

Frequency Attention for Knowledge Distillation

Cuong Pham¹ Van-Anh Nguyen¹ Trung Le¹ Dinh Phung¹
Gustavo Carneiro² Thanh-Toan Do¹

¹Department of Data Science and AI, Monash University, Australia

²Centre for Vision, Speech and Signal Processing, University of Surrey, United Kingdom

Abstract

Knowledge distillation is an attractive approach for learning compact deep neural networks, which learns a lightweight student model by distilling knowledge from a complex teacher model. Attention-based knowledge distillation is a specific form of intermediate feature-based knowledge distillation that uses attention mechanisms to encourage the student to better mimic the teacher. However, most of the previous attention-based distillation approaches perform attention in the spatial domain, which primarily affects local regions in the input image. This may not be sufficient when we need to capture the broader context or global information necessary for effective knowledge transfer. In frequency domain, since each frequency is determined from all pixels of the image in spatial domain, it can contain global information about the image. Inspired by the benefits of the frequency domain, we propose a novel module that functions as an attention mechanism in the frequency domain. The module consists of a learnable global filter that can adjust the frequencies of student's features under the guidance of the teacher's features, which encourages the student's features to have patterns similar to the teacher's features. We then propose an enhanced knowledge review-based distillation model by leveraging the proposed frequency attention module. The extensive experiments with various teacher and student architectures on image classification and object detection benchmark datasets show that the proposed approach outperforms other knowledge distillation methods.

1. Introduction

Convolutional Neural Networks (CNNs) have been widely applied and achieved a myriad of successes in various computer vision tasks, such as image classification, object detection, and image segmentation. However, these CNNs often require expensive memory and computation resources, making them unsuitable for applications with lim-

ited resources. Different approaches have been proposed to learn efficient deep neural networks, such as pruning [5, 10, 16], knowledge distillation [6, 7, 19, 23], and quantization [8, 20, 35, 37]. Among them, knowledge distillation (KD) is an attractive approach to reduce the computational cost of CNNs. In knowledge distillation, a smaller student network is trained to mimic the behaviour of a larger teacher network.

Different approaches have been proposed for knowledge distillation [2, 6, 7, 11, 19, 23, 30, 36]. Among them, intermediate feature-based KD is a popular approach because it is flexible to design different distillation mechanisms such as layer to layer distillation [6, 11, 23] and layer fusion distillation [19]. Attention-based KD [9, 11, 26, 28] is a specific form of intermediate feature-based knowledge distillation. In those works, the attention is performed in the spatial domain and they use attention maps to help the student to focus on the most informative information from the teacher. However, in [9, 11, 26, 28] each value of the attention map is calculated from a local region of the input feature map. This focus of the local regions may not be sufficient to effectively transfer knowledge from teacher model to student model in knowledge distillation when we need to capture the broader context or global information necessary for effective knowledge transfer.

Our goal is to encourage student model to capture both detailed and higher-level information such as object parts from the teacher model. This can be accomplished by processing the student's features in the frequency domain instead of the spatial domain. The frequency domain is useful for understanding images with repetitive or periodic patterns that may be difficult to discover using traditional spatial domain techniques. By capturing the intensity changes and patterns in the image, the frequency domain can identify different regions associated with objects, and each frequency could correspond to some specific structures, e.g., high frequencies correspond to large changes in image intensity over a short pixel distance (e.g., edges).

With the above benefits of the frequency domain, we propose a Frequency Attention Module (FAM), which has a

learnable global filter in the frequency domain. The global filter can be seen as a form of attention in the frequency domain, which can adjust the frequency of student’s feature maps. We then invert the attending features in frequency domain back to the spatial domain and minimize them with the teacher’s features. By updating the parameters of the learnable filter based on the guidance of the teacher, we can encourage the transformed student’s features to have similar patterns as the teacher’s features.

Given the proposed frequency-based attention module, we propose two enhanced architectures for layer-to-layer [6, 11, 23] and knowledge review distillation [19]. We extensively demonstrate the effectiveness of our proposed method with various teacher and student architectures on benchmark datasets for image classification and object detection. The experimental results show that the proposed approach outperforms other knowledge distillation methods. In summary, our contributions are:

- We propose a novel module, which is our main contribution in which we explore Fourier frequency domain for knowledge distillation. The module consists of a learnable global filter that can adjust frequency of the student’s features, which encourages student’s features to mimic patterns from the teacher’s features.
- We propose an enhanced layer-to-layer knowledge distillation model and an enhanced knowledge review-based distillation model by leveraging the proposed FAM module.
- Our method outperforms other knowledge distillation methods for classification on CIFAR-100 and ImageNet datasets and object detection on MS COCO dataset.

