mask, limiting its ability to represent frequency features effectively.

Fig. 1. Comparative illustration of the AFFNet block's theoretical framework (a) and its practical application (b), highlighting the discrepancy between the conceptual design and the actual implementation.

Consider feature tensor $X \in R^{H\times W\times C}$ which is mapped from an input image, with spatial resolution $H \times W$ and channel number C. A token $x \in R^{1\times 1\times C}$ is a restriction of X at a fixed spatial location. Token mixing is a key operation in evolving X through a deep network. A general expression is: $x_q^* := \sum_{i\in N(x_q)} \omega_{i,q}\, \varphi(x_i)$, where x_q^* is the transformed token, $N(x_q)$ is a neighborhood of x_q of certain size, $\omega_{i,q}$ the weight matrix, and $\varphi(\cdot)$ is embedding function. This formula is an abstraction of both CNN and transformer with suitable choices of N, φ and ω. Towards a computationally efficient, semantically adaptive and globally reaching token mixer desirable for lightweight networks, AFF [12] proposed to 1) (global) fast Fourier transform X in (h, w) to $F(X)$, 2) (local and adaptive on Fourier domain) mask it nonlinearly in the point-wise sense, 3) inverse Fourier back to the image domain (Fig. 1):

$$X^* = F^{-1}[M(F(X)) \odot F(X)], \tag{1}$$

where $M(\cdot)$ is implemented as subnetwork consisting of a group 1×1 convolution (linear) layer, followed by a ReLU function and another group linear layer; $\odot$ is elementwise multiplication (Hadamard product). The authors argued through convolution theorem that the AFF block (1) is global, adaptive token mixing and is mathematically equivalent to adopting a large-size dynamic convolution kernel as the weights for token mixing. An advantage of (1) is that the resulting model is attention free, CNN based, and lightweight with competitive performance on ImageNet-1K, though less so on downstream or dense prediction tasks (object detection and segmentation).

Through checking the authors' Github codes, we found that *the masking function M actually only acts on the channel dimension while being an identity map on the frequency plane*, which limits its performance and spatial resolution. More precisely, the actually implemented AFF block is:

$$X^*_{aff} = F^{-1}[M_C(F(X)) \odot F(X)], \tag{2}$$

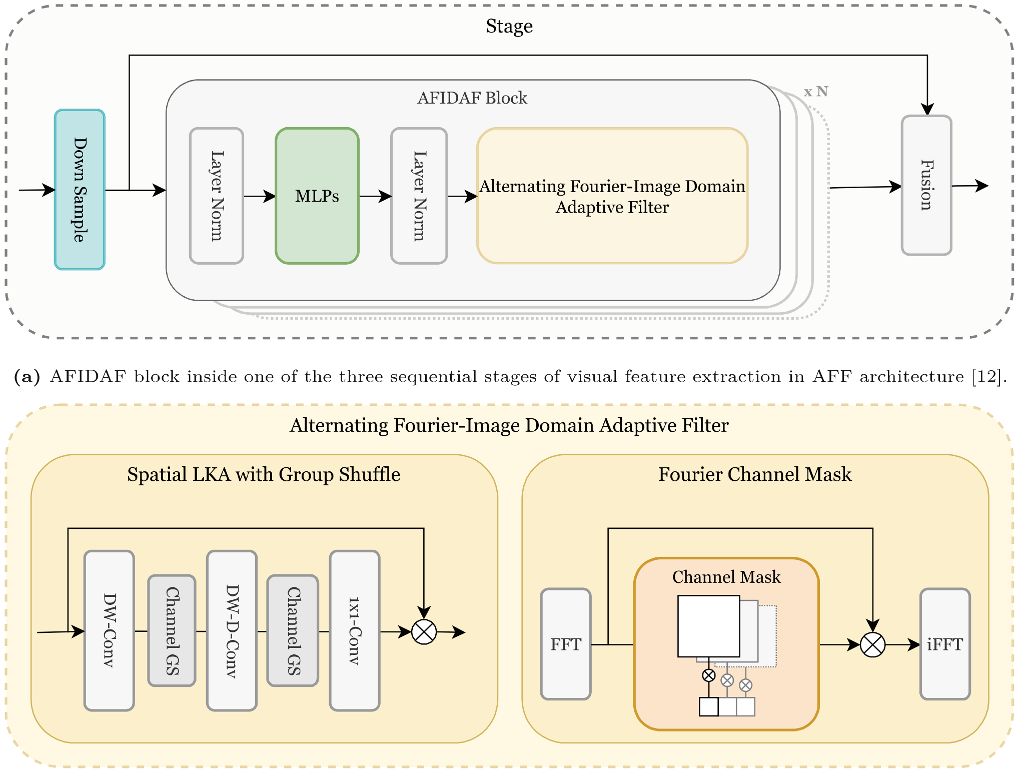
where $M_C(\cdot)$ is a subnet in the channel dimension leaving frequency dimensions unchanged.

