AFF-CAM: Adaptive Frequency Filtering based Channel Attention Module

DongWook Yang[0000−0002−3899−2987], Min-Kook Suh[0000−0002−1654−3689], and Seung-Woo Seo[0000−0003−4890−8563]

Seoul National University, Seoul, Korea
{ab3.yang, bluecdm, sseo}@snu.ac.kr

Abstract. Locality from bounded receptive fields is one of the biggest problems that needs to be solved in convolutional neural networks. Meanwhile, operating convolutions in frequency domain provides complementary viewpoint to this dilemma, as a point-wise update in frequency domain can globally modulate all input features involved in Discrete Cosine Transform. However, Discrete Cosine Transform concentrates majority of its information in a handful of coefficients in lower regions of frequency spectrum, often discarding other potentially useful frequency components, such as those of middle and high frequency spectrum. We believe valuable feature representations can be learned not only from lower frequency components, but also from such disregarded frequency distributions. In this paper, we propose a novel **A**daptive **F**requency **F**iltering based **C**hannel **A**ttention **M**odule (AFF-CAM), which exploits non-local characteristics of frequency domain and also adaptively learns the importance of different bands of frequency spectrum by modeling global cross-channel interactions, where each channel serves as a distinct frequency distribution. As a result, AFF-CAM is able to re-calibrate channel-wise feature responses and guide feature representations from spatial domain to reason over high-level, global context, which simply cannot be obtained from local kernels in spatial convolutions. Extensive experiments are conducted on ImageNet-1K classification and MS COCO detection benchmarks to validate our AFF-CAM. By effectively aggregating global information of various frequency spectrum from frequency domain with local information from spatial domain, our method achieves state-of-the-art results compared to other attention mechanisms.

1 Introduction

Recently, convolutional neural networks (CNN) have achieved remarkable progress in a broad range of vision tasks, e.g. image classification, object detection, and semantic segmentation, based on their powerful feature representation abilities. The success has mainly been fueled by strong prior by inductive bias and ability to model local relationship through large number of kernels in convolutional layers. To further enhance the performance of CNNs, recent researches have investigated to create networks that are *deeper* [1, 2], *wider* [3], and also to contain more *cardinality* [4, 5] by creatively stacking multiple convolutional layers. Even with the increase in performance with aforementioned attempts, there still exists a limitation of *locality* inherited in nature of CNNs that roots from localized receptive fields (RF) due to small kernels, e.g., 3x3 kernels [2] in most image-oriented tasks. Thus, it is of the utmost importance that we guide CNN to better extract global information, also known as *long-range* dependency. Theoretically, long-range dependencies can be acquired from deeper networks. Deeper networks allow the buildup of larger and more complex RFs because stacking multiple layers increases RFs linearly or exponentially. However, recent study [6] has proven that not all pixels in a RF contribute equally to an output unit's response and that effective RF only occupies a fraction of the full theoretical RF.

To efficiently and effectively implement non-local RFs to better acquire long-range dependencies, we introduce spectral transform theory, in particular Discrete Cosine Transform (DCT). We propose to adopt DCT in our network for the following reasons. First, as stated in spectral convolution theorem in Fourier Theory, updating a single value in frequency domain globally influences all the input features associated in Fourier Transform. We take advantage of this fact and enable CNN to implement the effect of having non-local RFs even from earlier layers that have localized RFs. Second, from 2D image perspective, DCT expresses a finite sequence of data points by the series of harmonic cosine functions wavering at distinct frequencies. In other words, DCT expresses the phenomena of an image in terms of different bands, e.g., low, middle, or high, of frequency components. For example, removing high frequency components blurs the image and eliminating low frequency components leaves us with edges. This indicates that by adequately modulating the amount of different frequency details through DCT, we are able to pick the most important frequency components and discard the rest. We link this to a well-known concept known as *attention* mechanism. Attention is a tool that permits the network to utilize the most relevant parts of a given input feature in a flexible manner. By fusing the characteristics of DCT and attention mechanism, we formulate the network to give distinct attention to different bands of frequency components. Although majority of the information is stored in just a few DCT coefficients, particularly those of lower frequencies, we believe useful information can be found not only in lower frequencies but also in middle or high frequency spectrum.

With aforementioned motivations, we introduce a novel **A**daptive **F**requency **F**iltering based **C**hannel **A**ttention **M**odule (AFF-CAM), that leverages non-local characteristics of frequency domain through DCT and formulates channel-wise attention map, which explores and learns the importance of distinct frequency distributions, to modulate the local feature representations from spatial domain. As depicted in Fig. 1, AFF-CAM is composed of three main sub-modules after going through DCT: i) **G**lobal **U**pdate **M**odule (GUM) that targets feature maps to acquire long-range dependencies with its non-local RFs. The effects of non-local RFs are implemented by 1x1 convolutional layers in frequency domain because, as mentioned above, point-wise update in frequency domain globally

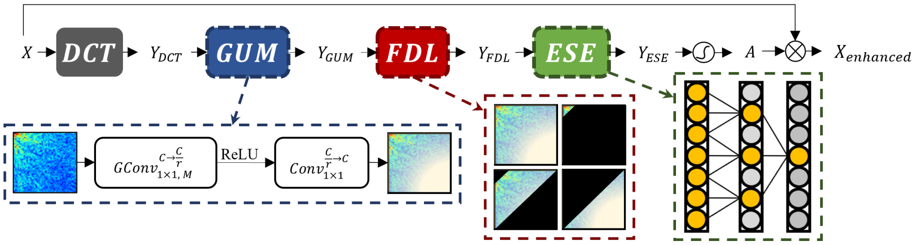
X
DCT
Y_{DCT}
GUM
Y_{GUM}
FDL
Y_{FDL}
ESE
Y_{ESE}
A
$X_{enhanced}$
$GConv^{C \to \frac{C}{r}}_{1\times1, M}$
ReLU
$Conv^{\frac{C}{r} \to C}_{1\times1}$