2. Related work

Knowledge Distillation (KD) has received substantial attention recently due to its versatility in various applications. In KD, the student model can benefit from the guidance of various forms from the teacher model to achieve better performance. This could be soft logit-based distillation [7, 36], relation-based distillation [2, 17, 30], or intermediate feature-based distillation [6, 11, 19, 23]. Among them, the feature-based knowledge distillation allows flexibility in designing distillation mechanisms. Particularly, in FitNets [23] given a student layer (guided layer) and a teacher layer (hint layer), the authors minimize the L_2 distance between the transformed student’s features and the teacher’s features. Following FitNets, AT [11], PKT [18], and SP [31] transfer knowledge through activation maps, feature distributions, and pairwise similarities, respectively. In OFD [6], the authors propose margin ReLU applied on

teacher’s feature maps to select information used for distillation. In [19], the authors introduce the review mechanism to enrich student features. They show that lower-level features of teacher are useful in supervising the higher-level features of student. They propose to fuse different levels of student features before mimicking teacher knowledge.

In [9, 11, 26, 28], the attention is performed in the spatial domain and they use attention maps to help the student focus on the most informative information from the teacher. Specifically, spatial attention maps in AT [11] can be computed using the sum of absolute values across the channel dimension. AFD [9] also transfers knowledge from teacher to student through spatial attention maps that are computed through channel-wise average pooling layer. They then maximize the similarity between attention maps of student’s features and the attention maps of teacher’s features. Meanwhile, [28] computes the spatial attention maps using average pooling and fully connected layers. However, with the attention in the spatial domain used [9, 11, 26, 28], weights of the attention map are usually calculated from local regions of the feature maps. The attention weights (i.e., values in the attention map) indicate the importance of the corresponding local regions. Due to its local property, a change in a value of the attention map (in backpropagation) only affects the corresponding local region.

Fourier frequency domain and attention in the frequency domain. In digital image processing, Fourier frequency domain represents an image with a set of sinusoidal waves, with each wave representing a different level of intensity in the whole image. The frequency domain is a helpful way to understand images that have repetitive or periodic patterns [3]. It is more effective than traditional spatial domain techniques in capturing geometric structures that are difficult to extract. By capturing the intensity changes in the image, the frequency domain can identify distinct regions that are associated with objects.

Each frequency in the frequency domain is determined by all the pixels in the image in the spatial domain. Frequencies can correspond to particular structures in the spatial domain. For instance, high frequencies correspond to significant changes in image intensity over a small distance between pixels, such as edges. Therefore, focusing on the frequency domain can be seen as a form of global attention. Meanwhile, attention in the spatial domain [9, 11, 28] primarily affects local regions in the input feature map, which may be insufficient for capturing the global structure of the feature map. By contrast, attention in the frequency domain can be especially useful for identifying global information or geometric structures of the feature map that may be difficult to detect using traditional spatial domain techniques. A change in a frequency of the attention frequency map can impact the entire input feature, compared to the effect on

the local regions when changing a value in attention map in the spatial domain.

In this work, we explore the Fourier frequency domain for the knowledge distillation problem. We propose a frequency attention module (FAM) that has a learnable global filter, which acts as an attention in the frequency domain. Based on the guidance from the teacher, FAM will encourage the student’s features to have similar patterns as teacher’s features.

3. Proposed method

This section first details the frequency attention module (FAM) that encourages the student to better mimic the teacher. We then present our design to integrate the FAM module into two popular knowledge distillation mechanisms, i.e., layer-to-layer feature-based distillation [6] and knowledge review-based distillation [19].

3.1. Frequency attention module

As shown in Figure 1, the FAM module consists of global and local branches. Specifically, given a feature map X with a dimension of $C_{in} \times H \times W$, in the global branch we first transform it into the frequency domain via Fast Fourier Transform (FFT). Here the FFT is applied to each channel separately. For the i^{th} channel X_i of the feature map X , the 2-D discrete FFT of X_i denoted by \mathcal{X}_i is expressed as:

$$\mathcal{X}_i(u, v) = \sum_{k=0}^{H-1} \sum_{l=0}^{W-1} X_i(k, l) e^{-i2\pi(\frac{uk}{H} + \frac{vl}{W})}. \quad (1)$$

To adjust the frequencies of \mathcal{X}_i , we apply a learnable global filter K which can be seen as a form of attention on \mathcal{X}_i .

Global filtering. It is worth noting that in feature distillation, we want the feature map resulting from the FAM module to have the same dimension as the dimension of a given teacher feature map where the knowledge will be distilled. Therefore, we design the global filter K with the dimension of $C_{out} \times C_{in} \times H \times W$, where C_{out} is the number of channels of the teacher’s feature map. Each kernel in the global filter K has the same size as the 3D input tensor \mathcal{X} with the size $C_{in} \times H \times W$. This kernel performs element-wise multiplied with the 3D input tensor \mathcal{X} , resulting in a 3D feature map with the same size as the input feature map. Next, the 3D frequency feature maps of the output are then summed up, (i.e., sum-pooling in each $C_{in} \times 1 \times 1$ block), resulting in a 2D output with the size $H \times W$. The above operation is performed for C_{out} kernels of the global filter, resulting in a 3D feature map with a size of $C_{out} \times H \times W$ as the output.