The main contribution of our paper is to realize that additional spatial filtering on image (or an equivalent on frequency) domain on top of (2) can improve AFF while keeping model size in the lightweight range. Instead of doing so in the frequency domain alone (or directly on (2)), we propose an *alternating adaptive filtering methodology between image domain and Fourier domain* (AFIDAF). Abstractly, the AFIDAF block is:

$$X^*_{afidaf} = F^{-1}[M_C(F(M_I(X)) \odot F(M_I(X))], \tag{3}$$

where $M_I(\cdot)$ is a large kernel approximation (LKA, Fig. 4 of [9]) with additional grouping and shuffling. The LKA [9] consists of depth-wise convolutions in the $H \times W$ domain followed by a CNN type multiplicative attention. To be more efficient, we further downsize M_I with group convolutions and shuffle operations, see the left subplot of Fig. 2(b). The alternating strategy (3) is a *splitting method* to handle token mixing in all $H \times W \times C$ dimensions. One potential difficulty to find a selective mask M in Eq. (1) on the frequency domain is that it must be properly localized to correspond to a large receptive field of view in the image domain by the uncertainty principle of Fourier transform. On the other hand, to resolve high frequency well, the mask must also cover the corresponding part of the frequency plane. In the AFIDAF approach (3), local and high frequency features of an image (edges/corners/textures etc.) are resolved by large kernel convolutions inside $M_I(\cdot)$ on the image domain; the low frequency and non-local features outside of individual kernel's reach are captured by channel mixing (2) on the Fourier domain. So Eq. (3) is a local-global image feature extractor. It is an interesting problem for a future study to localize (2) and decrease kernel size of $M_I(\cdot)$ (hence also localize in the image domain) to reduce AFIDAF model parameter size. We present our model design next.

3.2 AFIDAF Architecture



(a) AFIDAF block inside one of the three sequential stages of visual feature extraction in AFF architecture [12].

(b) An alternating image domain filtering (efficient large kernel convolution) and Fourier domain channelwise filtering to form a basic AFIDAF block (Fig. 2a). DW=depthwise, GS= group shuffle, DWD=depthwise-dilated.

Fig. 2. Illustration of AFIDAF in block and stage views.

Image Domain Adaptive Filtering. To compensate for the lack of spatial filtering in AFF implementation, we propose adding a full-size kernel convolution as an adaptive filter in the image domain, prior to the Fourier domain AFF filter, then repeat this block in each of the three stages of visual feature extractions in the AFF architecture. However, employing large kernel convolution can be computationally expensive.

To mitigate the high computational cost, we implement a decomposed large kernel convolution [9] combined with a channel-wise group shuffle [25,47]. This approach aims to reduce the computational overhead and large number of parameters typically associated with large kernel convolutions while still capturing long-range dependencies.

We adopt a convolution decomposition which includes three components: depth-wise spatial local convolution, depth-wise dilated convolution, and 1×1 channel convolution. The depth-wise spatial local convolution focuses on proximate features, maintaining spatial locality. The depth-wise dilated convolution extends spatial coverage to capture a broader context. Finally, the 1×1 channel convolution integrates channel-wise features, facilitating inter-channel interactions. It is often referred to as "attention" in the convolutional setting (see [9] and references therein). Thus, we arrive at:

$$\text{Attention}_{conv} := \text{Conv}_{1\times 1}(GS_{chan}((\text{DWD-Conv}(GS_{chan}(\text{DW-Conv}(X))))), \tag{4}$$

where X denotes input features, GS_{chan} is channel-wise group shuffle.

