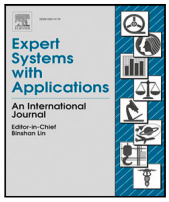
Expert Systems With Applications 250 (2024) 123892

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Coordinate Attention Guided Dual-Teacher Adaptive Knowledge Distillation for image classification

Dongtong Ma [a], Kaibing Zhang [a,b,*], Qizhi Cao [a], Jie Li [c], Xinbo Gao [d]

[a] School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China
[b] Shaanxi Key Laboratory of Clothing Intelligence and State-Province Joint Engineering and Research Center of Advanced Networking and Intelligent Information Services, School of Computer Science, Xi'an Polytechnic University, Xi'an, 710048, Shaanxi, China
[c] Video and Image Processing System Laboratory, School of Electronic Engineering, Xidian University, Xi'an 710071, China
[d] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

ARTICLE INFO

Keywords:
Dual-teacher knowledge distillation
Coordinate attention mechanism
Adaptive knowledge distillation

ABSTRACT

Knowledge distillation (KD) refers to transferring the knowledge learned from a teacher network with complex architecture and strong learning ability to another student network with light-weight and weak learning ability through a specific distillation strategy. However, most existing KD approaches to image classification often employ a single teacher network to guide the training of the student network. When the teacher network makes an erroneous prediction, the transferred knowledge will deteriorate the performance of the student network. To address or mitigate the above issue, we develop a novel KD approach called Coordinate Attention Guided Dual-Teacher Adaptive Knowledge Distillation (CAG-DAKD), to deliver more discriminative and comprehensive knowledge obtained from two teacher networks to a compact student network. Specifically, we integrate the positive prediction distribution of two teacher networks according to whether the two teacher networks predict correctly and the magnitude of the cross-entropy to deliver better output distribution to guide the student network. Furthermore, to distill the most valuable knowledge from the first teacher network that has a similar architecture to the student network, a coordinate attention mechanism is introduced into different layers of the first teacher network so that the student network can effectively learn more discriminative feature representations. We conduct extensive experiments on three standard image classification datasets: CIFAR10, CIFAR100, and ImageNet to verify the superiority of the proposed method over other state-of-the-art competitors. Code will be available at https://github.com/mdt1219/CAG-DAKD.git/

1. Introduction

Image classification (Fu, Li, Liu, & Yang, 2021) is a classical computer vision task and has many potential applications such as face recognition (Song et al., 2022), image retrieval (Feng, Wang, & Tang, 2021), and object detection (Dai et al., 2021). Most current image classification research is based on large-scale deep neural networks (DNNs) or their ensembles, such as Swin Transformer (Liu et al., 2022), Diffusion Model (Yang et al., 2023), EfficientNet (El-Dahshan, Bassiouni, Khare, Tan, & Acharya, 2024), and SpectralGPT (Hong et al., 2023), which usually contain millions of model parameters for compelling results. For example, Manzari, Ahmadabadi, Kashiani, Shokouhi, and Ayatollahi (2023) proposed a highly robust yet efficient Transformer model for image classification. Whereas Huang, Su, Wu, and Chen (2023) proposed an image classification model based on EfficientNet and triple attention mechanism, enabling the model to fully capture the channel and spatial attention information. Hong et al. (2023) employed a novel 3D generative pretrained transformer to build a universal remote sensing (RS) foundation model with over 600 million parameters, which accommodates input images with varying sizes, resolutions, time series, and regions in a progressive training manner. However, with the increase of model parameters, the training process would incur a large amount of computation cost, which makes it challenging to deploy a trained deep network on other mobile terminals or embedded devices in practical applications. In response to the above issue, numerous researchers have elaborated on various techniques to compress network parameters (Wang et al., 2021). The typical approaches include quantization or binarization (Fei, Dai, Li, Zou, & Xiong, 2022), factorization (Chen, Jiang, Liu, & Zhou, 2021), network pruning (Blalock, Gonzalez Ortiz, Frankle, & Guttag, 2020), and knowledge distillation (KD) (Gou, Yu, Maybank, & Tao, 2021). Among them, the KD-based

