



Knowledge Distillation Through Low-Frequency Logits

Takumi Kobayashi^{1,2}(✉)

¹ National Institute of Advanced Industrial Science and Technology (AIST),
Tsukuba, Japan

`takumi.kobayashi@aist.go.jp`

² University of Tsukuba, Tsukuba, Japan

Abstract. Deep convolutional neural networks (CNNs) are applied in various pattern recognition fields with great success, though demanding considerable computation cost due to the large-scale network models. Toward efficient deep models, knowledge distillation (KD) is a promising approach to effectively transfer discriminative power of a large-capacity teacher model into a small student model. In the KD framework, logits—final outputs of CNN—are useful clues to establish simple and effective distillation, though lacking capability to exploit spatial characteristics of the models. In this paper, we propose a logit-based KD method to effectively distill detailed spatial knowledge of the models. We leverage a global average pooling mechanism embedded in CNNs to construct a logit *map* which contains spatial dimensions and then extract the spatial characteristics by applying discrete cosine transform (DCT) to the map. From the viewpoint of frequency, the proposed method incorporates a standard logit-based KD loss with a frequency-based loss in a unified manner; the frequency loss is intriguingly built upon the lowest frequency logit—1st-order DCT components of the logit map. In the experiments on KD tasks for image classification, the proposed method produces favorable performance in comparison to the other approaches.

Keywords: knowledge distillation · logit map · frequency analysis

1 Introduction

Deep convolutional neural networks (CNNs) prevail in various fields of pattern recognition including computer vision and signal processing [32]. Their high performance can be attributed to the large capacity of the deep models which though impedes deployment on edge devices of less computation resources. To remedy the issue, approaches of pruning [5, 7] and weight quantization [12, 25] work for compressing the model sizes, and in particular knowledge distillation (KD) [9] is a promising approach to boost the performance of smaller-capacity models without touching the architecture and representation of network parameters. KD leverages an effective distillation loss to facilitate training of the smaller

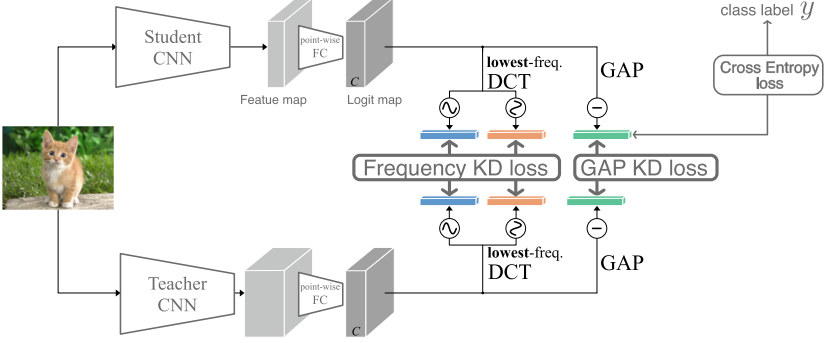


Fig. 1. Overview of the proposed KD method. We leverage DCT decomposition of logit maps to integrate frequency-based logits and global average pooling (GAP) logits in a unified manner; it should be noted that the GAP corresponds to direct-current (0-frequency) process of DCT.

model, so-called *student*, with the help of a *teacher* model of larger capacity; it has been successfully applied to diverse tasks including classification [9], semantic segmentation [30] and image generation [29].

KD is formulated to minimize discrepancy between the student and teacher models which is measured mainly from two perspectives of features at intermediate layers and logits at the last layer. The feature-based KD approaches [3, 8, 18, 21, 27] focus on difference between the student and teacher feature maps at intermediate layers through the lens of feature representation and attention maps. Though being flexible in terms of designing distillation criteria, they are required to cope with inconsistent feature maps of different dimensionalities as well as complicated layer-wise relationships between the different models; they are usually addressed by introducing extra modules and/or mechanisms tailored for distillation. On the other hand, logit-based distillation [9, 20, 24, 31] deals with high-level logit representation which is the final output of the models to render task-related knowledge of high abstraction. Thereby, the logit-based approach effectively distills knowledge of the teacher model into the student in a simple and efficient manner without adding extra modules to their network architectures. The logit representation, however, is too high-level abstract to lose detailed characteristics of the teacher model, such as about spatial information.

In this paper, we propose an efficient logit-based KD approach to exploit detailed spatial knowledge of the teacher CNN model through the *logits* in an effective way; the proposed method is summarized in Fig. 1. To reveal spatial characteristics of logits, we unfold the logit vector into a logit *map* [33] at the process of global average pooling [14], a common mechanism embedded in CNNs. Then, the logit map is disentangled from a frequency viewpoint to effectively encode spatial patterns on the high-level abstract representation of the logit map. Since the logit frequencies lack class-wise discriminative characteristics which are transformable to softmax probabilities, the standard KL-divergence [9] com-

monly used in KD losses is less applicable to the frequency representation. Thus, we introduce a dual KL-divergence loss to effectively measure discrepancy of the logit-frequency features. Specifically, the logit representation exhibits intriguing synergy with *lower*-frequency bands, contributing to efficient computation as well as performance boost. Our main contributions are summarized as follows:

- We propose an efficient logit-based knowledge distillation based on the logit-frequency representation which effectively encodes spatial patterns to render detailed knowledge of teacher models via discrete cosine transform (DCT).
- We propose a dual Kullback-Leibler divergence loss to measure the discrepancy of the logit frequency components for facilitating the proposed logit-frequency KD.
- In the experiments on ImageNet-1K dataset, the efficacy of the proposed method is validated in comparison to the other KD approaches. Particularly, the empirical evaluation demonstrates intriguing synergy between logits and lowest-frequency DCT in terms both of performance and computation cost. We also demonstrate the general applicability of our method to the function-matching KD framework [1].