Fig. 1: **General overview of AFF-CAM.** AFF-CAM is composed of three sub-modules: i) Global Update Module (GUM), which gathers long-range dependencies with non-local receptive fields and globally updates input feature representations, ii) Freqeuncy Distribution Learner (FDL), which allows the network to learn the importance of different bands of frequency spectrum through mask filtering, and iii) Enhanced Squeeze-and-Excitation (ESE) that efficiently acquires global cross-channel interactions to re-weigh channel responses.

affects input features associated in DCT. ii) **F**requency **D**istribution **L**earner (FDL) that partitions input feature maps in frequency domain into M different distributions/bands of frequency spectrum with a set of learnable frequency filters so that various kinds of advantageous information can be aggregated from M different distributions. We adopt this idea because DCT concentrates majority of its information in a handful of coefficients, mainly lowest frequency (DC) component, often ignoring information from other frequency distributions (e.g. middle, and high). FcaNet [7], which most resembles our method, also exploits DCT to construct multi-spectral channel attention map. However, FcaNet only uses a few pre-selected (e.g. 1, 2, 4, 8, 16, or 32) DCT coefficients that are considered profitable. By doing so, FcaNet cannot flexibly learn richer feature representations beyond fixed frequency components. With FDL, our network is not limited to specific, prefixed frequency components but provies more adaptation to learn useful details from different frequency distributions. iii) **E**nhanced **S**queeze-and-**E**xication (ESE) that produces a channel descriptor by aggregating feature maps across their spectral dimensions ($height \times width$) by "*squeeze*" operation and outputs a collection of per-channel modulation weights via "*excitation*" operation. While squeeze operation was originally conducted with Global Average Pooling (GAP) in SENet [8], we utilize combination of GAP and Global Max Pooling (GMP) operations to extract richer feature representations like in CBAM [9]. Excitation operation is generally implemented with two fully-connected (FC) layers to capture non-linear cross-channel interactions and also to control model complexity through dimension reduction. However, according to ECANet [10], using mutliple FC layers is not only memory intensive but also dimension reduction in FC layers destroys the direct correspondence between channel and its weight. ECANet proposes to replace FC layers with a single lightweight 1D convolution with adaptive kernel size k and captures *local* cross-channel interaction. Inspired by this, ESE also adopts 1D convolution layer but enhances the idea by introducing **1D-D**ilated **C**onvolution **P**yramid (1D-DCP), which stacks multiple dilated [11] 1D convolutional layers with different dilation rates and successfully obtains *global* cross-channel interactions. This way, 1D-DCP is able to implement the effect of a FC layer with the model complexity of a 1D convolution.

Contributions. To recap, main contributions of our AFF-CAM can be summarized as follows:

1. From GUM, we are able to obtain long-range dependencies that conventional convolutional layers cannot acquire through simply stacking multiple layers.
2. From FDL, we are not limited to a few, pre-defined frequency components but able to adaptively learn richer feature representations from wide range of frequency distributions.
3. From ESE, we attain global cross-channel interactions with substantially low computational complexity.

2 Related Work

2.1 Attention

By proposing to model the importance of features, attention mechanism has been a promising tool in enhancing the performance of CNNs. Attention modules facilitate the networks to learn "*what*" and "*where*" to look in spatial and channel dimensions, respectively, by focusing on important features and suppressing non-useful details through activations. **SENet** [8] first introduces "*Squeeze-and-Excitation*" (SE) block that adaptively re-calibrates channel-wise feature responses by explicitly modeling inter-dependencies between channels. With channel attention, SE block facilitates network to realize "*what*" is more meaningful representation in a given image. **ECANet** [10] reinstates the importance of efficient channel attention proposed by SENet and proposes cheaper alternative. ECANet further explains that because of channel reduction ratio r in Multi-Layer Perceptron (MLP) of SE block, relationship mapping between channels is indirect and thus, non-optimal. To alleviate such problems, ECA block replaces MLP with 1D convolution with an adaptive kernel size, k. Inspired by this, our AFF-CAM also replaces MLP with 1D convolution with additional improvements ("*1D-Dilated Convolution Pyramid*" of Section 3.2). **CBAM** [9] enhances SE block with "*Channel Attention Module*" (CAM) by replacing GAP with combination of GAP and GMP to preserve richer contextual information, leading to finer channel attention. Our AFF-CAM's "*squeeze*" operation shares similar architecture, as frequency feature representations are compressed with combination of GAP and GMP pooling operations ("*Enhanced Squeeze-and-Excitation*" of Section 3.2). CBAM also introduces "*Spatial Attention Module*" (SAM) by using 2D convolutions of kernel size $k \times k$ to guide network "*where*" to focus, which is complementary to the channel attention. Our method does not follow up on spatial-wise attention because conducting DCT destroys the pixel-to-pixel correspondence. For example, whereas first pixel of a spatial image might represent a sky, first pixel in a DCT image represents low frequency details of a whole image, and not just a sky. Thus, re-weighing feature response of each pixel position in DCT produces meaningless results. **AANet** [12] mixes "*self-attention*"[13] with

SE block. AANet does not re-calibrate the channel-wise feature responses, but creates completely new feature maps through self-attention mechanism, which simultaneously exploits spatial and feature sub-spaces, and then goes through channel-wise compression. **GCNet** [14] mixes SENet with simplified Non-local Network (NLN) [15], where NLN aggregates query-specific global context to each query position to capture long-range spatial dependencies. Simply put, GCNet models global context of a single pixel by aggregating the relational information of every other pixels in a given image, which is computationally intensive. Our proposed method eliminates this computation burden by acquiring long-range dependencies through modulating different frequency components globally with lightweight 1×1 convolution operations ("*Global Update Module*" of Section 3.2).