It is worth noting that the proposed filter acts in the frequency domain. Each frequency in the frequency domain is determined by all the pixels in the spatial domain; hence, although each element of each kernel attends to a particular frequency, the filter still achieves the global effects.

After that, we further suppress low frequencies, which encourages the student to de-focus from the non-salient regions. To this end, we add a high pass filter (HPF) after the learnable global filter to eliminate part of the lowest frequency components. The HPF is applied to each channel separately. Specifically, for each channel, we adopt the ideal HPF, which suppresses 1 percent of the lowest frequencies.

We then transform the frequency domain back to the spatial domain via the inverse Fast Fourier Transform (IFFT). Given $\bar{\mathcal{X}}$ which is the frequency feature map after the HPF, for the i^{th} channel $\bar{\mathcal{X}}_i$ of the frequency feature map $\bar{\mathcal{X}}$, the 2-D IFFT of $\bar{\mathcal{X}}_i$ denoted by \bar{X}_i is expressed as:

$$\bar{X}_i(k, l) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \bar{\mathcal{X}}_i(u, v) e^{i2\pi(\frac{uk}{H} + \frac{vl}{W})}. \quad (2)$$

Formally, let $g(\mathcal{X}, K)$ be the output of the global filtering as above, h be the high pass filter, \mathbb{F} and \mathbb{F}^{-1} be the FFT and the inverse IFFT, respectively, the output of the global branch is calculated as:

$$\mathcal{F}_{global}(X) = \mathbb{F}^{-1}(h(g(\mathbb{F}(X), K))), \quad (3)$$

where \mathbb{F} , h , \mathbb{F}^{-1} are applied in a channel-wise fashion.

The FAM module also consists of a local branch, which is a 1×1 convolutional layer in the spatial domain. This layer aims to leverage the information of features in the spatial domain. Let $\mathcal{F}_{local}(X)$ be the output of the local branch, and the output of the frequency attention module is calculated as below:

$$\mathcal{F}_{out} = \gamma_1 * \mathcal{F}_{global} + \gamma_2 * \mathcal{F}_{local}, \quad (4)$$

where γ_1 and γ_2 are the learnable weighting parameters of the global and local branches, respectively.

Computational complexity of the FAM module. The global branch comprises a fast Fourier transform (FFT), an inverse fast Fourier transform (IFFT), a global filter, and a high pass filter (HPF).

The complexity of the FFT of an image with dimensions $H \times W$ is $\mathcal{O}(HW \log(HW))$. Similarly, the complexity of the inverse fast Fourier transform (IFFT) of a frequency image with dimensions $H \times W$ is $\mathcal{O}(HW \log(HW))$. Therefore, the complexities of the FFT, global filter, HPF, and IFFT components in the FAM module are $\mathcal{O}(C_{in}HW \log(HW))$, $\mathcal{O}(C_{out}C_{in}HW)$, $\mathcal{O}(C_{out}HW)$, and $\mathcal{O}(C_{out}HW \log(HW))$, respectively.

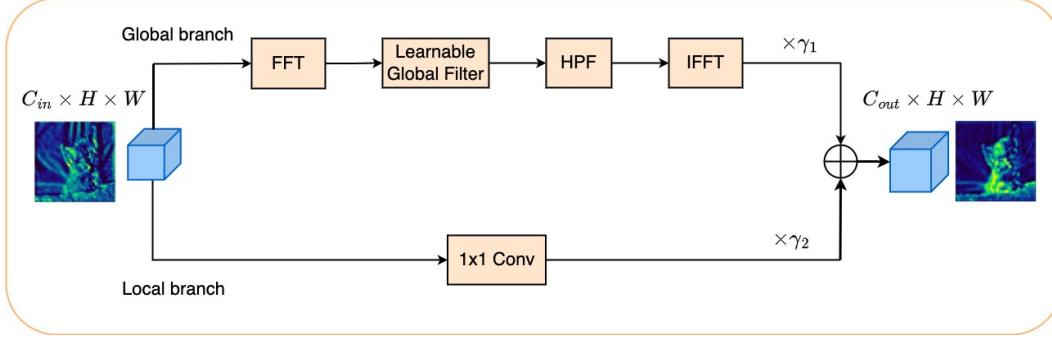


Figure 1. Fourier Frequency Attention Module. HPF stands for a high pass filter. In the global branch, the input student’s feature map is transformed to the frequency domain using the FFT. The frequency is then adjusted by a learnable global filter. A high pass filter is then applied to the adjusted frequency map to filter out lowest frequencies. The local branch consists of a 1×1 convolutional layer in the spatial domain. The outputs of the global and local branches are added and the resulting feature map is compared with the teacher’s feature map. γ_1 and γ_2 are the learnable weighting parameters of the global and local branches, respectively.