* Corresponding author at: School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China.
E-mail addresses: dongtongma@stu.xpu.edu.cn (D. Ma), kaibingzhang@xpu.edu.cn (K. Zhang), qizhicao@stu.xpu.edu.cn (Q. Cao), leejie@mail.xidian.edu.cn (J. Li), gaoxb@cqupt.edu.cn (X. Gao).

https://doi.org/10.1016/j.eswa.2024.123892
Received 8 April 2023; Received in revised form 24 March 2024; Accepted 1 April 2024
Available online 6 April 2024

approaches which base on the teacher–student paradigm have been recognized as an effective way to train a lightweight yet powerful student network. Fundamentally, the student network can simulate the soft target output of the large-scale teacher network via knowledge transfer. Through KD (Gou et al., 2021), a lightweight neural network model can yield comparable performance while maintaining significantly lower computational complexity. The optimized student network makes the model much easier to be configured on many mobile terminals and embedded devices, which is pivotal for practical applications especially in those resource-limited scenarios.

According to how many teacher networks applied in the KD, the existing methods can be roughly categorized into two subclasses: one-teacher distillation (Huang, Guo, & Wang, 2024; Pham et al., 2024; Zhou, Aysa, Ubul, et al., 2024) and multi-teacher distillation (Luo, Zeng, & Zhong, 2024; Ma, Jiang, Guan, & Yi, 2023; Yuan et al., 2021). In the framework under one-teacher distillation, the student network acquires knowledge from one teacher network in different ways, so it can be further divided into three major subclasses: response-based distillation, feature-based distillation, and relation-based distillation. Zhou et al. (2024) addressed the shortcomings of traditional feature-based distillation algorithms by proposing that the student network learns to mimic the teacher network in generating semantically strong feature maps at the feature level, while also performing response-based distillation at the prediction heads of the network. Pham et al. (2024) propose an enhanced knowledge review-based distillation model by leveraging the proposed frequency attention module. Huang et al. (2024) introduced the diffusion model into KD and proposed to denoise student features using a diffusion model trained by teacher features. Considering that a single teacher network cannot deliver sufficient knowledge to the student network and is not beneficial to those complicated classification tasks, multi-teacher distillation-based approaches that provide more knowledge from different teachers have received intensive attention. For example, Luo et al. (2024) proposed a progressive distillation scheme, using an assistant network to alleviate the huge knowledge gap between the teacher and the student and designing an Adaptive Choose Teacher module to optimize the distillation. Yuan et al. (2021) systematically developed a reinforced method to dynamically assign weights to teacher models for different training instances and optimize the performance of student model. Ma et al. (2023) proposed a novel Multi-teacher Knowledge Distillation approach, which effectively integrates multiple teachers with importance weights to provide guidance for the accurate anomaly detection of students.

Although promising performance has been achieved by existing one-teacher or multi-teacher distillation methods, there remain several noticeable limitations: (1) The response-based KD methods only consider the soft output probability score of the response layer as knowledge but ignore the valuable knowledge learned in the intermediary layers. As a result, if the teacher network makes an erroneous prediction, the knowledge transferred to the student network may severely degrade its performance. (2) The popular feature-based KD methods directly match the intermediate layer representations between the teacher network and the student network. However, they cannot effectively explore the most discriminative features of the student network. (3) Most existing multi-teacher KD methods randomly select one teacher's soft target or average the soft targets of multiple teachers as knowledge. Such a simple distillation scheme is challenging to efficiently gathering complementary knowledge from different teacher networks.

In this paper, against the above issues, we propose a novel Coordinate Attention Guided Dual-teacher Adaptive Knowledge Distillation, termed CAG-DAKD, to learn more discriminative and valuable knowledge from two teacher networks to a student network for better classification. Unlike most existing multi-teacher KD methods which randomly select the soft target of one teacher or average the soft targets of multiple teachers as knowledge, our method transfers abundant knowledge learned from two different teacher networks through different distillation strategies to jointly guide the training of a lightweight student network. Specifically, the positive prediction distributions of two teacher networks are integrated through an adaptive weighting scheme to provide a better output distribution to guide the student network. Furthermore, the knowledge of intermediate features in the teacher network is informative and facilitates distilling more discriminative feature representations to the student network. As distinct from most feature-based KD methods that directly match the intermediate layer representations, we introduce a coordinate attention mechanism into different layers from the first teacher network with a similar architecture to the student network. Our primary contributions are outlined as follows:

- We develop an adaptive KD strategy based on a dual-teacher network framework, which delivers complementary knowledge according to whether the two teacher networks predict correctly and the magnitude of their respective cross-entropy.
- We apply a coordinate attention mechanism to different layers at the first teacher network and the student network to convey the most discriminative features to the student network.
- We perform thorough compared experiments and ablation study on three benchmark databases to validate the merit of the proposed CAG-DAKD framework over other state-of-the-art competitors.