2.2 Frequency Analysis

With eminent breakthroughs of CNN, there have been wide range of works [16, 17, 7, 18] that tried to incorporate frequency analysis, more specifically Fourier Transform, into deep learning frameworks. Because of the duality between convolution in spatial domain and element-wise multiplication in frequency domain, computing convolution in frequency domain has been considered as a replacement for vanilla convolution in spatial domain to solve an issue of heavy computational expense. Simply put, properties of Fourier Transform for CNN can be denoted as $F(x * y) = F(x) \odot F(y)$, where x and y represent two spatial signals (e.g. images) and $*$ and $\odot$ the operators of the convolution and the Hadamard product, respectively. Mathieu et al. [16] first carries out the convolutional operations in frequency domain with discrete Fourier Transform (DFT) and discovers that Fourier Transform of filters and the output gradients can be reused, leading to faster training and testing process. Additionally, operating convolutions in spectral domain can alleviate the problem of "*locality*" presented in vanilla convolutions due to their local receptive fields, as point-wise modulation in frequency domain globally updates input features associated in Fourier Transform. **FFC** [17] is inspired to capsulate both local context and long-range context with a local branch that conducts ordinary small-kernel convolution and a semi-global/global branch that manipulates image-level spectrum via DFT, respectively. However, FFC naively adds responses from local and semi-global/global branches. Concept of merging local and global information through frequency analysis is similar to our AFF-CAM. However, because outputs of spatial domain and frequency domain encode different levels of feature representation, our method does not simply add but utilizes output of frequency domain as a guidance to enhance low-level details captured in spatial domain via attention mechanism. **FcaNet** [7] utilizes DCT to construct multi-spectral channel attention. FcaNet adopts DCT, rather than DFT, for computational efficiency with its ability to pre-compute basis function (Eq.2) from real-valued cosine functions. The idea of updating channel features through DCT is similar to our AFF-CAM. However, N number of DCT weights/filters in FcaNet are hand-picked and fixed before training, thus cannot be learned and optimized. We believe useful information can be found and learned in different bands of frequencies. Therefore, we add learnable filters to fixed base filters to learn and provide more adaptation to select the frequency of interest beyond the fixed base filters ("*Frequency Distribution Learner*" of Section 3.2). **FNet** [18] replaces transformers' self-attention sublayers with Fourier Transform. While incorporating DFT to create attention maps is similar to our AFF-CAM, the way such attention maps are used differs. AFF-CAM utilizes attention maps to modulate the local descriptors of spatial feature representations to reason over high-level, global context. However, FNet disregards local information, but primarily considers long-range dependencies that is modeled via Fourier Transform.

3 Method

In this section, we demonstrate the core concepts of our proposed AFF-CAM. The main contributions are threefold: i) Global Update Module (GUM), which obtains long-range dependencies by globally updating feature representations associated in DCT, ii) Frequency Distribution Learner (FDL), which learns importance of different frequency distributions, and iii) Enhanced Squeeze-and-Excitation (ESE) that efficiently acquires global cross-channel interactions to re-calibrate channel-wise feature responses. The general overview of our proposed AFF-CAM is shown in Fig. 1.

3.1 Preliminaries: Discrete Cosine Transform

We begin by introducing DCT. DCT is a powerful tool used in field of digital signal processing for transforming a spatial-temporal signal or an image into spectral sub-bands of different importance. Simply put, it is a linear transformation of measurements in time/spatial domain to the frequency domain. DCT has the property that most of the visually significant information about a given image is concentrated and stored in just a few coefficients. For this reason, DCT is often used in image compression application. Even though all the properties can be extended to higher input dimensions, we constrain ourselves to the 2D DCT for simplicity. Like any Fourier-related transform, DCT expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes. The general equation for 2D DCT F for a given input feature map $f_{x,y} \in \mathbb{R}^{H\times W\times C}$, where $f_{x,y}$ represents the pixel value of $H \times W$ image at point (x, y), is defined as:

$$F_{u,v} = C(u)C(v)\sum_{x=0}^{W-1}\sum_{y=0}^{H-1} f_{x,y}B_{u,v}^{x,y} \tag{1}$$

$$\text{s.t.} \quad C(u) = \begin{cases} \frac{1}{\sqrt{W}} & \text{if } u = 0 \\ \sqrt{\frac{2}{W}} & \text{otherwise} \end{cases} \qquad C(v) = \begin{cases} \frac{1}{\sqrt{H}} & \text{if } v = 0 \\ \sqrt{\frac{2}{H}} & \text{otherwise} \end{cases}$$

$$B_{u,v}^{x,y} = cos(\frac{\pi u}{2W}(2x+1))cos(\frac{\pi v}{2H}(2y+1)) \tag{2}$$

where $B_{u,v}^{x,y}$ is a basis function of DCT. In terms of CNN, basis function can be regarded as filters/weights for convolution operations. Basis function can be pre-computed and simply looked up in DCT computation.