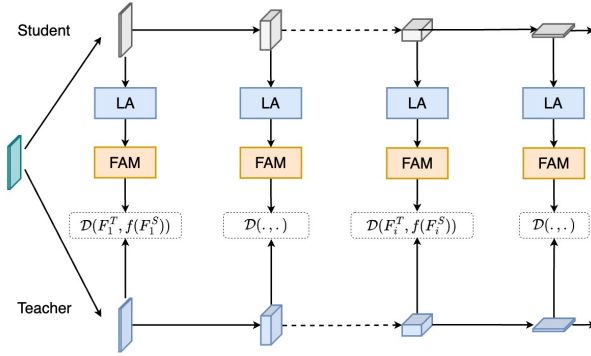


Figure 2. The proposed enhanced layer-to-layer knowledge distillation. LA is the local attention and FAM is the proposed frequency attention module. \mathcal{D} is the distance function. F^T and F^S represent the feature maps of teacher and student, respectively.

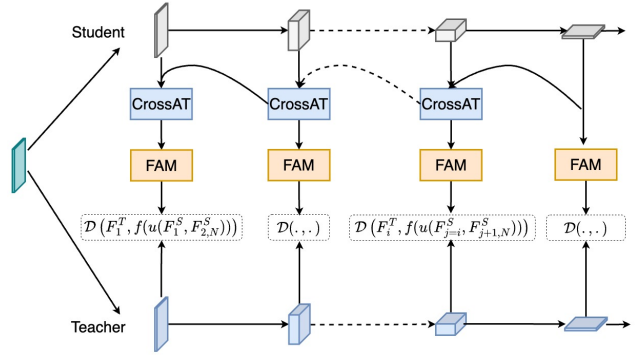


Figure 3. The proposed enhanced knowledge review distillation. CrossAT is the cross attention and FAM is the proposed frequency attention module. \mathcal{D} is the distance function. F^T and F^S represent the feature maps of teacher and student, respectively.

The FAM module also consists of a local branch, which is a 1×1 convolutional layer in the spatial domain. The local branch has the complexity of $\mathcal{O}(C_{out}C_{in}HW)$. Overall, the FAM module has the complexity of $\mathcal{O}(C_{out}C_{in}HW)$.

3.2. Applying FAM to knowledge distillation

3.2.1 Layer-to-layer intermediate feature-based knowledge distillation

Let \mathcal{I} be the selected layer indices from the teacher for intermediate feature-based distillation. The layer-to-layer knowledge distillation loss is defined as

$$\mathcal{L}_{feat} = \sum_{i \in \mathcal{I}} \mathcal{D}(F_i^T, f(F_j^S)), \quad (5)$$

where F_j^S is the feature map from the j^{th} layer of the student selected for receiving the knowledge from the feature

map F_i^T from i^{th} layer of the teacher; f is a transformation applied on the student’s feature map. In our work, f is the FAM module. \mathcal{D} is a distance function. In this work, we use L_2 distance as the distance function. It is worth noting the teacher is fixed in our framework, i.e., there are no transformations applied on teacher’s feature maps.

In order to make FAM to better mimic teacher, we find that it would be beneficial to also enhance local structures in the spatial domain. To this end, we place an attention layer after student’s feature maps before feeding it through the FAM module, as shown in Figure 2. To avoid increasing model complexity, we use local self-attention (LA) layer introduced by [21]. In LA, the self-attention is applied only to a small neighbourhood around each position.

3.2.2 Knowledge review distillation

We also integrate the FAM module into the knowledge review distillation mechanism [19], as shown in Figure 3.

In [19], the authors propose a knowledge review mechanism that uses teacher’s low-level features to supervise deeper student’s features. They fuse different levels of the student’s features before mimicking knowledge from the teacher. In knowledge review mechanism [19], the distillation loss is defined as follows:

$$\mathcal{L}_{feat} = \mathcal{D}(F_M^T, f(F_N^S)) + \sum_{i=M-1}^1 \mathcal{D}(F_i^T, f(u(F_{j=i}^S, F_{j+1,N}^S))), \quad (6)$$

where M and N are the numbers of selected intermediate layers of teacher and student used for knowledge distillation. We note that in intermediate feature-based KD, the student and the teacher models are often divided into stages. The number of stages is the same for the teacher and the student, i.e., $M = N$. The last layers in each stage are used for distillation. $u(\cdot, \cdot)$ is a fusion function that recursively fuses student features. $F_{j+1,N}^S$ denotes the fusion of features from F_{j+1}^S to F_N^S ; $u(F_{j=i}^S, F_{j+1,N}^S) = u(F_{j=i}^S, u(F_{j+1}^S, F_{j+2,N}^S))$; f is the FAM module.