1.1 Related Works

Feature-Based KD. Feature maps at intermediate layers of CNNs are effective representation for distillation. FitNet [19] constructs feature-based distillation by comparing layer-wise feature maps between teacher and student CNNs. AT [27] focuses on the attention representation built upon the feature maps as a basic ingredient for distillation. In [16, 21], relationships of the feature representations among samples are taken into account in the distillation. OFD [8] provides sophisticated activation function to facilitate distillation from teacher feature maps to student ones. In ReviewKD [3], one-to-one layer correspondence is extended by aggregating multi-layer feature maps. In those methods, while detailed spatial characteristics are exploited for KD, some transformation mechanisms (modules) are additionally embedded into the networks to align the feature-map representation between teacher and student layers, which increases complexity of training. In contrast, we address the spatial-aware KD by means of *logits* without imposing any additional parameterized modules on the networks.

Logit-Based KD. Logits are final outputs of CNN models to encode task-related characteristics regarding class categories which are useful knowledge to be distilled; the logits are usually fed into a softmax cross-entropy loss on classification tasks. In logit-based KD, while mean-squared error (MSE) is simply applicable to measure discrepancy between the teacher and student logits [24], Kullback-Leibler (KL) divergence with tempered softmax [9] is a prevailing loss function to facilitate KD training. It is sophisticated to DKD [31] by disentangling the KL divergence in terms of target (ground-truth) class and the others, while the tempering process in the softmax is analyzed in [20]. The logit-based

KD is established by designing a loss function to measure discrepancy of logits, due to which the KD training is performed in a simple and efficient way. The logit representation, however, lacks spatial information in contrast to the feature maps at the intermediate layers. In this study, we incorporate spatial characteristics of teacher/student models into the logit-based KD by unfolding the logit (vector) to a spatial logit map.

Logit Map. Global average pooling (GAP) [14] is embedded in most CNNs [6, 10, 11], followed by a linear FC classifier to finally output the logits (vector). By swapping the processing order of GAP and FC, CNN provides a logit *map* which contains spatial dimensions as in a feature map. While the logit map is utilized to enhance explainability of CNNs as in CAM [33], it is also useful for regularizing network training [22] and knowledge distillation [13, 17]; the local and global logit representation are distilled through the map in GLD [13] and the logit map is dealt with as an attention map like CAM by CAT-KD [17]. In contrast, to effectively encode the spatial characteristics of the logit maps in a KD framework, we resort to frequency-based representation of the map via DCT.

Frequency Representation. Frequency-based analysis, especially via DCT, has been widely applied in computer vision. While DCT-based pre-processing for input images can be leveraged to facilitate training deep CNNs [26], frequency-based approaches are also incorporated in a KD framework. WaveletKD [29] constructs a KD loss in a wavelet space of images to enhance image generation, and frequency representation of feature maps is leveraged for KD-based semantic segmentation [30] and feature-based KD [18]. Those methods tackle the challenge to exploit effective *high*-frequency components of *feature* maps by means of wavelet [29], learnable tokens [30] and extra trainable modules [18]. In contrast, we delve into frequency-based representation of *logit maps* to construct an effective KD loss based on the logit frequencies. Besides, we also present effectiveness of *lower* frequency logits, in contrast to the above-mentioned works that explore higher-frequency feature maps.

2 Method

This section details our KD method which deals with spatial frequencies of logit maps. We first briefly describe how the logit maps are constructed in CNN and then propose a KD loss function to measure discrepancy between teacher and student logits in a frequency domain.

2.1 Logit Map

An input image $\mathcal{I}$ is processed by a deep CNN model ϕ which is equipped with global average pooling (GAP) [14] at the penultimate layer followed by the fully-

connected (FC) linear classifier, to produce a logit vector $\bar{\mathbf{z}}$ as follows;

$$\text{Feature extraction: } \mathbf{X} = \phi(\mathcal{I}) = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^{HW} \in \mathbb{R}^{d \times H \times W}, \quad (1)$$

$$\text{GAP: } \bar{\mathbf{x}} = \frac{1}{HW} \sum_{i=1}^{HW} \mathbf{x}_i \in \mathbb{R}^d, \quad (2)$$

$$\text{FC classifier: } \bar{\mathbf{z}} = \mathbf{W}^\top \bar{\mathbf{x}} + \mathbf{b} \in \mathbb{R}^C, \quad (3)$$

where (H, W) indicates a spatial size of the feature map $\mathbf{X}$ produced by CNN ϕ ; a d -dimensional feature vector at the i -th position is denoted by $\mathbf{x}_i \in \mathbb{R}^d$. Those features are aggregated via GAP (2), and then finally projected into the logit vector $\bar{\mathbf{z}}$ by the FC classifier (3) which is parameterized by $\mathbf{W} \in \mathbb{R}^{d \times C}$ and $\mathbf{b} \in \mathbb{R}^C$ for C -class classification. It is noteworthy that this process (1–3) including GAP is employed in most CNNs [6, 10].