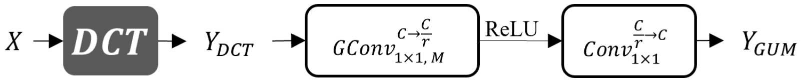
X
DCT
Y_{DCT}
$GConv_{1\times1,M}^{C\to\frac{C}{r}}$
ReLU
$Conv_{1\times1}^{\frac{C}{r}\to C}$
Y_{GUM}

Fig. 2: **Overview of GUM.**

3.2 Adaptive Frequency Filtering based Channel Attention Module (AFF-CAM)

Channel Attention. Channel attention mechanism attempts to assign different significance to the channels and reduce channel redundancy of given feature map by capturing inter-channel relationship. General equation for acquiring channel-wise attention map A, given a feature map f, can be denoted as:

$$A = \sigma(Network(compress(f)) \quad (3)$$

where σ is a sigmoid activation function, $compress$ is a operation to aggregate spatial information into a single global value, i.e., $\mathbb{R}^{H\times W\times C} \longmapsto \mathbb{R}^{1\times1\times C}$, and $Network$ is a mapping function, i.e., fully-connected layer or 1D convolutional layer. There have been studies [7, 18] that tried to acquire channel attention map using Fourier Transform. FcaNet is most similar to our proposed AFF-CAM, as it utilizes DCT to construct multi-spectral channel attention map. However, FcaNet exhibits three drawbacks: i) DCT feature representations are not updated globally using non-local RFs, ii) fixed number of "hand-picked" DCT filters that cannot be learned and optimized through training, and iii) large number of network parameters due to two fully-connected layers as $Network$ in Eq. 3. In the subsequent subsections, we discuss how each of our proposed submodules in AFF-CAM irons out such drawbacks.

Global Update Module (GUM). To alleviate the first issue, we propose GUM as shown by Fig. 2. Given a spatial feature map $X \in \mathbb{R}^{H\times W\times C}$, we first transform it into frequency domain by performing 2D DCT along the spatial dimensions:

$$Y_{DCT} = F[X] \in \mathbb{R}^{H\times W\times C} \quad (4)$$

where $F[\cdot]$ denotes the 2D DCT (Eq. 1). As Y_{DCT} represents the frequency spectrum of X, we then globally modulate the spectrum by performing 1x1 grouped convolution [19] with a group size of M ($GConv_{1\times1,M}^{c\to\frac{c}{r}}$), ReLU activation function, and 1x1 convolution operation ($Conv_{1\times1}^{\frac{c}{r}\to c}$):

$$Y_{GUM} = Conv_{1\times1}^{\frac{C}{r}\to C}(ReLU(GConv_{1\times1,M}^{C\to\frac{C}{r}}(Y_{DCT}))) \in \mathbb{R}^{H\times W\times C} \quad (5)$$

We control the number of parameters by adopting $C \to \frac{C}{r}$ and $\frac{C}{r} \to C$, which indicate channel reduction and channel restoration by ratio r, respectively. We set $r = 16$ for larger models, e.g., ResNet50 [2], and $r = 8$ for smaller models, e.g., ResNet-18/34 [2]. Global modulation can be implemented with 1x1 convolutional layers because point-wise update of spectrum Y_{DCT} can globally affect input feature representations X involved in DCT. We use grouped convolution with

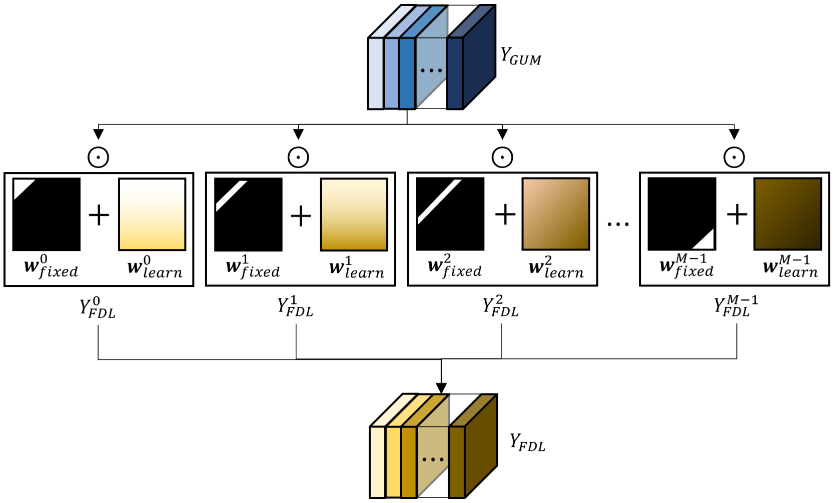
Y_{GUM}
⊙
⊙
⊙
⊙
$\mathbf{w}^0_{fixed}$ + $\mathbf{w}^0_{learn}$
$\mathbf{w}^1_{fixed}$ + $\mathbf{w}^1_{learn}$
$\mathbf{w}^2_{fixed}$ + $\mathbf{w}^2_{learn}$
...
$\mathbf{w}^{M-1}_{fixed}$ + $\mathbf{w}^{M-1}_{learn}$
Y^0_{FDL}
Y^1_{FDL}
Y^2_{FDL}
Y^{M-1}_{FDL}
Y_{FDL}

Fig. 3: **Detailed overview of FDL.**

a group size of M for the first convolution operation $GConv_{1\times1,M}^{C\to\frac{C}{r}}$, as we will partition frequency spectrum into M different bands of frequency components in Frequency Distribution Learner. This way, each of the M different frequency distributions can be globally updated separately. Second convolution operation $Conv_{1\times1}^{\frac{C}{r}\to C}$ is achieved by conventional 1×1 convolutional layer as sole purpose of this operation is to restore channel dimension back to C. As a result, Y_{GUM} is able to attain long-range dependencies through the effect of using non-local RFs.