The rest of the paper is structured the following way. Section 2 surveys the work closely related to this paper. Section 3 presents the details about the proposed CAG-DAKD. Whereas Section 4 demonstrates the experimental results on three benchmark datasets to prove the superiority of the proposed CAG-DAKD over compared method. Finally Section 5 concludes this paper.

2. Related work

In this section, we provide a quick overview of three related works, i.e., KD (Gou et al., 2021; Zhou et al., 2024), Multi-Teacher KD (Dvornik, Schmid, & Mairal, 2019; Luo et al., 2024; You, Xu, Xu, & Tao, 2017), and Attention Distillation (Zagoruyko & Komodakis, 2016a; Zhou, Zhuge, Guan, & Liu, 2020), respectively.

2.1. Knowledge distillation

The fundamental idea behind KD is to learn a lightweight student network to attain better performance by utilizing the supervision information provided by one or more teacher networks (Tzelepi, Passalis, & Tefas, 2021). According to different knowledge distilled, the existing methods can be mainly split into three categories, namely response-based KD, feature-based KD, and relation-based KD, respectively. The core idea of response-based KD is to integrate the actual category labels of samples with the output results of the teacher network as prior knowledge to guide the training of student network. This strategy was first proposed by Hinton, Vinyals, and Dean (2015), which guides the training of the student network by softening the output of the teacher network. Whereas the feature-based KD approaches utilize the features learned from intermediate layers as knowledge to direct the learning of the student network. One representative method is the FitNets model proposed by Romero et al. (2014), where the features from the middle layer of the teacher network are extracted as the output target of the middle layer of the student network to guide the student network training. Different from the previous two types of KD methods, the relation-based KD approaches obtain knowledge from the relationship between different layers of the teacher network or various data samples in a mini-batch. By defining the Gram matrix, Yim, Joo, Bae, and Kim (2017) explored the relationships between different feature layers as knowledge to supervise the student network.