Due to the linearity of GAP and FC, we can swap them [33] while keeping the final logits $\bar{\mathbf{z}}$;

$$\text{Point-wise FC classifier: } \mathbf{z}_i = \mathbf{W}^\top \mathbf{x}_i + \mathbf{b} \in \mathbb{R}^C, \forall i \in \{1, \dots, HW\}, \quad (4)$$

$$\text{GAP: } \bar{\mathbf{z}} = \frac{1}{HW} \sum_{i=1}^{HW} \mathbf{z}_i \in \mathbb{R}^C. \quad (5)$$

For knowledge distillation, we focus on the logit map [13, 17, 33] of

$$\mathbf{Z} = \{\mathbf{z}_i \in \mathbb{R}^C\}_{i=1}^{HW} \in \mathbb{R}^{C \times H \times W}. \quad (6)$$

In contrast to intermediate feature maps, such as $\mathbf{X}$, the logit map $\mathbf{Z}$ contains a consistent channel dimension C which is shared by teacher and student CNNs, so that the teacher and student maps are directly comparable without introducing additional transformation. Besides, the spatial dimensions (H, W) in the logit map render more detailed characteristics of CNN models than logit vectors used in logit-based KDs [9, 31]. We hereafter denote the teacher and student logit maps by $\mathbf{Z}^t$ and $\mathbf{Z}^s$, respectively.

2.2 Frequency-Based Distillation

On the assumption that $\mathbf{Z}^t$ and $\mathbf{Z}^s$ have identical spatial dimensions (H, W) ¹, the discrepancy between those logit maps is straightforwardly measured by

$$\mathcal{L}^{MSE}(\mathbf{Z}^s, \mathbf{Z}^t) = \frac{1}{HW} \sum_{i=1}^{HW} \frac{1}{2C} \|\mathbf{z}_i^s - \mathbf{z}_i^t\|_2^2, \quad (7)$$

$$\mathcal{L}^{KL}(\mathbf{Z}^s, \mathbf{Z}^t) = \frac{1}{HW} \sum_{i=1}^{HW} \text{kldiv}(\mathbf{z}_i^s, \mathbf{z}_i^t; T), \quad (8)$$

¹ As shown in the later, the proposed method is applicable to different-sized logit maps.

where we use the following KL-divergence [9] with a temperature T ,

$$\text{kldiv}(\mathbf{z}^s, \mathbf{z}^t; T) = T^2 \sum_{c=1}^C \sigma(\mathbf{z}^t/T)_c \log \frac{\sigma(\mathbf{z}^t/T)_c}{\sigma(\mathbf{z}^s/T)_c}, \quad (9)$$

$$\text{where softmax } \sigma(\mathbf{z}) = \left\{ \frac{\exp(z_c)}{\sum_{c'=1}^C \exp(z_{c'})} \right\}_{c=1}^C \in \mathbb{R}^C. \quad (10)$$

These losses are constructed by directly applying MSE and a standard KD loss [9] to point-wise logit pairs $(\mathbf{z}_i^s, \mathbf{z}_i^t)$, respectively. It should be noted that, due to the FC classifier (4), the point-wise logits $\mathbf{z}_i$ exhibit the class-wise discriminative nature that a target-class logit score would be higher, as in the GAP logit $\bar{\mathbf{z}}$ (5); it is a rationale to apply the KL-divergence (9) to the point-wise KD (8). These loss functions (7, 8), however, measure discrepancy of the logit maps *locally* in the point-wise manner without paying attention to rather *global* structure of the map which is related to robust and primary characteristics of the *spatial* logit maps.

Frequency Representation of Logit Maps. To extract global characteristics of the logit maps, we transform the maps into a frequency domain where spatial (periodic) patterns are well encoded. The logit map $\mathbf{Z}$ is converted to the frequency representation $\mathbf{F} = \{\mathbf{f}_j \in \mathbb{R}^C\}_{j=1}^{HW}$ by discrete cosine-transform (DCT);

$$\mathbf{f}_j = \{\langle \mathbf{Z}_{c,:}, \mathbf{B}(f_j^h, f_j^w) \rangle\}_{c=1}^C \in \mathbb{R}^C, \quad (11)$$

where $\mathbf{Z}_{c,:} \in \mathbb{R}^{H \times W}$ is the c -th channel slice of the logit map $\mathbf{Z} \in \mathbb{R}^{C \times H \times W}$ and the j -th DCT basis with vertical and horizontal frequencies of $(f_j^h, f_j^w) \in \{(0, 0), \dots, (H-1, W-1)\}$ is formulated as

$$\mathbf{B}(f_j^h, f_j^w) = \mathbf{B}_H(f_j^h) \mathbf{B}_W(f_j^w)^\top \in \mathbb{R}^{H \times W}, \quad (12)$$

$$\mathbf{B}_N(f) = \begin{cases} \frac{1}{N} \mathbf{1}_N & (f = 0) \\ \frac{\sqrt{2}}{N} \cos \frac{\pi f (2\mathbf{p}_N + 1)}{2N} & (f \in \{1, \dots, N-1\}) \end{cases} \in \mathbb{R}^N, \quad (13)$$

where $N \in \{H, W\}$, $\mathbf{1}_N = [1, \dots, 1]^\top \in \mathbb{R}^N$ and $\mathbf{p}_N = [0, 1, \dots, N-1]^\top \in \mathbb{R}^N$; the DCT bases are shown in Fig. 2 and note that $\mathbf{f}_1 = \bar{\mathbf{z}}$, GAP logits in (5), due to $(f_1^h, f_1^w) = (0, 0)$.