In [19], $u(\cdot, \cdot)$ is an attention-based fusion (ABF [19]) function that learns two attention maps for two inputs and uses attention maps to aggregate two inputs. In this work, we hypothesize that the similarities between teacher and student features are often higher when they are at the same level (i.e., same stage). It could mean that a student’s feature of the same level as the teacher’s feature has a more important role than others. Emphasizing the importance of those features could be more beneficial for knowledge distillation. Therefore, instead of using the ABF attention used in [19], we want an attention mechanism that directly emphasizes the importance of the student’s feature map that is at the same level as the teacher’s feature map. To achieve this goal, we propose using cross attention [32] in which the low-level feature map is considered as the value and key and the high (fused) feature map is considered as the query when fusing student’s feature maps at different levels. Specifically, let $F' = u(F_{j+1}^S, F_{j+2,N}^S)$

$$u(F_j^S, F') = \text{softmax}((\mathbf{W}_Q F')(\mathbf{W}_K F_j^S)^T) \mathbf{W}_V F_j^S, \quad (7)$$

where \mathbf{W}_Q , \mathbf{W}_K , and \mathbf{W}_V represent learnable parameters for query, key, and value, respectively. In summary, compared to [19], firstly, to emphasize the importance of the student’s feature map that is at the same level as the teacher’s feature map, our enhanced KD review architecture uses cross attention instead of ABF [19]. Then, we feed the output of cross attention to the FAM module to adjust frequencies before computing the distance function \mathcal{D} .

The overall loss consists of the task loss (i.e., cross-entropy loss for classification task) and the feature distillation loss:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \mathcal{L}_{feat} \quad (8)$$

4. Experiments

4.1. Experimental setup

Datasets. We evaluate our approach on CIFAR-100 [12] and ImageNet [24] datasets for image classification task, and COCO dataset [14] for object detection task. The CIFAR-100 dataset consists of 60,000 images for 100 classes, in which, 50,000 and 10,000 images are used for training and validation sets, respectively. ImageNet is a challenging dataset with 1000 classes. This dataset contains 1.2 million images for training and 50,000 images for validation, which is used as a test set in our experiments. For object detection task, COCO is a standard dataset with multiple objects in an image. In total, this dataset contains 1.5 million object instances of 80 object categories in 118,000 training and 5,000 validation images.

Implementation details. We apply our method across various teacher-student architecture pairs, as shown in Table 1, Table 2, Table 3, and Table 4. For a fair comparison, we do experiments on standard teacher/student pairs following other distillation methods [2, 19, 30, 36] and base on the public distiller code-base [36]. This includes the distillation when teachers and students are in the same architecture and in different architectures. For training, we use the standard training procedure following [19, 36] and pre-trained teachers in all settings for both classification and object detection tasks. We employ L_2 distance as a distance function \mathcal{D} when calculating the \mathcal{L}_{feat} losses (Eq. (5) and Eq. (6)). The implementation details for CIFAR-100, ImageNet, and MS-COCO datasets and the values of the hyper-parameter α (Eq. 8) for each teacher/student pair are provided in supplementary materials due to page limit.

4.2. Comparison with the state of the art

4.2.1 Image classification

Comparative results on CIFAR-100. We present top-1 classification accuracy on the CIFAR-100 by various teacher-student pairs, both from the same network family (Table 1) and from the different network family (Table 2). The selected networks comprise ResNet [4], WideResNet [34], ShuffleNet [15], MobileNetV2 [25], and VGG [29]. The results of competitors are cited from [2, 19, 36].

Overall, our method FAM-KD (review) consistently outperforms all compared methods in all settings. In some cases, i.e., WRN-40-2/WRN-16-2, ResNet110/ResNet32, WRN-40-2/ShuffleNet-V1, students’ performance even surpasses the teachers.

Regarding layer-to-layer distillation, our method FAM-KD (layer-to-layer) outperforms all other methods belonging to the same category. Our method consistently outperforms the most competitor WCoRD [2] on all settings.