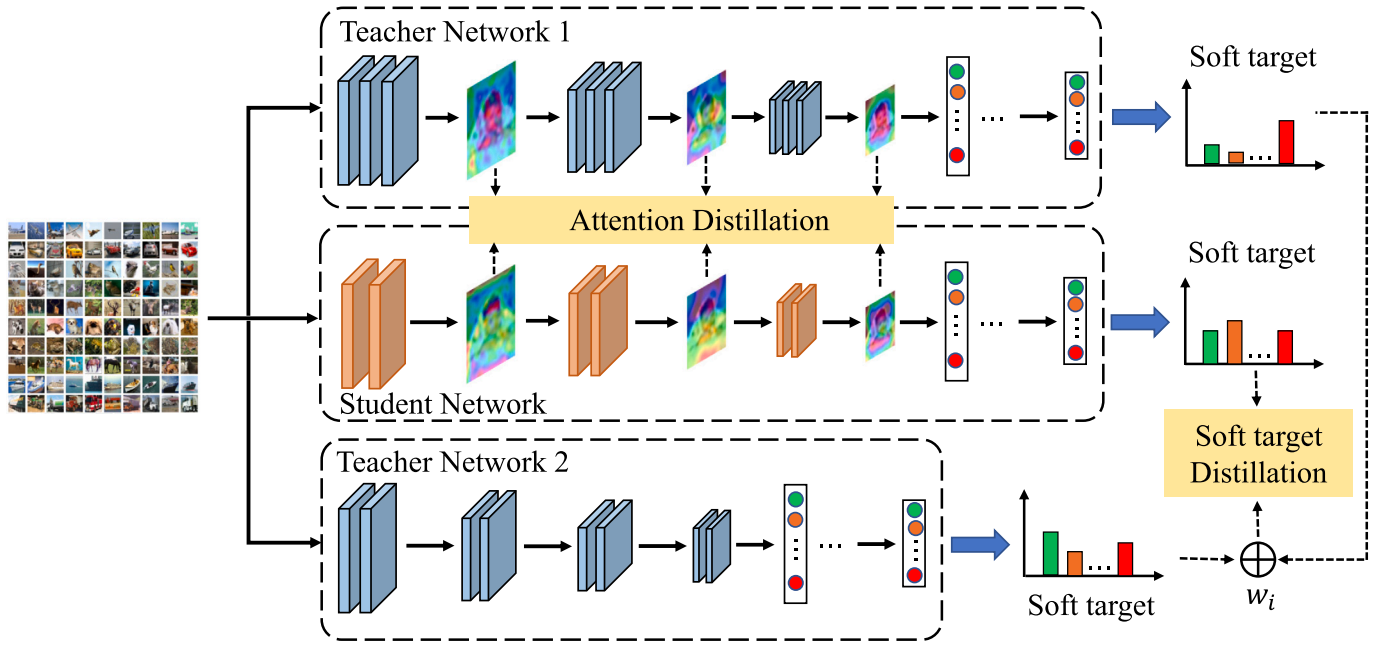


Fig. 1. The framework of proposed CAG-DAKD. The architecture consists of two different teacher networks which conduct various types of knowledge guidance to the student network through the coordinate attention module and adaptive distillation module.

2.2. Multi-teacher knowledge distillation

Multi-teacher KD refers to the knowledge learned from multiple teacher networks provided to the student network. There are numerous ways to transfer knowledge from multiple teacher networks to a student network. The most straightforward approach is to average the outputs of all teacher networks. For example, Dvornik et al. (2019) proposed integrating and averaging the predicted results of multiple teacher networks and then delivering them to the student network through the KL divergence by comparing the predicted probability of the student network to the average predicted probability of multiple teacher networks. This strategy is straightforward and easy to implement, different teacher networks have varied model structures and different training frameworks, so they can provide distinct knowledge. If there is a poor teacher network, it may degrade the learning performance of the guided student network. To release the difficulty, Zhu et al. (2020) proposed a more effective way to weight the knowledge of different teacher networks and elaborately designed a gate network to obtain the weights of different teacher networks, showing more impressive results than those single teacher based networks. Chen, Su, and Zhang (2019) proposed a framework containing two teacher networks trained with different strategies. Although the multi-teacher KD networks can enrich the knowledge, further research is still worthy of further investigating on how to learn more complementary yet comprehensive knowledge from multiple teachers.

2.3. Attention distillation

Attention mechanisms (Niu, Zhong, & Yu, 2021) have been extensively applied in computer vision tasks including image classification, semantic segmentation, object detection and so on. It improves computational efficiency by effectively filtering and removing certain irrelevant information while focusing limited attention on the most critical regions. Therefore, attention mechanisms have been increasingly exploited in KD in recent years. Zagoruyko and Komodakis (2016a) first introduced the attention mechanism into KD and proposed two types of attention distillation: activation-based attention transfer and gradient-based attention transfer. Whereas Zhou et al. (2020) applied the channel attention to the feature maps before each sub-sampling so that the student network could learn the activation attention feature maps at different layers in the teacher network. Ji, Heo, and Park (2021) proposed an attentional-based meta-network by learning the relative similarity between the learned features and applying the identified similarity to control the distillation intensity of all possible teacher–student feature pairs.

3. The proposed method

3.1. Overview of network structure

Inspired by the advantage of multi-teacher KD and attention distillation, we propose a novel CAG-DAKD framework for image classification. Fig. 1 displays the overall flow chart of the proposed CAG-DAKD framework containing two different teacher networks which conduct various types of knowledge guidance to the student network through attention distillation and adaptive distillation. To be more explicit, we extract the coordinate attention features from the first teacher network and then use them as knowledge to facilitate the student network to effectively mine discriminative features. We further integrate the positive prediction distribution of two teacher networks to deliver better output distribution to guide the student network, depicted as the adaptive distillation module.