In this frequency domain, the MSE loss (7) is rewritten to

$$\mathcal{L}^{MSE}(\mathbf{Z}^s, \mathbf{Z}^t) = \sum_{j=1}^{HW} \frac{1}{2C} \|\mathbf{f}_j^s - \mathbf{f}_j^t\|_2^2 = \sum_{k=0}^{H+W-2} \sum_{j|f_j^h+f_j^w=k} \frac{1}{2C} \|\mathbf{f}_j^s - \mathbf{f}_j^t\|_2^2, \quad (14)$$

where we introduce a total frequency $k = f^h + f^w \in \{0, \dots, H+W-2\}$ as depicted in Fig. 2. The MSE loss (14) is composed of all the differences across

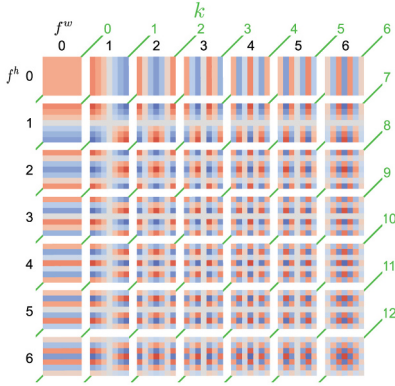


Fig. 2. DCT bases for $H = W = 7$. Total frequency is sum of vertical and horizontal frequencies; $k = f^h + f^w$. Best viewed in color.

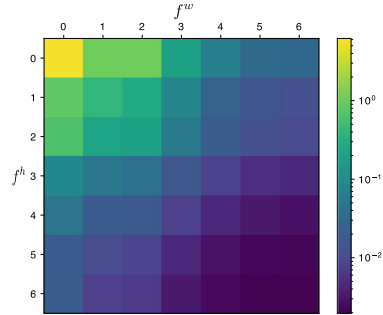


Fig. 3. DCT frequency powers of logit maps produced by ResNet-34 [6] on ImageNet dataset.

whole frequencies including higher ones. As the high-level abstract representation of the logit maps are dominated by low frequencies (Fig. 3) and higher-frequency components are susceptible to noises, it is effective to restrict our focus on the lower-frequency ones by

$$\mathcal{L}^{freq}(\mathbf{Z}^s, \mathbf{Z}^t) = \sum_{k=0}^K \sum_{j | f_j^h + f_j^w = k} \mathbb{E} \ell_k(\mathbf{f}_j^s, \mathbf{f}_j^t), \quad (15)$$

where we measure discrepancies on up to K frequency, and ℓ_k denotes a loss function for the k -th frequency as detailed in the next section. The logit scores are disentangled in terms of frequency and thus we can determine the frequency-wise loss function ℓ_k independently.

2.3 Frequency Loss ℓ_k

As $k = 0$ -frequency basis $\mathbf{B}(0, 0)$ in (12) is equivalent to GAP, the $\mathbf{f}_1$ corresponds to the logit vector $\bar{\mathbf{z}}$ in (5). Thus, we can set $\ell_{k=0}$ as off-the-shelf logit-based KD loss, such as [9, 31], which we denote by **kd**.

On the other hand, the logit frequency representation of $k > 0$ exhibits distinct characteristics than the GAP logits of $k = 0$. Namely, due to cosine weighting over the spatial domains, the class-category discriminative nature embedded in the GAP logit $\bar{\mathbf{z}} = \mathbf{f}_1$ is no longer valid; the logit frequency of the target class could be higher or lower according to its spatial patterns. Therefore, it is less effective to apply a KD loss **kd**, especially KL-divergence **kldiv** (9), based on softmax $\sigma(\mathbf{f})$ to measure the discrepancy since it highly focuses on higher score in $\mathbf{f}$, ignoring lower ones as shown in Fig. 4b. The MSE loss used in (14) also works ineffectively, being less consistent with $\ell_{k=0} = \mathbf{kd}$, the off-the-shelf KD loss

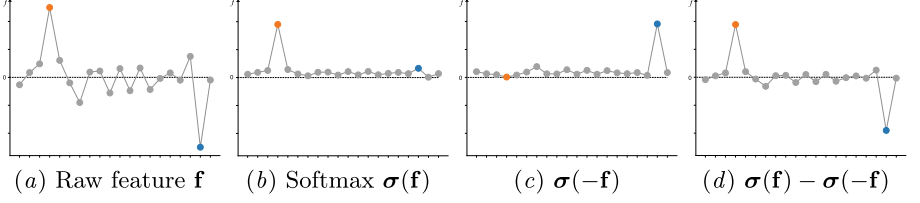


Fig. 4. Softmax-based representations of logit frequency features $\mathbf{f}$. Components of larger magnitude are highlighted. Best viewed in color.

based on KL-divergence. MSE uniformly measures difference for any logit frequency features, although the frequency components of larger magnitude would be more informative than those of smaller magnitudes.