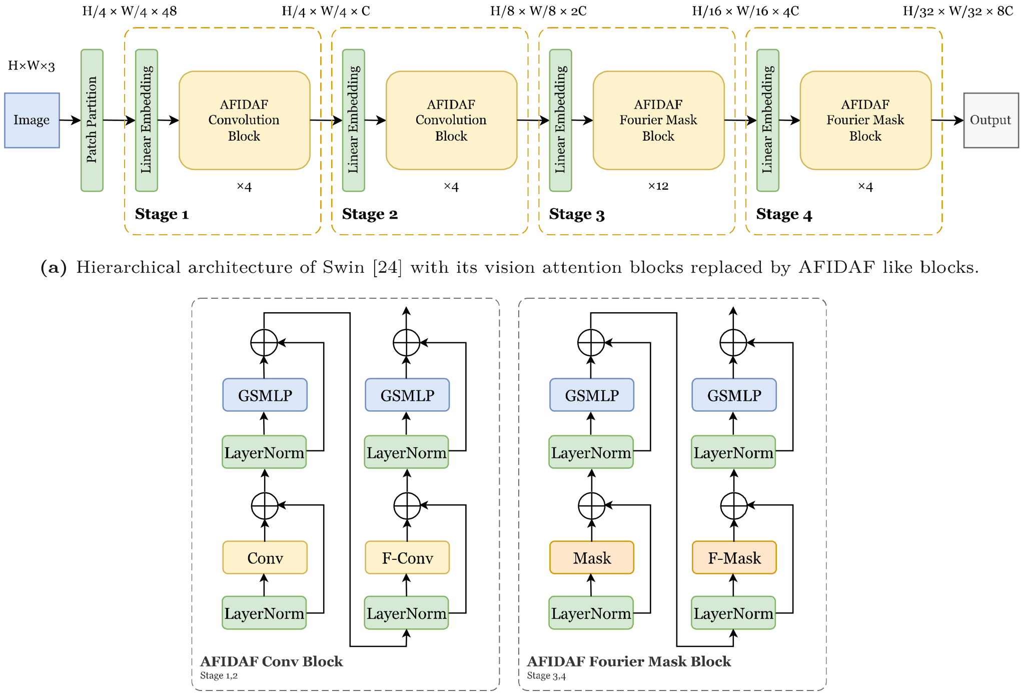
The channel-wise group shuffle further optimizes performance by reordering the channels in each group, ensuring effective feature mixing and reducing redundancy. This step enhances the learning process by promoting diverse feature representations without significantly increasing computational costs.

By incorporating these techniques, we achieve a balanced approach that leverages both spatial and Fourier domain filters, enhancing the AFF architecture's ability to efficiently and accurately extract meaningful visual features. This approach allows us to maintain reasonable computational efficiency while achieving the desired adaptive filtering effects, see Fig. 2b for the block view and Fig. 2a for the block in a stage which repeats three times from input to output.

Alternating Fourier-Image Domain Filtering. By integrating spatial and Fourier filters, we enhance the AFF architecture via a local-global approximation structure, enabling it to effectively and accurately extract significant visual features. This dual approach ensures that we maintain computational efficiency while achieving the desired adaptive filtering outcomes. Figure 2b and 2a illustrate this concept, showing the block view and its repetitive three-stage process from input to output, respectively.

3.3 Hierarchical AFIDAF

As another contribution of this paper, we improve the efficiency of existing ViTs with the dual domain alternating architecture. The Swin transformer has demonstrated a good performance with relatively low complexity among ViTs. However, window attention computations are known to be less device-friendly than convolutions. The subsequent MLPs also rapidly increase the model size. We shall maintain the hierarchical framework of Swin (Fig. 3a), while replacing its transformer blocks with our design of hierarchical AFIDAF (HAFIDAF) blocks (Fig. 3b).



(a) Hierarchical architecture of Swin [24] with its vision attention blocks replaced by AFIDAF like blocks.

(b) An alternating image filtering and Fourier channelwise mask to form a hierarchical AFIDAF block (Fig. 3a). F-Conv=frequency domain convolution, F-Conv(X)=iFFT(Conv(FFT(X))). F-Mask=Fourier Channel Mask, F-Mask(X)=iFFT(Mask(FFT(X))), where Mask=M_C as in Eq. 2. GSMLP is group shuffled multi-layer perception.

Fig. 3. Overview of HAFIDAF acting on Swin [24] and the resulting compressed hierarchical architecture.