3.2. Attention distillation

A wide range of research reveals that the channel attention mechanism (Hu, Shen, & Sun, 2018) can considerably strengthen the model's performance. Unfortunately, this method solely takes into account the channel relationship but fails to consider the location information, which is critical for constructing a spatial selective attention map. To tackle this issue, Woo, Park, Lee, and Kweon (2018) proposed the CBAM attention mechanism, which uses a global pooling on channels to code the location information. Albeit the effectiveness, this approach only benefits from capturing the local information. Hou, Zhou, and Feng (2021) proposed a coordinated attention mechanism by incorporating the location coordinates within the channel attention,

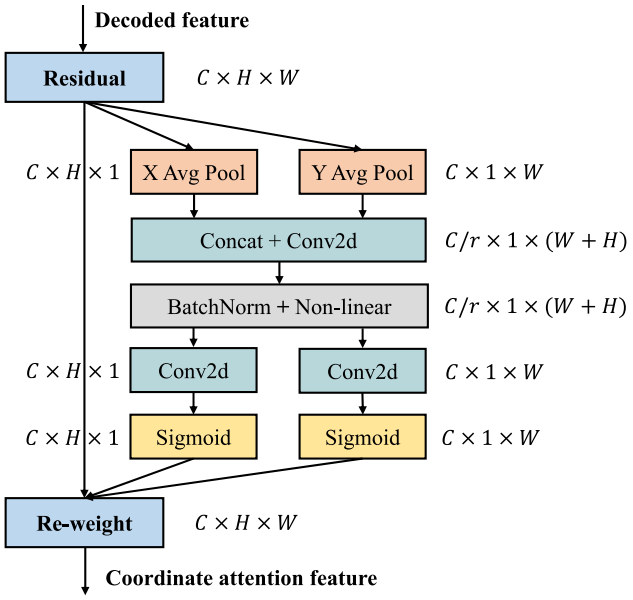


Fig. 2. The illustration of the coordinate attention module.

demonstrating that the proposed coordinated attention mechanism is beneficial to image classification as well as other target detection and semantic segmentation. Motivated by the effectiveness of the coordinated attention mechanism, we introduce it to the corresponding layers in the first teacher network and the student network to obtain more discriminative feature maps with different dimensions. The loss of attention distillation at the ith layer can be expressed as:

$$\mathcal{L}_{AD}^{i}\left(f_T, f_S\right) = \mathcal{D}\left(\mathcal{A}\left(f_T^i\right), r\left(\mathcal{A}\left(f_S^i\right)\right)\right), \tag{1}$$

where f_T^i and f_S^i denote the outputs of the ith block of the teacher network and the student network, respectively, $\mathcal{A}(\cdot)$ is the coordinate attention operation, $r(\cdot)$ is a 1×1 convolution to increase the dimensional size, and $D(\cdot)$ is the MSE loss to determine the similarity between the two attention features.

The computational process of the coordinated attention mechanism is similar to that of channel attention, with the exception that it considers the location information. The spatial coordinate information can be effectively integrated into the feature map by splitting channel attention into two dimensions. For the intermediate feature $X = \left[x_1, x_2, \ldots, x_C\right] \in \mathbb{R}^{C \times H \times W}$, the average pooling kernel of $(H, 1)$ and $(1, W)$ is utilized to compress each channel in both horizontal and vertical directions to obtain the output feature maps along the two directions, which are expressed by:

$$\begin{cases} z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \\ z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \end{cases}, \tag{2}$$

where $z_c^h(h)$ and $z_c^w(w)$ denote the results of horizontal and vertical average pooling, respectively. Through these two transformations, the global receptive field can be well obtained and the position information can be accurately encoded.

Once obtaining $z_c^h(h)$ and $z_c^w(w)$, then we concatenate them to construct an intermediate feature map $f = \delta\left(F_1\left(z^h, z^w\right)\right)$, which contains the spatial information in both the horizontal and vertical directions by a 1×1 convolution. Following that, f is separated into two independent components along the spatial dimension through two 1×1 convolutions, yielding the following representations as:

$$\begin{cases} g^h = \sigma\left(F_h\left(f^h\right)\right) \\ g^w = \sigma\left(F_w\left(f^w\right)\right) \end{cases}, \tag{3}$$

where $\sigma(\cdot)$ represents the sigmoid function, f^h and f^w represent the intermediate feature maps of the spatial information in the horizontal and vertical directions, respectively, and F_h and F_w denotes two different 1×1 convolutions.