To effectively measure the discrepancy of logit frequencies, we propose a dual KL-divergence as

$$\ell_k(\mathbf{f}_j^s, \mathbf{f}_j^t; T) = \frac{1}{2} \{ \text{kldiv}(\mathbf{f}_j^s, \mathbf{f}_j^t; T) + \text{kldiv}(-\mathbf{f}_j^s, -\mathbf{f}_j^t; T) \}. \quad (16)$$

By combining two types of KL-divergences, the logit frequencies both of larger positive and larger negatives contribute to the discrepancy measurement (Fig. 4), which makes it possible to exploit various prominent spatial patterns for knowledge distillation.

The dual KL-divergence loss (16) is regarded as an extension of MSE loss as follows. The gradient of MSE loss w.r.t $\mathbf{f}_j^s$ is given by

$$\ell_k^{MSE} = \frac{1}{2C} \|\mathbf{f}_j^s - \mathbf{f}_j^t\|_2^2, \quad \frac{\partial}{\partial \mathbf{f}_j^s} \ell_k^{MSE} = \frac{1}{C} (\mathbf{f}_j^s - \mathbf{f}_j^t), \quad (17)$$

while the loss gradient of (16) is written as

$$\frac{\partial}{\partial \mathbf{f}_j^s} \ell_k = T \left[\frac{\boldsymbol{\sigma}(\mathbf{f}_j^s/T) - \boldsymbol{\sigma}(-\mathbf{f}_j^s/T)}{2} - \frac{\boldsymbol{\sigma}(\mathbf{f}_j^t/T) - \boldsymbol{\sigma}(-\mathbf{f}_j^t/T)}{2} \right], \quad (18)$$

which replaces $\mathbf{f}$ in (17) with $\boldsymbol{\sigma}(\mathbf{f}/T) - \boldsymbol{\sigma}(-\mathbf{f}/T)$ to enhance positive and negative components of larger magnitudes as shown in Fig. 4d. Thus, (16) renders a distillation loss by focusing on those distinctive feature components in a superior manner to MSE. Following discussion in [9], connection between (17) and (18) is further clarified by

$$\frac{\partial}{\partial \mathbf{f}_j^s} \ell_k \xrightarrow{T \rightarrow \infty} \frac{1}{C} (\mathbf{f}_j^s - \mathbf{f}_j^t) = \frac{\partial}{\partial \mathbf{f}_j^s} \ell_k^{MSE}, \quad (19)$$

where we use $\boldsymbol{\sigma}(\mathbf{f}/T) \xrightarrow{T \rightarrow \infty} \frac{1}{C} (1 + \mathbf{f}/T)$. Therefore, we can say that (16) is generally applicable to measure discrepancy of any feature vectors as an extension of MSE.

Proposed KD Loss. Our KD loss is finally formulated by

$$\mathcal{L}^{our}(\mathbf{Z}^s, \mathbf{Z}^t) = \lambda_{kd} \text{kd}(\mathbf{f}_1^s, \mathbf{f}_1^t) + \lambda \sum_{k=1}^K \mathbb{E}_{j | f_j^h + f_j^w = k} \frac{\text{kldiv}(\mathbf{f}_j^s, \mathbf{f}_j^t; T) + \text{kldiv}(-\mathbf{f}_j^s, -\mathbf{f}_j^t; T)}{2}, \quad (20)$$

where λ_{kd} and λ are weights for the off-the-shelf KD loss kd based on GAP logits $\mathbf{f}_1 = \bar{\mathbf{z}}$ and the proposed frequency-based KD loss of $k > 0$, respectively.

Considering that the logit maps of abstract representation exhibit concentration on lower frequencies (Fig. 3), we can set $K = 1$ which is also empirically validated in the experimental results (Sect. 3). Thereby, the loss is efficiently computed for any sized logit maps by simply measuring GAP ($\text{B}(0, 0)$) as well as $k = 1$ -frequencies by means of DCT with $\text{B}(0, 1)$ and $\text{B}(1, 0)$.

3 Results

We apply the proposed method to knowledge distillation tasks on ImageNet-1K [4] classification.

Training Protocol. Given a pretrained teacher model, we train a student model from scratch via the KD loss (20) together with a standard supervised cross-entropy loss by following the standard training recipe in this literature; we use SGD optimizer with 0.9 momentum, 512 batch size, 0.0001 weight decay and 0.2 initial learning rate which is decayed by a factor of 0.1 at 30, 60 and 90-th epoch over 100 training epochs, while applying data augmentation composed of resized random cropping and horizontal flipping. It is implemented on PyTorch by using 4 NVIDIA TitanV GPUs.

3.1 Ablation Study

We analyzed the proposed method from various aspects by using a pair of ResNet-34 [6] teacher and ResNet-18 student; both models produce a logit map $\mathbf{Z}$ of $H = W = 7$. In the loss (20), we use a standard KL-divergence loss kldiv (9) for the $k = 0$ -frequency (GAP) loss kd with $\lambda_{kd} = 0.5$ and the $k > 0$ -frequency loss ℓ_k (16) with $\lambda = 1$. The temperature is set to $T = 1$ in both types of losses.