HAFIDAF Blocks. Though in principle like AFIDAF, the hierarchical AFIDAF (HAFIDAF) blocks differ in the following sense. First, HAFIDAF is for the purpose of model compression. Second, the approximations on the spatial domain and the channel domain are made in separate stages (Fig. 3a).

The first two stages are AFIDAF convolution blocks. Such a block consists of an alternating-type spatial/frequency convolution, which resembles the large kernel approximation M_I in the setting of Fig. (2b). Here, convolution is performed in place of window attention, as a more friendly alternative to mobile devices. Moreover, Fourier convolution is performed every 2 blocks. It has two main advantages over other simple architectures, e.g. the Hadamard product on image domain. First, F-conv acts on small frequency kernels, allowing for entries of similar frequency modes to connect. In comparison, Hadamard product acts only on single pixels. Second, compared to the Hadamard product, F-Conv (Fig. (3b)) is smaller in size and thus more efficient. A group shuffle MLP then follows to contribute to higher efficiency as well.

The latter two stages contain AFIDAF Fourier Mask blocks. Each block consists of the channelwise Fourier mask, which resembles the channelwise operator M_C in the setting of Sect. 3.1. A group shuffle MLP follows afterward. In all stages, Layer Normalizations are performed beforehand, and shortcut connections are present for ease of training the deep layers.

4 Experiments

4.1 Image Classification

Setting. The ImageNet-1K dataset [35], containing over 1.2 million images across 1000 object categories, is utilized for training our models from scratch to validate the effectiveness and efficiency of our proposed AFIDAF network in image classification. We trained AFIDAF from scratch for 300 epochs using 256 × 256 pixel images on 8 NVIDIA RTX A6000 GPUs with a batch size of 1024. The learning rate schedule follows a cosine decay, starting at 2e−3 and decreasing to a minimum of 2e−4, with the AdamW optimizer (weight decay of 0.05) and cross-entropy loss.

The input features from preprocessing have a size of 256^2 with 1 block and 16 output channels, passing through the network composed of 1 Conv Stem Layer and 3 Down Sample AFIDAF Blocks, concluding with the output. The Conv Stem Layer yields an output size of 64^2, encompassing 4 blocks, and 32 output channels. The first Down Sample AFIDAF Block produces an output of 32^2 with 2 blocks, and 96 channels for AFIDAF-T (128 for AFIDAF). The second Down Sample AFIDAF Block outputs 16^2 with 4 blocks and 160 channels for AFIDAF-T (256 for AFIDAF). The third Down Sample AFIDAF Block results in an output size of 8^2 with 3 blocks, and 192 channels for AFIDAF-T (320 for AFIDAF).

Results. We compare our proposed AFIDAF model with other state-of-the-art lightweight models in Table 1. Our AFIDAF demonstrates superior performance, achieving 80.9% Top-1 accuracy with 6.5M parameters and 1.5G FLOPs, outperforming other lightweight networks of similar sizes. Additionally, AFIDAF-T achieves 77.6% Top-1 accuracy with just 3.0M parameters and 0.8G FLOPs.

Ablation on Alternating Domain Filtering. To validate the effectiveness of our alternating Fourier and image domain filtering approach, we compare AFIDAF with AFFNet [12] and IDAF (replacing the AFF block with image domain LKA [9]) on ImageNet-1K. The results, shown in the last 3 lines of Table 1, demonstrate the superiority of our alternating domain approach over single-domain methods.

Table 1. Lightweight network classification comparison on ImageNet-1K dataset. IDAF (image domain adaptive filtering only) replaces AFF block's channel mixing with image domain LKA [9].

Model	Params (M)	Flops (G)	Top-1 (%)
MViT-XS [29]	2.3	1.0	74.8
EFormer-S0 [20]	3.5	0.4	75.7
VAN-B0 [9]	4.1	0.9	75.4
EdgeNext-XS [27]	2.3	0.5	75.0
AFFNet-T [12]	2.6	0.8	77.0
AFIDAF-T	3.0	0.8	**77.6**
MNetv2 [28]	6.9	0.6	74.7
ShuffleNetV2 [26]	5.5	0.6	74.5
MNetv3 [28]	5.4	0.2	75.2
T2T-ViT [46]	6.9	1.8	76.5
DeiT-T [36]	5.7	1.3	72.2
CoaT-Lite-T [43]	5.7	1.6	77.5
LeViT-128 [7]	9.2	0.4	78.6
GFNet-Ti [34]	7.0	1.3	74.6
Mformer [17]	9.4	0.2	76.7
EfficientViT [1]	7.8	0.7	79.1
EdgeViT-XS [30]	6.7	1.1	77.5
MViT-S [29]	5.6	2.0	78.4
EdgeNext-S [27]	5.6	1.3	79.4
MViTv2-1.0 [19]	4.9	1.8	78.1
tiny-MOAT-1 [44]	5.1	1.2	78.3
MixFormer-B1 [3]	8	0.7	78.9
RepViT-M1.0 [38]	6.8	1.1	80.3
AFFNet [12]	5.5	1.5	79.8
IDAF	6.2	1.4	80.3
AFIDAF	6.5	1.5	**80.9**