By expanding g^h and g^w to the attention weights, the coordinate attention module may generate an attention feature map with the same size as the input, which is described as:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j). \tag{4}$$

Finally, the coordinate attention module accomplishes both horizontal and vertical attentions, which is essentially a type of channel attention. The whole procedure is depicted in Fig. 2.

3.3. Adaptive knowledge distillation

The construction of the proposed adaptive weighting module is depicted in Fig. 3. The logits of the teacher networks are delivered to the softmax layer for the prediction probability. Then we integrate the positive prediction distribution of two teacher networks according to whether the two teacher networks predict correctly and the magnitude of the respective cross-entropy. Finally, the adaptive KD between the soft prediction of the teacher network and the student network is calculated through the KL divergence.

Hinton et al. (2015) originally put out the concept of KD to obtain the soft probability prediction fraction by defining a temperature coefficient τ. as below:

$$q_i = \frac{\exp\left(z_i/\tau\right)}{\sum_j \exp\left(z_j/\tau\right)}, \tag{5}$$

where z_i represents the logits of the network and τ represents the temperature coefficient. Note that a higher temperature means a smoother output of the network with less difference between classes.

The vanilla KD loss is comprised of two components: the cross-entropy loss that measures the classification accuracy of the student work with the ground truth label, and the KL divergence between the soft classification probability of the teacher network and the student network by using the same temperature coefficient τ. Then the final loss is expressed as:

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{CE}\left(X_S, y\right) + (1 - \alpha)\tau^2 \mathcal{H}\left(X_T^\tau, X_S^\tau\right), \tag{6}$$

where α is a hyperparameter for controlling the weight of the soft prediction, $\mathcal{L}_{CE}(\cdot)$ measures the classification accuracy of the student work, and $\mathcal{H}(\cdot)$ measures the KL divergence between the soft classification probability $X_T^\tau = \mathrm{softmax}\left(\frac{z_T}{\tau}\right)$ and $X_S^\tau = \mathrm{softmax}\left(\frac{z_S}{\tau}\right)$.

Considering that the response layers of two different teacher networks are utilized concurrently in our framework, X_S^τ in Eq. (6) indicates the weighting of the soft classification probability of the dual-teacher network, which is expressed as follows:

$$X_T^\tau = w_1 X_{T1}^\tau + w_2 X_{T2}^\tau, \tag{7}$$

where w_1 and w_2 are adaptive weights which vary from $[0, 1]$ with the cross-entropy of the two teacher networks.

Long, Cao, Wang, and Jordan (2018) have proved that the entropy of soft labels in the teacher network can approximately reflect the accuracy of its prediction of output categories. To put it more bluntly, the soft label with a higher entropy will result in an incorrect prediction. This view can also be proved by the study of Kwon, Na, Lee, and Kim (2020), which argues that soft labels with lower entropy are more reliable. Inspired by the philosophy of guided KD proposed by Zhou et al. (2020), we propose to adopt an adaptive distillation strategy to compute the final output of two teacher networks.

When the predicted results of two teacher networks are correct, the teacher network with the lower cross-entropy is assigned a larger

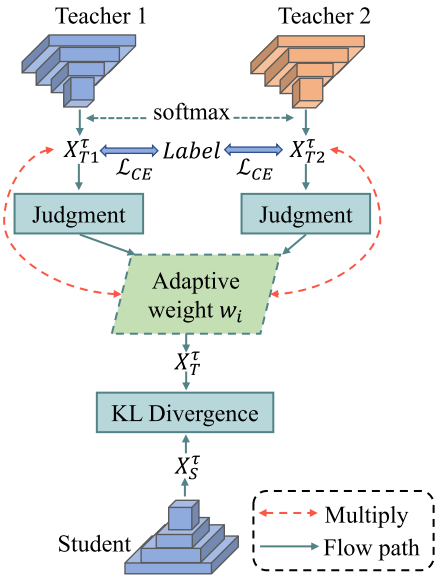


Fig. 3. The illustration of the adaptive distillation module.

weight and vice versa. The adaptive weights w_1 and w_2 are respectively calculated by:

$$\begin{cases} w_1 = 1 - \frac{\mathcal{L}_{CE}(X_{T1},y)}{\mathcal{L}_{CE}(X_{T1},y)+\mathcal{L}_{CE}(X_{T2},y)} \\ w_2 = 1 - \frac{\mathcal{L}_{CE}(X_{T2},y)}{\mathcal{L}_{CE}(X_{T1},y)+\mathcal{L}_{CE}(X_{T2},y)} \end{cases}, \tag{8}$$

where $\mathcal{L}_{CE}(\cdot)$ represents the cross-entropy function that measures the classification accuracy of the teacher network, and w_1 and w_2 represent the adaptive weights of the two teacher networks, respectively.