Teacher	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet32x4
Student	WRN-16-2	WRN-40-1	ResNet20	ResNet32	ResNet8x4
Teacher	75.61	75.61	72.34	74.31	79.42
Student	73.26	71.98	69.06	71.14	72.50
Soft logit-based distillation					
KD [7]	74.92	73.54	70.66	73.08	73.33
DKD [36]	76.24	74.81	71.97	74.11	76.32
Layer to layer-based distillation					
FITNET [23]	73.58	72.24	69.21	71.06	73.50
AT [11]	74.08	72.77	70.55	72.31	73.44
VID [1]	74.11	73.30	70.38	72.61	73.09
RKD [17]	73.35	72.22	69.61	71.82	71.90
CRD [30]	75.48	74.14	71.16	73.48	75.51
WCoRD [2]	75.88	74.73	71.56	73.81	75.95
OFD [6]	75.24	74.33	70.98	73.23	74.95
FAM-KD (layer-to-layer) - Ours	76.03	74.88	72.03	74.03	76.24
Layer to layer + Soft logit-based distillation					
WCoRD + KD [2]	76.11	74.72	71.92	74.20	76.15
Knowledge review-based distillation					
ReviewKD [19]	76.12	75.09	71.89	73.89	75.63
FAM-KD (review) - Ours	76.47	75.40	72.15	74.45	76.84

Table 1. Results on the CIFAR-100 validation set. Teachers and students are in the same architecture. FAM-KD (layer-to-layer) and FAM-KD (review) refer to our proposed methods in Section 3.2.1 and Section 3.2.2, respectively. Our reported results are an average of three trials.

Teacher	ResNet32x4	WRN-40-2	ResNet32x4	VGG13
Student	ShuffleNet-V1	ShuffleNet-V1	ShuffleNet-V2	MobileNet-V2
Teacher	79.42	75.61	79.42	74.64
Student	70.50	70.50	71.82	64.60
Soft logit-based distillation				
KD [7]	74.07	74.83	74.45	67.37
DKD [36]	76.45	76.70	77.07	69.71
Layer to layer-based distillation				
FITNET [23]	73.59	73.73	73.54	63.16
AT [11]	71.73	73.32	72.73	59.40
VID [1]	73.38	73.61	73.40	65.56
RKD [17]	72.28	72.21	73.21	64.52
CRD [30]	75.11	76.05	75.65	69.73
WCoRD [2]	75.40	76.32	75.96	69.47
OFD [6]	75.98	75.85	76.82	69.48
FAM-KD (layer-to-layer) - Ours	77.15	77.33	77.64	69.96
Layer to layer + Soft logit based-distillation				
WCoRD + KD [2]	75.77	76.68	76.48	70.02
Knowledge review-based distillation				
ReviewKD [19]	77.45	77.14	77.78	70.37
FAM-KD (review) - Ours	77.76	77.57	78.41	70.88

Table 2. The comparative results on the CIFAR-100 validation set. Teachers and students are in the different architectures. FAM-KD (layer-to-layer) and FAM-KD (review) refer to our proposed methods in Section 3.2.1 and Section 3.2.2, respectively. Our reported results are an average of three trials.

Setting		Teacher	Student	KD [7]	AT [11]	OFD [6]	CRD [30]	WCoRD [2]	DKD [36]	ReviewKD [19]	FAM-KD (Ours)
(a)	Top-1	73.31	69.75	70.66	70.69	70.81	71.17	71.49	71.70	71.61	71.91
	Top-5	91.42	89.07	89.88	90.01	89.98	90.13	90.16	90.41	90.51	90.53
(b)	Top-1	76.16	68.87	68.58	70.69	70.81	71.17	-	72.05	72.56	73.33
	Top-5	92.86	88.76	88.98	90.01	89.98	90.13	-	91.05	91.00	91.44

Table 3. Top-1 and top-5 accuracy (%) on the ImageNet validation set. (a) ResNet34 and ResNet18 and (b) ResNet50 and MobileNetV1 are used as the teacher and student architectures. Our results (FAM-KD) are with the enhanced knowledge review-based distillation (Section 3.2.2). Our reported results are an average of three trials.

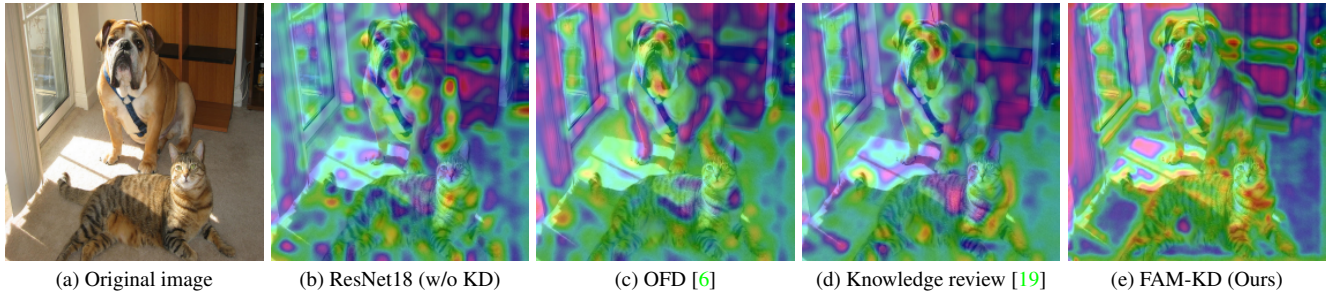


Figure 4. (a) Original image. (b) - (e) Grad-CAMs [27] from layer 9 of ResNet18 model when training (b) without knowledge distillation, (c) with OFD [6], (d) with knowledge review [19], and (e) with FAM-KD (ours), respectively. When training with distillation, ResNet34 is used as the teacher. The figure shows that our FAM-KD (e) has better focus on the object than using OFD [6] and knowledge review [19].