4.2 Object Detection

Setting. Experiments on object detection are conducted using the MS-COCO 2017 [21] dataset, a widely-used benchmark for object detection, instance segmentation, and keypoint detection tasks. The dataset includes 118K training images, 5K validation images, and 20K test-dev images, covering 80 object categories annotated with bounding boxes, masks, and keypoints. The objects in this dataset are diverse and challenging, ranging from people and animals to vehicles and household items.

Table 2. Comparison of AFIDAF variants Object detection on MS-COCO 2017 dataset.

Model	Param(M)	mAP(%)
AFFNet-T [12]	3.0	25.3
AFIDAF-T	3.1	**25.4**
AFFNet [12]	5.6	28.4
IDAF	5.9	28.2
AFIDAF	6.2	**30.2**

Table 3. AFIDAF variants vs. other LW backbones Semantic segmentation on PASCAL VOC 2012 dataset.

Model	Params (M)	mIOU(%)
AFFNet-T [12]	3.5	77.8
MViTv2-0.75 [19]	6.2	75.1
AFIDAF-T	3.9	**79.6**
AFFNet [12]	6.9	80.5
EdgeNext [27]	6.5	80.2
IDAF	7.5	81.1
AFIDAF	7.8	**81.6**

Following the common practice in [12,27,29], we compare lightweight backbones, AFIDAF and AFIDAF-T, using the SSD [22] framework. We initialize the backbone with ImageNet-1K pre-trained weights and fine-tune the entire model on MS-COCO for 200 epochs with a 320 × 320 input resolution. The training uses a cosine learning rate scheduler with a base learning rate of 7e−4, a minimum learning rate of 7e−5, and AdamW optimizer (weight decay 0.05) with the Ssd Multibox loss function.

Results. As shown in Table 2, the detection models equipped with AFIDAF outperform other lightweight transformer-based detectors in terms of mean Average Precision (mAP). Specifically, AFIDAF surpasses the second-best AFFNet [12] by 1.8% in mAP. Consistently, AFIDAF-T edges out AFFNet-T by 0.1% in mAP with 0.1M more parameters.

4.3 Semantic Segmentation

Setting. We perform semantic segmentation experiments on the PASCAL VOC 2012 benchmark dataset [5]. This dataset, widely utilized for object recognition, detection, and segmentation tasks, comprises of over 11,000 images with pixel-level annotations across 20 object categories. It presents significant challenges due to the high variability in object appearances, occlusions, and clutter. Following common practices [27], we augment the dataset using MS-COCO 2017 [21], incorporating additional annotations and data to enhance our experiments.

We use the DeepLabv3 [2] framework for semantic segmentation with AFIDAF and AFIDAF-T backbones. Images are resized to 512 × 512, and models are initialized with ImageNet-1K pretrained weights. Models are trained for 50 epochs on the VOC dataset, using a cosine learning rate scheduler with a base rate of 5e−4, a minimum rate of 1e−6, and optimizer AdamW with a weight decay of 0.05. The loss function employed is cross-entropy loss.

Results. In Table 3, AFIDAF demonstrates superior performance compared to other lightweight networks for semantic segmentation. Specifically, AFIDAF achieves a mean Intersection over Union (mIoU) of 81.6%, surpassing the second-best lightweight network, AFFNet, by 1.1%. Additionally, AFIDAF-T exceeds AFFNet-T by 1.8% in mIoU.

Table 4. Comparison of HAFIDAF with middleweight networks on ImageNet-1K classification. HAFIDAF achieves competitive performance with fewer parameters.