When the result predicted by only one teacher network is correct, the weight corresponding to this teacher network is assigned to 1 while the weight of the other teacher network is assigned to 0. Besides, when both teacher networks predict erroneously, the weights of both teacher networks are assigned to 0. In this case, the student network is only supervised by the cross-entropy loss for classification.

As a result, the final adaptive KD loss of the proposed dual-teacher network is represented as:

$$\mathcal{L}_{AKD} = \mathcal{L}_{CE}\left(X_S, y\right) + \alpha\tau^2\mathcal{H}\left(w_1 X_{T1}^{\tau} + w_2 X_{T2}^{\tau}, X_S^{\tau}\right). \tag{9}$$

3.4. Total loss function

By combining the attention distillation loss $\mathcal{L}_{AD}$ and the adaptive KD loss $\mathcal{L}_{AKD}$, the overall loss designed for optimization in the training stage is formulated as below:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \underbrace{\mathcal{L}_{CE}\left(X_S, y\right)}_{\text{CE loss}} + \underbrace{\alpha\tau^2\mathcal{H}\left(w_1 X_{T1}^{\tau} + w_2 X_{T2}^{\tau}, X_S^{\tau}\right)}_{\text{KL distillation}} \\ & + \underbrace{\beta\mathcal{D}\left(\mathcal{A}\left(f_T^i\right), r\left(\mathcal{A}\left(f_S^i\right)\right)\right)}_{\text{attention distillation}} \end{aligned}, \tag{10}$$

where α and β are the trade-off parameters to control the strength of the adaptive KD and the attention distillation.

3.5. Summary of proposed CAG-DAKD algorithm

The overall procedure of the proposed CAG-DAKD for image classification is summarized in Algorithm 1.

Algorithm 1 : CAG-DAKD model

Input: The training data $\mathbb{D} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{|D|}$, the epoch number N, the batch size b, the learning rate γ.
Output: The student network parameter θ.

for epoch = 1,2,...,N **do**
  for a batch-size of training data $\mathbb{D} = \left\{\left(x_i, y_i\right)\right\}_{i=1}^{b}$ **do**
    Initialize $\mathcal{L}_{\text{batch}} = 0$,
    $logit_{T_1}$, $feat_{T_1}^i = T_1(x)$,
    $logit_{T_2} = T_2(x)$,
    $logit_S$, $feat_S^i = S(x)$
    1)Calculate the attention distillation loss:
    Calculate the coordinate attention by Eq. ((2)-(4)).
    Calculate the attention distillation loss by Eq. (1).
    2)Calculate the adaptive distillation loss:
    $out_{T_i} = F.softmax(logit_{T_i})$, $i \in \{1,2\}$
    $pred_{T_i} = top1(out_{T_i})$, $i \in \{1,2\}$
    if $pred_{T_1}$==True and $pred_{T_2}$==True **then**
      Calculate the respective CrossEntropyLoss.
      Calculate the adaptive weights by Eq. (8).
    else if $pred_{T_1}$==True or $pred_{T_2}$==True **then**
      $w_i = \begin{cases} 1, & pred_{T_i} == \text{True} \\ 0, & pred_{T_i} == \text{False} \end{cases}$
    else
      $w_i = 0$
    end if
    Calculate the soft target loss by Eq. (5) with $\tau = 4$.
    Calculate the adaptive distillation loss by Eq. (9).
    3)Calculate the total loss:
    Calculate the total loss by Eq. (10).
    Update $\mathcal{L}_{batch} = \mathcal{L}_{batch} + \frac{1}{b}\mathcal{L}_{\text{total}}$.
  end for
  Update the network parameter $\theta = \theta - \gamma \cdot \nabla_\theta \mathcal{L}_{batch}$.
end for

4. Experimental results and analysis

In the section, we first describe our experimental configuration and the validation datasets for the performance evaluation. Next, the efficiency of the proposed approach is validated on various datasets by comparing it to previous distillation methods under various network settings. Finally, we do thorough ablation experiments to demonstrate the contribution of each component and the selection of hyperparameters in the proposed method.