Compare to WCoRD + KD [2], our method achieves competitive results, despite that we only use feature-based distillation. The highest improvement over WCoRD + KD [2] is 1.38% with the ResNet32x4/ShuffleNet-V1 setting. It is worth noting that even with the layer-to-layer setting, the FAM-KD achieves comparable results with the current state-of-the-art feature distillation method using the review mechanism [19].

Regarding the knowledge review mechanism (FAM-KD (review)), we outperform compared methods for all teacher-student distillation pairs. Compare to [19], our method outperforms ReviewKD [19] in all cases. Compare to DKD [36], which is a soft logit-based distillation method, our method FAM-KD (review) also outperforms DKD [36] in all settings. The highest improvement is 1.34% with the ResNet32x4/ShuffleNet-V2 setting. The promising results have shown the effectiveness of the FAM module, supporting students to perform better.

Comparative results on ImageNet. We validate our approach on the large-scale dataset ImageNet [24] in the case of integrating the FAM module into the knowledge review distillation mechanism (FAM-KD). Table 3 presents the top-1 and top-5 classification accuracy on the ImageNet validation set of various distillation methods. When both teacher and student have the same architecture, we employ ResNet34/ResNet18 as the teacher/student pair. Meanwhile, when teacher and student have different architectures, we use ResNet50/MobileNetV1 as the teacher/student pair. Our approach yields the high-

est performance on both top-1 and top-5 accuracy. For ResNet34/ResNet18, compared to the vanilla KD [7], the FAM-KD improves by a large margin of 1.24% top-1 accuracy. Meanwhile, compared to ReviewKD [19] and DKD [36], the FAM-KD improves 0.3% and 0.21% top-1 accuracy, respectively. The relative improvements¹ over DKD and ReviewKD are considerable at 20.2% and 31.6%, respectively. For ResNet50/MobileNetV1, our approach yields a significant improvement, i.e., the improvements of 1.28% and 0.77% over DKD and ReviewKD in top-1 accuracy, and the corresponding relative improvements over DKD and ReviewKD are 36.9% and 19.3%, respectively.

Figure 4 visualizes Grad-CAMs [27] (gradient-weighted class activation mapping) extracted from layer 9 of the ResNet18, which serves as student’s architecture, with ResNet34 used as the teacher. The figure shows that the Grad-CAM when using our FAM-KD (e) has better focus on the object than using OFD [6] and knowledge review [19].

4.2.2 Object detection

Table 4 presents the object detection accuracy on the MS COCO dataset. We use the FasterRCNN [22] with FPN [13] as the detector, and use teacher/student pairs ResNet101/ResNet18, ResNet101/ResNet50 for the backbones. The results show that our method FAM-

¹Similar to [2], we compute the relative improvement as $\frac{\text{Ours}-A}{A-\text{KD}}$, where A is the method we are comparing to. For each method, the corresponding accuracy of the student is used for the calculation.

Method	ResNet101 & ResNet18			ResNet101 & ResNet50		
	AP	AP_50	AP_75	AP	AP_50	AP_75
Teacher	42.04	62.48	45.88	42.04	62.48	45.88
Student	33.26	53.61	35.26	37.93	58.84	41.05
KD [7]	33.97	54.66	36.62	38.35	59.41	41.71
FitNet [23]	34.13	54.16	36.71	38.76	59.62	41.80
FGFI [33]	35.44	55.51	38.17	39.44	60.27	43.04
ReviewKD [19]	36.75	56.72	34.00	40.36	60.97	44.08
DKD [36]	35.05	56.60	37.54	39.25	60.90	42.73
DKD + ReviewKD [36]	37.01	57.53	39.85	40.65	61.51	44.44
FAM-KD (ours)	37.20	57.86	40.01	40.77	61.42	44.49

Table 4. Comparative object detection accuracy on the MS-COCO dataset. We use the two-stage method Faster RCNN [22] with FPN [13] as the detector. On the student side, ResNet18 and ResNet50 models are selected as backbones, while teacher models use ResNet101 as a backbone. Our results (FAM-KD) are with the enhanced knowledge review-based distillation (Section 3.2.2). Our reported results are an average of three trials.