Model	Params (M)	Flops (G)	Top-1 (%)	Top-5 (%)
Swin-T [24]	28	4.5	81.2	95.5
SpectFormer-XS [32]	20	4.0	80.2	94.7
PoolFormer-S12 [45]	12	1.8	77.2	-
PoolFormer-S24 [45]	21	3.4	80.3	-
GFNet-XS [34]	16	2.9	78.6	94.2
HAFIDAF	14.8	4.45	79.8	95.0

4.4 Experimental Evaluation of HAFIDAF

We conduct experiments to evaluate our proposed Hierarchical AFIDAF (HAFIDAF) model, comparing it with state-of-the-art vision transformers models across image classification, semantic segmentation, and object detection tasks.

Image Classification. Table 4 compares HAFIDAF with other middleweight networks on ImageNet-1K. Based on the Swin-T architecture, HAFIDAF reduces parameters by 47% (14.8M vs. 28M) while only decreasing Top-1 and Top-5 accuracy by 1.4% and 0.5%, respectively. This demonstrates HAFIDAF's efficiency in balancing model size and performance against recent Vision Transformers.

Object Detection. We evaluate HAFIDAF using the Cascade Mask R-CNN framework (Table 5). With consistent training settings across models, HAFIDAF achieves a 17% reduction in model size compared to Swin-T, with only a 1.6% drop in AP^{box}. This showcases HAFIDAF's ability to balance compression and accuracy in detection tasks.

Semantic Segmentation. Using the UperNet framework, we compare HAFIDAF and Swin-T on the Pascal VOC dataset (Table 6). HAFIDAF reduces parameters by 24% while improving all three accuracy metrics (mIoU, mAcc, and aAcc), highlighting its effectiveness in dense prediction tasks.

Table 5. Object detection performance on COCO dataset. HAFIDAF maintains competitive performance with significantly fewer parameters compared to larger models.

Model	Param (M)	AP^{box} (%)	$\text{AP}^{\text{box}}_{50}$ (%)	$\text{AP}^{\text{box}}_{75}$ (%)
R-50 [10]	82	46.3	64.3	50.5
DeiT-S [36]	80	48.0	67.2	51.7
Swin-T [24]	86	50.5	69.3	54.9
HAFIDAF	72	48.9	67.6	53.4

Table 6. Semantic segmentation performance on Pascal VOC 2012 dataset. HAFIDAF outperforms Swin-T baseline across all metrics with 24% fewer parameters.

Model	Param (M)	mIoU (%)	mAcc (%)	aAcc (%)
Swin-T [24]	60	71.1	77.9	93.4
HAFIDAF	**46**	**72.4**	**80.3**	**93.8**

5 Conclusion

We found that the channel direction filtering in AFFNet limited its performance and proposed to alternate an efficient image domain large kernel convolution approximation with AFFNet block. The dual domain feature extraction approach (AFIDAF) and its tiny version AFIDAT-T achieved consistent improvements over AFFNet and other state of the art lightweight networks in classification and downstream CV tasks. The hierarchical version HAFIDAT successfully compressed ViT benchmark Swin-T [24], reducing parameter size while maintaining performance in similar CV tasks.

Acknowledgments. The work was partially supported by NSF grants DMS-2151235, DMS-2219904, and a Qualcomm Gift Award.

References

1. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: lightweight multi-scale linear attention for high-resolution dense prediction. In: ICCV (2023)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Chen, Q., et al.: Mixformer: mixing features across windows and dimensions. In: CVPR (2022)
4. Dong, X., et al.: Cswin transformer: a general vision transformer backbone with cross-shaped windows. In: CVPR (2022)
5. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**, 98–136 (2015)
6. Fan, H., et al.: Multiscale vision transformers. In: ICCV (2021)
7. Graham, B., et al.: Levit: a vision transformer in convnet's clothing for faster inference. ICCV (2021)
8. Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B.: Adaptive fourier neural operators: efficient token mixers for transformers. ICLR (2022)
9. Guo, M., Lu, C., Liu, Z., Cheng, M., Hu, S.: Visual attention network. Comput. Vis. Media **9**(4), 733–752 (2023)