Frequency Distribution Learner (FDL). To solve the second issue, we propose FDL as demonstrated by in Fig. 3. It has been established that spectral distribution of DCT is non-uniform and most of the energy are concentrated in just a few coefficients in low frequency area. While the general consensus is to utilize low frequency components to compress and represent an image, we believe useful details can be found and learned from other frequency distributions.

To carry out this task, we first construct M different binary fixed filters $[w^i_{fixed}]^M_{i=1}$, where $w^i_{fixed}\in\mathbb{R}^{H\times W\times1}$. As low-frequency components are placed in top-left corner and higher frequency components in the bottom-right corner of DCT spectrum, each of the M binary fixed filters are split diagonally to exploit different frequency distribution. We then add learnable filters $[w^i_{learn}]^M_{i=1}$, where $w^i_{learn}\in\mathbb{R}^{H\times W\times1}$, to the binary fixed filters to grant more access to select the frequency of interest beyond the fixed filters. By doing so, we obtain $w^i=[w^i_{fixed}+w^i_{learn})]^M_{i=1}$, in which we clip w^i_{learn} using Hyperbolic Tangent (tanh) to keep w^i_{learn} in a range of -1 to 1.

To apply each of obtained w^i to Y_{GUM}, we split Y_{GUM} into M parts along the channel dimension. Each of the split feature maps is denoted as:

$$Y_{GUM}=[Y^0_{GUM};Y^1_{GUM};\cdots,Y^{M-1}_{GUM}] \tag{6}$$

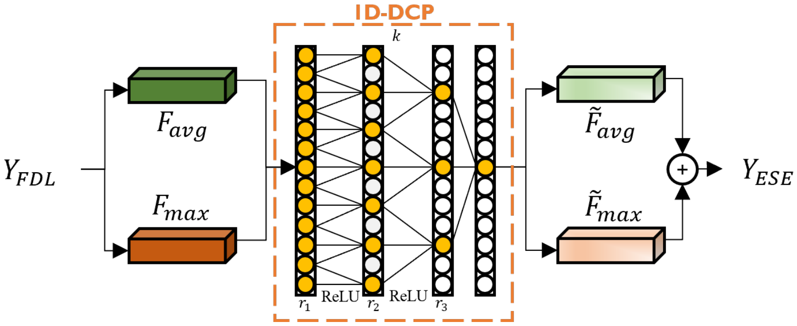

Fig. 4: **Detailed overview of ESE.**

where $Y^i_{GUM} \in \mathbb{R}^{H \times W \times \frac{C}{M}}, i \in \{0, 1, \cdots, M-1\}$ and C should be divisible by M. For each Y^i_{GUM}, corresponding w^i is assigned as:

$$Y^i_{FDL} = Y^i_{GUM} \odot w^i \in \mathbb{R}^{H \times W \times \frac{C}{M}} \tag{7}$$

where $\odot$ represents element-wise multiplication. Each of the Y^i_{FDL} is then concatenated to produce the final output:

$$Y_{FDL} = concat([Y^0_{FDL}, Y^1_{FDL}, \cdots, Y^{M-1}_{FDL}]) \in \mathbb{R}^{H \times W \times C} \tag{8}$$

As a result, each $Y^i_{FDL} \in \mathbb{R}^{H \times W \times \frac{C}{M}}$ holds different information from M distinct frequency distributions.

Enhanced Squeeze-and-Excitation (ESE) To make use of the information from different frequency distributions, we introduce ESE, as depicted by Fig. 4. ESE first aggregates, or so called "squeezes", different frequency bands (e.g. low, middle, high) information of Y_{FDL} along its spectral dimensions ($height \times width$) via GAP and GMP to produce two channel context descriptors: $F_{avg} \in \mathbb{R}^{1 \times 1 \times C}$ and $F_{max} \in \mathbb{R}^{1 \times 1 \times C}$.

After, F_{avg} and F_{max} are separately forwarded to a shared network to capture cross-channel interactions and produce enhanced descriptors: $\tilde{F}_{avg}$ and $\tilde{F}_{max}$. While two FC layers are most widely used to implement a shared network, they are computationally expensive, which is our third issue. ECANet replaces two FC layer with a single 1D convolution to avoid high computation and dimension reduction that occur in typical FC frameworks [8, 9, 12, 20, 14, 7]. While ECANet solves the computation issue, it can only capture *local* cross-channel interaction. So it can be viewed as a "partial" FC layer. We propose 1D-Dilated Convolution Pyramid (1D-DCP), as shown in orange dashed box in Fig. 4, which stacks multiple 1D convolutional layers with different dilation rates r_1, r_2, r_3 to capture *global* cross-channel interaction. Except for the last layer, each 1D convolutional layer is followed by ReLU activation function. Dilation convolution grants larger RF size by inserting spaces, i.e. zeros, between the kernel elements. For a 1D input signal $x[i]$, the output $x_{dilated}[i]$ of dilated convolution with filter $w[k]$ of length K is defined as $x_{dilated}[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k]$ where r is the dilation