Frequency. The key ingredient of the proposed method is the frequency-wise loss, the second term in (20) using K frequencies; in this case of $H = W = 7$, we can consider $K \in \{1, \dots, 12\}$ as shown in Fig. 2.

We first evaluate performance in a multi-frequency framework (20) to sum up $k \in \{1, \dots, K\}$ -frequency losses, reporting the results in Table 1a. The performance is maximized at the lowest frequency ($K = 1$), then decreasing as the higher ones are sequentially incorporated. The efficacy of the lowest frequency is

Table 1. Performance results using various frequencies. (a) Multiple $k \in \{1, \dots, K\}$ frequency losses are summed up in our loss (20). Instead, (b) only a single $k = K$ -th frequency loss is used. Top-1 and Top-5 classification accuracies (%) are reported.

(a) Multiple frequencies 1~K														(b) Single K-th frequency													
Top-1	Top-5	1	2	3	4	5	6	7	8	9	10	11	12	Top-1	Top-5	1	2	3	4	5	6	7	8	9	10	11	12
71.86	90.68	✓												71.86	90.68	✓											
71.81	90.64	✓	✓											71.69	90.49	✓											
71.81	90.63	✓	✓	✓										71.53	90.29		✓										
71.81	90.53	✓	✓	✓	✓									71.44	90.21			✓									
71.72	90.61	✓	✓	✓	✓	✓								71.28	90.15				✓								
71.63	90.51	✓	✓	✓	✓	✓	✓							71.42	90.27					✓							
71.73	90.42	✓	✓	✓	✓	✓	✓	✓						71.44	90.23						✓						
71.65	90.59	✓	✓	✓	✓	✓	✓	✓	✓					71.42	90.32							✓					
71.68	90.63	✓	✓	✓	✓	✓	✓	✓	✓	✓				71.39	90.23								✓				
71.61	90.64	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			71.41	90.24									✓			
71.67	90.61	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		71.29	90.16										✓		
71.62	90.56	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	71.10	90.15											✓	

further highlighted through performance comparison in a single-frequency approach; only $k = K$ -th frequency is used in the second term in (20), replacing $\sum_{k=1}^K$ with $\sum_{k=K}^K$, for directly comparing respective frequency losses as shown in Table 1b. The lowest $K = 1$ frequency produces superior performance to the other higher ones. The result is intriguingly contrastive against the other feature-based KD [18, 29, 30] that explores effective *higher*-frequency representations as described in Sect. 1.1. The logit representation is abstract and high-level about class categories, exhibiting primary spatial patterns in lower frequency domains, while higher frequency logits may be dominated by noisy patterns. The lowest $K = 1$ frequency representation extracts such a rough (abstract) first-order spatial patterns, being efficiently incorporated in our method as an effective clue to boost KD performance. Besides, application of only $K = 1$ frequency renders computational efficiency; we just compute two DCT components of logit maps (Fig. 2) in addition to the standard GAP one. We apply the lowest frequency $K = 1$ to our loss (20) in the following experiments.

Loss. Then, we evaluate a functional form of the frequency-wise loss ℓ_k . Table 2 shows performance comparison over various types of loss functions mentioned in Sect. 2.3. The proposed dual KL-divergence loss (16) outperforms standard KL-divergence (9) and MSE (17) losses. The KL-divergence losses are superior to MSE which uniformly measures discrepancies at any logit magnitudes while KL-divergence focuses on logits of larger magnitudes as shown in Fig. 4. Specifically, our dual KL-divergence (16) is capable of dealing with logit-frequency features

Table 2. Comparison across various types of loss formulation.

Loss	Top-1	Top-5
Frequency-wise loss $\ell_{k=1}$		
Dual KL-div (16)	71.86	90.68
KL-div (9)	71.68	90.57
MSE (17)	71.34	90.41
Spatial loss $\mathcal{L}$		
KL-div (8)	71.61	90.37
MSE (7)	70.74	90.20

Table 3. Performance analysis for hyper-parameters of our loss (20).

(a) Temperature T			(b) Weight λ		
T	Top-1	Top-5	λ	Top-1	Top-5
0.5	71.82	90.56	0.5	71.64	90.34
1	71.86	90.68	1	71.86	90.68
2	71.57	90.54	2	71.81	90.65
4	71.52	90.52			

both of positively and negatively larger magnitudes (Fig. 4d) to effectively distill lowest-frequency logits of the teacher model for facilitating KD. In contrast, the standard KL-divergence (9) emphasizes only the larger logits on a positive side, suppressing the negative ones (Fig. 4b). Since the lowest-frequency logits can exhibit larger magnitudes both positively and negatively, the dual KL-divergence is required to cope with those frequency-based logit representations in the KD framework.