Setting	Global branch	Local branch	Top-1
(a)		✓	73.90
(b)	✓		74.17
(c)	✓	✓	74.45

Table 5. Impact of the global branch and local branch of the FAM. ResNet110 and ResNet32 are used as the teacher and the student, respectively. The results are on the CIFAR-100 validation set.

KD consistently outperforms ReviewKD [19] and the recent work DKD [36] for both settings at all metrics. With the teacher/student pairs ResNet101/ResNet18 and ResNet101/ResNet50, the proposed method outperforms DKD 2.15 and 1.52 AP points, respectively. In [36], to boost the detection accuracy of the student, the authors combine their soft logit-based distillation DKD with the knowledge review-based distillation [19]. Compare to DKD+ReviewKD [36], despite that we only use intermediate feature-based distillation, our method outperforms DKD+ReviewKD for most metrics, except AP_50 with ResNet101/ResNet50 setting.

4.3. Ablation studies

In this section, we focus on investigating how different components of the FAM module contribute to the performance of FAM-KD. All experiments are conducted on CIFAR-100 dataset with ResNet110 as a teacher and ResNet32 as a student when integrating the FAM module into the knowledge review-based mechanism as presented in Section 3.2.2. Other experiments that inspect the effectiveness of the FAM module when integrating the FAM module to ReviewKD [19], given the same ReviewKD implementation in the public distiller code-base [36] are provided in supplementary materials.

With/without global and local branches in FAM. In Table 5, we present performance of FAM-KD with and without the global branch and local branch. The results show

Setting	Top-1
FAM-KD (w/o HPF)	73.82
FAM-KD	74.45

Table 6. Effect of the high pass filter (HPF) component in global branch of the FAM. ResNet110 and ResNet32 are used as the teacher and the student, respectively. The results are on the CIFAR-100 validation set.

that the global branch benefits the model better than the local one and having both branches gives the best result.

With/without high pass filter (HPF) in the global branch. The effectiveness of the HPF is presented in Table 6, i.e., having the HPF boosts the performance 0.63%. This shows the effectiveness of HPF, which helps to filter out the lowest frequency components and encourages the student to de-focus from the non-salient regions.

5. Conclusion

In this paper, we propose to use the frequency domain to encourage the student model to capture both detailed and higher-level information such as object parts based on a well-trained teacher’s guidance. We introduce a novel frequency attention module (FAM) for knowledge distillation that operates in the frequency domain and has a filter that can be adjusted to mimic the teacher’s features. This encourages the student’s features to have similar geometric structures to the teacher’s features. Moreover, we propose an enhanced knowledge review-based distillation by leveraging the proposed FAM and cross attention. We extensively evaluate our approach with different teacher and student models, and the proposed approach achieves significant improvements compared to other state-of-the-art methods for image classification on the CIFAR-100 and ImageNet datasets and for object detection on the MS COCO dataset.

References

- [1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019. 6
- [2] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *CVPR*, 2021. 1, 2, 5, 6, 7
- [3] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006. 2
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5
- [5] Yihui He, Xiangyu Zhang, and Jian Sun. Channel Pruning for Accelerating Very Deep Neural Networks. In *ICCV*, 2017. 1
- [6] Byeongho Heo, Jeesoo Kim, Sangdoon Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 1, 2, 3, 6, 7
- [7] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014. 1, 2, 6, 7, 8
- [8] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *The Journal of Machine Learning Research*, 2017. 1
- [9] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *AAAI*, 2021. 1, 2
- [10] Jianxin Wu Jian-Hao Luo and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *ICCV*, 2017. 1
- [11] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 1, 2, 6, 7
- [12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7, 8
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5
- [15] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 5
- [16] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *CVPR*, 2019. 1
- [17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 2, 6
- [18] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018. 2
- [19] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8
- [20] Cuong Pham, Tuan Hoang, and Thanh-Toan Do. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In *WACV*, pages 6435–6443, 2023. 1
- [21] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. *NIPS*, 2019. 4
- [22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2015. 7, 8
- [23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 1, 2, 6, 8
- [24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5, 7
- [25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 5
- [26] Bharat Bhusan Sau, Soumya Roy, Vinay P Namboodiri, and Raghu Seshu Iyengar. Deep knowledge distillation using trainable dense attention. 2021. 1, 2
- [27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7
- [28] Sungho Shin, Joosoon Lee, Junseok Lee, Yeonguk Yu, and Kyoobin Lee. Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition. In *ECCV*, 2022. 1, 2
- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5
- [30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 1, 2, 5, 6, 7
- [31] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 2
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 30, 2017. 5
- [33] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019. 8
- [34] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5
- [35] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. LQ-Nets: Learned Quantization for Highly Accurate and Compact Deep Neural Networks. In *ECCV*, 2018. 1
- [36] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 1, 2, 5, 6, 7, 8

- [37] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. *CoRR*, 2016. 1