10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hou, Q., Lu, C., Cheng, M., Feng, J.: Conv2former: a simple transformer-style convnet for visual recognition. IEEE Trans PAMI (2024). https://doi.org/10.1109/TPAMI.2024.3401450
12. Huang, Z., Zhang, Z., Lan, C., Zha, Z.J., Lu, Y., Guo, B.: Adaptive frequency filters as efficient global token mixers. In: ICCV (2023)
13. Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., Fu, B.: Shuffle transformer: ethinking spatial shuffle for vision transformer. arXiv:2106.03650 (2021)
14. Kolesnikov, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. ICLR (2021)
15. LeCun, Y., Bengio, Y.: Convolutional Networks for Images, Speech, and Time Series. MIT press (1998)
16. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with Fourier transforms. Proc. 2022 Conf. North Amer. Chapt. Assoc. Comput. Ling.: Human Language Tech pp. 4296 – 4313 (2022)
17. Li, J., Leng, Y., Song, R., Liu, W., Li, Y., Du, Q.: Mformer: taming masked transformer for unsupervised spectral reconstruction. IEEE Trans. Geosci Remote Sensing **61**, 1–12 (2023)
18. Li, K., et al.: Uniformer: unifying convolution and self-attention for visual recognition. IEEE Trans PAMI **45**(10), 12581–12600 (2023)
19. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. CVPR (2022)
20. Li, Y., et al.: Rethinking vision transformers for MobileNet size and speed. In: ICCV (2023)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (2014)
22. Liu, W., et al.: Ssd: single shot multibox detector. ECCV (2016)
23. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. CVPR (2022)
24. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV (2021)
25. Lyu, J., Zhang, S., Qi, Y.Y., Xin, J.: Autoshufflenet: learning permutation matrices via an exact Lipschitz continuous penalty in deep convolutional neural networks. KDD (2020)
26. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: practical guidelines for efficient cnn architecture design. ECCV (2018)
27. Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S., Anwer, R., Khan, F.: Edgenext: efficiently amalgamated CNN-transformer architecture for mobile vision applications. CADL at ECCV (2022)
28. Mehta, R., Sivaswamy, J.: M-net: a convolutional neural network for deep brain structure segmentation. In: IEEE Intern. Symposium Biomed Imaging, pp. 437–440 (2017)
29. Mehta, S., Rastegari, M.: Mobilevit: lightweight, general-purpose, and mobile-friendly vision transformer. ICLR (2022)
30. Pan, J., et al.: EdgeViTs: competing light-weight cnns on mobile devices with vision transformers. In: ECCV (2022)
31. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with hilo attention. In: NeurIPS (2022)

32. Patro, B.N., Namboodiri, V.P., Agneeswaran, V.S.: Spectformer: frequency and attention is what you need in a vision transformer. arXiv preprint arXiv:2304.06446 (2023)
33. Qin, D., et al.: Mobilenetv4 - universal models for the mobile ecosystem. arXiv preprint arXiv:2404.10518 (2024)
34. Rao, Y., Zhao, W., Zhu, Z., Zhou, J., Lu, J.: Gfnet: global filter networks for visual recognition. IEEE Trans PAMI **45**(9), 10960–10973 (2023)
35. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**, 211–252 (2015)
36. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. ICML (2021)
37. Vaswani, A., et al.: Attention is all you need. NeurIPS **30** (2017)
38. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: RepVit: revisiting Mobile CNN from ViT Perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15909–15920 (2024)
39. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: co-designing and scaling convnets with masked autoencoders. CVPR (2023)
40. Wu, H., et al.: CvT: introducing convolutions to vision transformers. In: ICCV (2021)
41. Wu, S., Wu1, T., Tan, H., Guo, G.: Pale transformer: a general vision transformer backbone with pale-shaped attention. AAAI (2022)
42. Xie, E., Wang, W., Yu, Z., Alvarez, A.A.J.M., Luo, P.: Segformer: simple and efficient design for semantic segmentation with transformers. NeurIPS (2021)
43. Xu, W., Xu, Y., Chang, T., Tu, Z.: Co-scale conv-attentional image transformers. In: ICCV (2021)
44. Yang, C., et al.: Moat: alternating mobile convolution and attention brings strong vision models. arXiv:2210.01820 (2022)
45. Yu, W., et al.: Metaformer is actually what you need for vision. In: CVPR, pp. 10819–10829 (2022)
46. Yuan, L., et al.: Tokens-to-Token ViT: training Vision Transformers from Scratch on ImageNet. In: ICCV (2021)
47. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: CVPR (2017)