The frequency-based approach is also compared with the *spatial* losses that naively apply KL-divergence and MSE to point-wise logit vectors in (8) and (7). Among those two types of losses, the KL-divergence works better than MSE as is the case with the above-mentioned comparison regarding a frequency-wise loss. As discussed in Sect. 2.2, the spatial KL-divergence loss (8) is deemed to contain all the frequencies, thereby producing performance close to those of broader-band frequency losses shown in the bottom rows of Table 1a. Through effective concentration on the lower-frequency logits, the proposed loss is superior to the spatial ones. As to computation, the proposed lowest-frequency loss consumes 1 (for $\mathbf{kd}$) + 2 (for $\ell_{k=1}$) = 3-times computation of KL-divergence, which is much smaller than HW -times computation of KL-divergence in the spatial KD loss (8). Thus, the proposed method establishes favorable logit-based KD in terms both of performance and computation efficiency.

Temperature. As the logit frequency ($k = 1$) renders orthogonal characteristics to that of the GAP logit ($k = 0$), the temperature of the dual KL-divergence (16) can be set differently from that of $\mathbf{kd}$ loss for the GAP logit which is set to $T_{\mathbf{kd}} = 1$ in these experiments. Table 3a shows performance results across various temperatures T in $\ell_{k=1}$, demonstrating that $T = 1$ produces the best performance while the higher temperatures degrade performance. It is accordance with the poor performance result of MSE shown in Table 2; MSE is regarded as a loss of high (infinite) temperature as discussed in (19). Our method is moderately robust against the lower temperature $T = 0.5$ which further enhances larger-magnitude logits, implying that the prominent low-frequency logits are important for effective KD. Based on these experimental results, we apply the dual KL-divergence with $T = 1$ for the frequency loss $\ell_{k=1}$ in our KD loss.

Table 4. Performance comparison. For comparison methods, we report performance scores shown in the respective papers.

Method	T: ResNet34 [6] S: ResNet18 [6]		T: ResNet50 [6] S: MobileNetV1 [10]	
	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
Feature-based KD				
AT [27]	70.69	90.01	69.56	89.33
OFD [8]	70.81	89.98	71.25	90.34
CRD [21]	71.17	90.13	71.37	90.41
ReviewKD [3]	71.61	90.51	72.56	91.00
FAM-KD [18]	71.91	90.53	73.33	91.44
Logit-based KD				
CAT-KD [17]	71.26	90.45	72.24	91.13
GLD [13]	71.63	90.53	-	-
LSKD [20]	71.42	90.29	72.18	90.80
KD [9]	70.66	89.88	68.58	88.98
DKD [31]	71.70	90.41	72.05	91.05
Ours+KD	71.86	90.68	72.05	90.83
Ours+DKD	72.19	90.70	72.94	91.33

Weight. We also explored the weight parameter λ for the frequency loss $\ell_{k=1}$ in (20). Table 3b shows that $\lambda = 1$ attains the best performance while the larger weight $\lambda = 2$ also provides favorable performance.

3.2 Comparison Results

We compare the proposed logit-based KD loss to the other KD approaches including feature-based and logit-based ones. Table 4 shows overall performance comparison on two types of teacher-student pairs in the ImageNet-1K classification framework: (a) ResNet34 [6] teacher and ResNet18 student, (b) ResNet50 teacher and MobileNetV1 [10] student. The proposed method (20) can naturally incorporate off-the-shelf logit-based KD loss, such as [9] and [31], into kd loss using the $k = 0$ (GAP) logit, to produce favorable performance. In particular, a naive KD loss [9], i.e., `kldiv` in (9), is effectively improved while keeping simplicity of the loss. The combination with DKD [31] further boosts the performance. As to the representation of logits, CAT-KD [17] and GLD [13] are based on the logit maps $\mathbf{Z} \in \mathbb{R}^{C \times H \times W}$ similarly to ours. Comparison to those methods highlights efficacy of our approach that incorporates frequency-based decomposition

Table 5. Performance results on function-matching KD framework [1] on two types of image classification tasks: ImageNet-100 dataset [23] and Food-101 dataset [2].

Loss	T: ResNet34 [6] S: ResNet18 [6]		T: ResNet50 [6] S: MobileNetV1 [10]		T: ResNet101 [6] S: ResNet34 [6]	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ImageNet-100						
Teacher	89.98	98.58	91.90	99.04	92.14	99.06
Student	84.72	96.06	85.84	96.46	89.98	98.58
KD [1]	89.38	98.48	90.14	98.62	90.88	98.84
Our+KD	89.96	98.54	90.66	98.94	91.72	99.10
Food-101						
Teacher	83.50	96.17	85.81	96.98	86.71	97.07
Student	80.67	93.98	83.45	95.18	83.50	96.17
KD [1]	84.61	96.68	86.73	97.27	86.31	97.29
Our+KD	84.91	96.73	87.03	97.47	87.24	97.60

of the logit maps into a KD framework and leverages GAP-based logits in a unified manner.

3.3 Function-Matching KD Framework [1]

In [1], efficacy of *MixUp* data augmentation [28] and *longer* training epochs is articulated in order to enhance matching of teacher and student models in a KD framework. The *MixUp* augmentation might impair natural spatial characteristics of logit maps through mixing two images, from which our frequency-based logit representation would suffer. Thus, for validating general applicability of the proposed method, this section evaluates our KD loss (20) in the *MixUp* based KD framework [1].