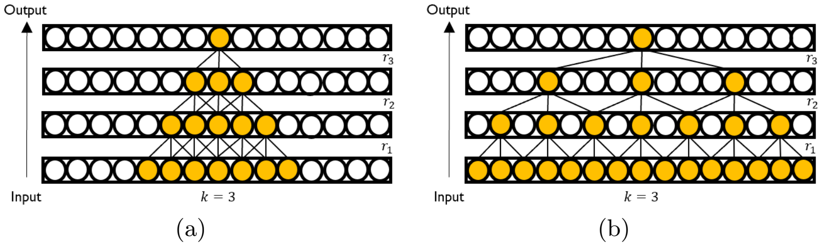

Fig. 5: **Illustrations of *conventional* 1D convolution and *dilated* convolution with kernel size of $k = 3$.** (a) Three-layer network using *conventional* 1D convolution operation with dilation rates $r_1 = r_2 = r_3 = 1$. (b) Three-layer network using *dilated* convolution operation with exponentially increasing dilation rates of $r_1 = 1, r_2 = 2, r_3 = 4$.

rate. It can be seen that when $r = 1$, dilated convolution is identical to the conventional 1D convolution operation. Fig. 5 demonstrates the difference between conventional convolution and dilated convolution operated on 1D signals. Like our proposed 1D-DCP, Fig. 5 is constructed with a 3-layer structure with kernel size $k = 3$ and dilation rates r_1, r_2, r_3. Fig. 5a shows that RFs of conventional 1D convolution increase linearly with the number of layers, resulting in RF of 7. However, Fig. 5b indicates that using exponentially increasing dilated rates, i.e., $r_1 = 1, r_2 = 2, r_3 = 4$, the RFs also increases exponentially to 15. Thus, by stacking multiple 1D convolutional layers with different dilation rates, our 1D-DCP can successfully capture global cross-channel interaction like a FC layer without its computation overhead. $\tilde{F}_{avg} \in \mathbb{R}^{1\times1\times C}$ and $\tilde{F}_{max} \in \mathbb{R}^{1\times1\times C}$ are then merged using element-wise summation to produce the output:

$$\tilde{F}_{avg} = DCP_{1D}(F_{avg}),\ \tilde{F}_{max} = DCP_{1D}(F_{max}) \tag{9}$$

$$Y_{ESE} = \tilde{F}_{avg} + \tilde{F}_{max} \in \mathbb{R}^{1\times1\times C} \tag{10}$$

Channel Attention Map. Obtained Y_{ESE} then goes through sigmoid activation function σ to provide final channel attention map A:

$$A = \sigma(Y_{ESE}) \tag{11}$$

It is clearly presented that our input spatial feature map X and resulting attention map A of AFF-CAM accommodate complementary feature information; X contains low-level details, whereas A encompasses high-level semantics. By enabling effective communication between those two feature representations, our proposed AFF-CAM can simultaneously capture short-range and long-range dependencies. AFF-CAM accomplishes this by multiplying input spatial feature representation X with acquired attention map A. This lets frequency information to be a guide and modulates spatial feature representation to reason over high-level, global context. In essence, the communication between X and A is formulated as:

$$X_{enhanced} = A \otimes X \in \mathbb{R}^{H\times W\times C} \tag{12}$$

Model	M	Top-1 (%)	Params (M)
ResNet50	-	75.44	25.56
A	1	76.87 (+1.43)	28.12
B	4	**77.62** (+2.18)	27.22
C	8	76.36 (+0.92)	27.12
D	16	76.16 (+0.72)	27.15

Table 1: **Ablation study on effectiveness of M.**

Model	Method	Top-1 (%)	Params (M)
B$_{GRC}$	**G**Conv-**R**eLU-**C**onv	**77.62**	27.22
B$_{CRC}$	**C**onv-**R**eLU-**C**onv	77.03	28.16
B$_{GRG}$	**G**Conv-**R**eLU-**G**Conv	75.98	26.27

Table 2: **Ablation study on GUM.** The results of utilizing different variations of *GConv* and *Conv*.

where $A \in \mathbb{R}^{1\times1\times C}$, $X \in \mathbb{R}^{H\times W\times C}$, and $\otimes$ denotes element-wise product.

4 Experiments

In this subsection, we evaluate our AFF-CAM on the widely used benchmarks: ImageNet-1k for image classification and MS COCO for object detection. We compare AFF-CAM with several state-of-the-art attention baselines built upon ResNet [2], including SENet, CBAM, GCNet, AANet, ECANet, FFC, and FcaNet.

4.1 Ablation Studies

We begin by conducting several ablation studies on ImageNet-1k dataset to empirically demonstrate the effectiveness of our network design with ResNet-50 as a baseline architecture. Initially, we experiment on how different number of partitions of frequency spectrum M in Eq. 6 affects our network in terms of accuracy and number of parameters in Table 1. For this experiment, we do not adopt our 1D-DCP sub-module into the network, as we solely want to see the effect of M. Instead, we use a single 1D convolutional layer like in ECANet. We test with M=1, 4, 8, and 16. Maximum value of M is 16 because of channel reduction ratio r and grouped convolution operation $GConv_{1\times1,M}^{C\to\frac{C}{r}}$ in Eq. 5, where $\frac{C}{r}$ needs to be divisible by M with $r = 16$. Even with added parameters from learnable filters w^i in Eq. 7, total number of parameters decreases when $M > 1$ because of the grouped convolution operation with a group size M in Eq. 5. We yield the best result of 77.62% when using $M = 4$, which significantly improves baseline ResNet50 by 2.18%. This indicates that commonly disregarded frequency distributions, such as those of middle or high frequencies, do provide meaningful feature representations and are worth giving attention to.

With obtained $M = 4$, we conduct the next experiment, where different variations of 1×1 convolution operations are adopted for GUM (Eq. 5). As demonstrated in Table 2, Model B$_{GRC}$ outputs the best result of 77.62%. This proves that i) separately updating M different frequency components is superior to updating the whole frequency spectrum (Model B$_{CRC}$), and ii) sole purpose of the second convolution operation is to restore channel dimension, thus depth-wise separable convolution is not needed (Model B$_{GRG}$).

Next, we analyze how using different dilation rates r_1, r_2, r_3 in 1D-DCP of ESE influences our network. For this experiment, we bring the best Model B$_{GRC}$ from above ablation study (Table 2) to be the baseline. Model B$^{1,1,1}_{GRC}$ introduces the idea of stacking multiple 1D convolutional layers on top of Model B$_{GRC}$ but

Model	r_1	r_2	r_3	Top-1 (%)	Params (M)
B_{GRC}	1	-	-	77.62	27.22
$\mathrm{B}_{GRC}^{1,1,1}$	1	1	1	77.77 (+0.15)	27.22
$\mathrm{B}_{GRC}^{1,2,4}$	1	2	4	77.98 (+0.36)	27.22
$\mathrm{B}_{GRC}^{1,3,9}$	1	3	9	**78.09** (+0.47)	27.22
$\mathrm{B}_{GRC}^{1,4,16}$	1	4	16	77.84 (+0.22)	27.22

Table 3: **Ablation study on 1D-DCT.** The effectiveness of different dilation rates r_1, r_2, r_3.

GUM	FDL	ESE	Top-1 (%)	Params (M)
-	-	-	75.44	25.56
		✓	75.91 (+0.47)	25.56
	✓	✓	76.46 (+1.02)	25.61
✓	✓	✓	78.09 (+2.65)	27.22

Table 4: **Ablation study on AFF-CAM.** The effectiveness of each sub-module.

does not use any dilations. Models $\mathrm{B}_{GRC}^{1,2,4}$, $\mathrm{B}_{GRC}^{1,3,9}$ and $\mathrm{B}_{GRC}^{1,4,16}$ apply exponentially increasing dilation rates. Each model's superscript indicates the applied dilation rates r_1, r_2, r_3 on to the baseline Model B_{GRC}. We generate the best result of 78.09% with Model $\mathrm{B}_{GRC}^{1,3,9}$, which improves baseline Model B_{GRC} by 0.47% without adding any trainable parameters. The result implies that stacking multiple 1D convolutions with different dilation rates enables the network to capture *global* cross-channel interactions just like a FC layer without its computation overhead.

Finally, with selected hyper-parameters $M = 4$ (Table 1) and $r_1 = 1, r_2 = 3, r_3 = 9$ (Table 3) from above ablation studies, we establish the strength of three sub-modules of our proposed AFF-CAM in Table 4. The result indicates that the largest performance gain of 1.63% (76.46% → 78.09%) comes from adding GUM, which enables the effect of having non-local RFs and acquires long-range dependencies. With biggest performance gain, GUM is also responsible for most of the added parameters in AFF-CAM. For future works, we plan on developing a more efficient module that obtains long-range dependencies. Second biggest performance gain of 0.55% (75.91% → 76.46%) comes from adding FDL, which exploits and gives attention to commonly discarded frequency distributions. This also matches our proposed motivation that profitable feature representations can be learned not only from low frequency distribution, where majority of the information is held, but also from commonly discarded frequency distributions.

4.2 Image Classification on ImageNet-1K

To evaluate the results of the proposed AFF-CAM framework on ImageNet, we employ three ResNet backbone architectures, e.g. ResNet-18, ResNet-34, and ResNet-50. We adopt the same data augmentation scheme as ResNet for training and apply single cropping with the size of 224x224 for testing. The optimizer is performed by Stochastic Gradient Descent (SGD) with momentum of 0.9 and weight decay of 1e-4. The learning rate starts with 0.1 and drops every 30 epochs. All models are trained for 90 epochs with mini-batch size of 1024 for smaller models (e.g. ResNet-18/34) and 512 for bigger models (e.g. ResNet-50) on each of the 4 A100-SXM GPUs. Table 5 summarizes the experimental results.

When using **ResNet-18** as backbone architecture, AFF-CAM outperforms the baseline ResNet-18 by 1.12% and surpasses current state-of-the-art CBAM by 0.79% for Top-1 accuracy. When using **ResNet-34** as backbone architecture,

Method	Backbone	Top-1 (%)	Top-5 (%)	Params (M)	FLOPs (G)
ResNet [2]	ResNet-18	70.40	89.45	11.69	1.814
SENet [8]		70.59	89.78	11.78	1.814
CBAM [9]		70.73	89.91	11.78	1.815
AFF-CAM		**71.52**	**90.36**	11.84	1.84
ResNet [2]	ResNet-34	73.31	91.40	21.80	3.66
SENet [8]		73.87	91.65	21.95	3.66
CBAM [9]		74.01	91.76	21.96	3.67
AANet [12]		74.70	92.00	20.70	3.56
ECANet [10]		74.21	91.83	21.80	3.68
FcaNet† [7]		74.29	91.92	21.95	3.68
AFF-CAM		**74.88**	**92.27**	22.06	3.71
ResNet [2]	ResNet-50	75.44	92.50	25.56	3.86
SENet [8]		76.86	93.30	28.09	3.86
CBAM [9]		77.34	93.69	28.09	3.86
GCNet [14]		77.70	93.66	28.11	4.13
AANet [12]		77.70	93.80	25.80	4.15
ECANet [10]		77.48	93.68	25.56	3.86
FFC [17]		77.80	-	27.70	4.50
FcaNet† [7]		77.29	93.67	28.07	4.13
AFF-CAM		**78.09**	**93.82**	27.22	4.39

*All results are reproduced with the same training settings.