Training Protocol. Following [1], given a pretrained teacher model, we train a student model from scratch only on the KD loss (20) (*without* supervised cross-entropy loss) by using AdamW [15] optimizer with 512 batch size, 0.0003 weight decay and 0.003 initial learning rate which is decayed via cosine-scheduling over 1200 training epochs including 20-epoch warming up. In addition to the standard data augmentation composed of resized random cropping and horizontal flipping, *MixUp* augmentation [28] is applied. In the function-matching KD framework [1], besides *MixUp* augmentation and longer training (1200 epochs), the cross-entropy loss based on ground-truth class labels is excluded to let a student model be trained solely on the KD loss, i.e., only the supervision that a teacher model provides.

Dataset. We evaluated KD losses on the following two datasets for image classification tasks. ImageNet-100 [23], a subset of ImageNet-1K [4], contains 120K training and 5K test images of 100 ImageNet class categories. Food-101 [2] covers 75,750 training and 25,250 test images across 101 food categories.

Model. We apply three types of teacher-student pairs; (a) ResNet-34 [6] teacher and ResNet-18 student, (b) ResNet-101 teacher and ResNet-34 student, and (c) ResNet-50 teacher and MobileNet-V1 [10] student.

Results. The performance results are shown in Table 5. The proposed KD loss (20) that incorporates a standard KL-divergence (9) as the GAP loss kd favorably improves performance of KD [9] by effectively exploiting the spatial knowledge of a teacher model even through the *MixUp*ed logit maps; the performances of student models are significantly boosted to surpass or be on par with those of teacher models. Due to linearity of the DCT process (13), the lowest-frequency ($k = 1$) logits also enjoy the MixUp-augmentation effect in a similar way to the GAP logits for boosting KD performance.

4 Conclusion

In this study, we have proposed a novel logit-based KD approach for CNNs equipped with GAP, to exploit spatial characteristics of a teacher model through logit maps. The proposed KD loss is built upon frequency-wise decomposition of the spatial logit map by means of DCT, thereby distilling spatial teacher knowledge through the lowest frequency of the logit map in an efficient way without complicated training processes. In the experiments on image classification tasks, the proposed KD method produces favorable performance in comparison to the other KD methods. Our future works include an extension from CNNs to vision transformers (ViTs) such as by constructing (logit) maps from attention map and/or token-wise classification.

References

1. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: a good teacher is patient and consistent. In: CVPR, pp. 10925–10934 (2022)
2. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101 – mining discriminative components with random forests. In: ECCV, pp. 446–461 (2014)
3. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: CVPR, pp. 5008–5017 (2021)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
5. Frankle, J., Carbin, M.: The lottery ticket hypothesis: finding sparse, trainable neural networks. In: ICLR (2019)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
7. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: ICCV, pp. 1389–1397 (2017)
8. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: ICCV, pp. 1921–1930 (2019)
9. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS 2014 Deep Learning Workshop (2014)
10. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. [arXiv:1704.04861](https://arxiv.org/abs/1704.04861) (2017)
11. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 2261–2269 (2017)
12. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Quantized neural networks: training neural networks with low precision weights and activations. *J. Mach. Learn. Res.* **18**, 1–30 (2018)
13. Kim, Y., Park, J., Jang, Y., Ali, M., Oh, T.H., Bae, S.H.: Distilling global and local logits with densely connected relations. In: ICCV, pp. 6290–6300 (2021)
14. Lin, M., Chen, Q., Yan, S.: Network in network. In: ICLR (2014)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
16. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR, pp. 3967–3976 (2019)
17. Park, W., Kim, D., Lu, Y., Cho, M.: Class attention transfer based knowledge distillation. In: CVPR, pp. 11868–11877 (2023)
18. Pham, C., Nguyen, V.A., Le, T., Phung, D.: Frequency attention for knowledge distillation. In: WACV, pp. 2277–2286 (2024)
19. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: hints for thin deep nets. In: ICLR, pp. 535–541 (2015)
20. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: CVPR, pp. 15731–15740 (2024)
21. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
22. Wang, C., Xiao, J., Han, Y., Yang, Q., Song, S., Huang, G.: Towards learning spatially discriminative feature representations. In: ICCV, pp. 1326–1335 (2021)
23. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML, pp. 9929–9939 (2020)
24. Wu, J., Leng, C., Wang, Y., Hu, Q., Cheng, J.: Do deep nets really need to be deep? In: NeurIPS (2014)
25. Wu, J., Leng, C., Wang, Y., Hu, Q., Cheng, J.: Quantized convolutional neural networks for mobile devices. In: CVPR, pp. 4820–4828 (2016)
26. Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K., Ren, F.: Learning in the frequency domain. In: CVPR, pp. 1740–1749 (2020)
27. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
28. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: ICLR (2018)
29. Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K.: Wavelet knowledge distillation: towards efficient image-to-image translation. In: CVPR, pp. 12464–12474 (2022)
30. Zhang, Y., Huang, T., Liu, J., Jiang, T., Cheng, K., Zhang, S.: Freekd: knowledge distillation via semantic frequency prompt. In: CVPR, pp. 15931–15940 (2024)

31. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR (2022)
32. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M.: A review of convolutional neural networks in computer vision. *Artif. Intell. Rev.* **57**(99), 1–30 (2024)
33. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016)