FcaNet† is reproduced as official code utilizes different training strategies.

Table 5: **Classification results on ImageNet-1K dataset.** The best Top-1/5 accuracy scores across all baselines are written in bold.

AFF-CAM outperforms the baseline ResNet-34 by 1.57% and surpasses current state-of-the-art AANet by 0.18% for Top-1 accuracy. When using **ResNet-50**as backbone architecture, AFF-CAM outperforms the baseline ResNet-50 by 2.65% and surpasses current state-of-the-art FFC by 0.29% for Top-1 accuracy.

4.3 Object Detection on MS COCO

In this subsection, we evaluate our AFF-CAM framework on object detection task to verify its general applicability across different tasks. We utilize Faster R-CNN [21] as baseline detector and ResNet-50 with Feature Pyramid Network (FPN) [22] as backbone architecture. For training implementation, we adopt MMDetection [23] toolkit and use its default settings as choice of hyperparameters. Optimizer is executed with SGD with momentum of 0.9 and weight decay of 1e-4. The learning rate is initialized to 0.01 and drops by the factor of 10 at 8th and 11th epochs. All models are trained for 12 epochs with mini-batch size of 4 on each of the 4 A100-SXM GPUs. As shown in Table 6, our AFF-CAM framework proves its generalization ability. Without bells and whistles, our AFF-CAM outperforms baseline ResNet-50 by 3.4% and surpasses current state-of-the-art FcaNet by 0.8% for Average Precision (AP).

5 Conclusions

In this paper, we propose a novel AFF-CAM that effectively explores the details of different frequency bands through DCT. While most information is stored

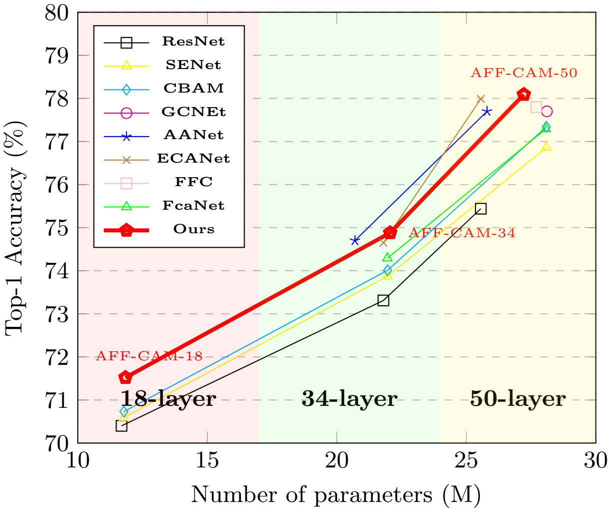

Fig. 6: **ImageNet-1K Top-1 Accuracy vs. Model Complexity.**

Method	Backbone	Detector	AP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
ResNet [2]			36.4	58.2	39.2	21.8	40.0	46.2
SENet [8]			37.7	60.1	40.9	22.9	41.9	48.2
ECANet [10]	ResNet-50	Faster-RCNN [21]	38.0	60.6	40.9	23.4	42.1	48.0
FcaNet [7]			39.0	60.9	42.3	23.0	42.9	49.9
AFF-CAM			**39.8**	60.7	43.6	22.8	42.4	51.0

*All results are reproduced with the same training settings.

Table 6: **Object detection results on MS COCO val 2017 dataset.** The best Average Precision score is written in bold.

in lower DCT coefficients, our method exploits other discarded frequency spectrum and adaptively re-calibrates channel-wise feature responses by efficiently modeling global inter-dependencies between channels. Furthermore, our method takes advantage of the fact that point-wise update in frequency domain globally affects input features associated in DCT. As a result, our method is able to implement the ensemble of local and non-local receptive fields in a single unit. Comprehensive experiments are conducted on ImageNet-1K classification and MS COCO detection datasets to demonstrate the applicability of AFF-CAM across different architectures, as well as different tasks. The results display consistent performance improvements that are clearly attributed to our proposed motivations.

Acknowledgements This research was supported by the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2022 (No.915034201).

References

1. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1–9
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
3. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1251–1258
5. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1492–1500
6. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. (2016) 4905–4913
7. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 783–792
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7132–7141
9. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). (2018) 3–19
10. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks (2020)
11. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
12. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 3286–3295
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
14. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2019) 0–0
15. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7794–7803
16. Mathieu, M., Henaff, M., LeCun, Y.: Fast training of convolutional networks through ffts. arXiv preprint arXiv:1312.5851 (2013)
17. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33** (2020)
18. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv:2105.03824 (2021)

19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
20. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D.: Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. arXiv preprint arXiv:2105.14447 (2021)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
22. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2117–2125
23. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)