4.1. Datasets

In our experiments, we validate the advantages of our proposed method on three commonly used datasets, i.e., CIFAR10, CIFAR100, and ImageNet datasets. The CIFAR10 dataset (Krizhevsky & Hinton, 2009) consists of 60,000 32 × 32 color images, of which 50,000 are used as the training images and the rest as the test images. There are 6000 images in each of the ten categories. The CIFAR100 dataset (Krizhevsky & Hinton, 2009) is similar to CIFAR10, in which 60,000 32 × 32 color images and 100 categories are contained. Each category has 600 images, 500 of which are used as training images and the rest as test images.

In addition, since all the images in the CIFAR10 and CIFAR100 datasets are the size of 32 × 32, they cannot represent images of natural scenes. We conduct experiments on the ImageNet dataset as well to verify the effectiveness of our proposed method. Here, we use the ILSVRC 2012 dataset (Russakovsky et al., 2015) which is a subset of the ImageNet dataset and consists of 1000 categories with about 1000 images per category. In total, approximately 1.2 million training

images, 50,000 of which are used for validation and 150,000 of which are used for test. Classification on ImageNet is a more difficult task than that on CIFAR100 due to the increased resolution and greater number of pictures.

4.2. Implementation details

In the experiments, we choose various types of networks, including WideResNet (Zagoruyko & Komodakis, 2016b), ResNet (He, Zhang, Ren, & Sun, 2016), and VGG (Simonyan & Zisserman, 2014), as our backbone networks and implement all experiments on a PyTorch platform deployed with a NVIDIA 3080 GPU. We incorporate the coordinate attention modules into all the intermediate layers of the first teacher model and the student model. The specific reasons for this configuration will be discuss in the subsequent ablation study section. A conventional data augmentation strategy (including padding, random crop, and horizontal flip) is adopted to normalize the input images with means and standard deviations for the CIFAR10 and CIFAR100 datasets (Krizhevsky & Hinton, 2009). Each model undergoes 200 iterations of training with 128 images in a mini-batch. The Stochastic Gradient Descent (SGD) optimizer is used for training with the following parameters: momentum is 0.9, the initial learning rate is 0.1 and decreased at 100 and 150 iterations respectively, and the weight decay factor is 5.0×10^{-4}. We follow the same evaluation metric as in Tian, Krishnan, and Isola (2019) to ensure that our results are comparable. As for the ILSVRC 2012 dataset (Russakovsky et al., 2015), all the images are cropped to the size of 224×224 for training and evaluation. Each model undergoes 200 iterations of training with 128 images in a mini-batch. Initially with a learning rate of 0.1 and increased by a factor of 0.1 after 30, 60, and 80 epochs. The parameter settings for the SGD optimizer are the same as for the CIFAR dataset.

4.3. Comparisons with state-of-the-art methods

Evaluation on CIFAR10 and CIFAR100: In this section, we select three groups of different network architectures to validate the generalizability of the proposed CAG-DAKD. Table 1 describes three different experimental settings, including the types of networks in each group, the experimental parameters, and the validation accuracy on the CIFAR10 and CIFAR100 datasets of each network.

To demonstrate that the method we have proposed is effective, we evaluate it in light of the following state-of-the-art methods: (1) Soft Target (Hinton et al., 2015), which transfers the dark knowledge by employing the softmax with a temperature coefficient, (2) Fitnets (Romero et al., 2014), which extracts the features from the different layers of the teacher network to instruct the corresponding layers of the student network, (3) Attention Transfer (Zagoruyko & Komodakis, 2016a), which proposed to combine attention with KD and leveraged the attentional features of teacher networks as knowledge, (4) Factor Transfer (Kim, Park, & Kwak, 2018), which performs a coding and decoding process on the output of the model and extracts the factor matrix, using the factor of the teacher network to guide the factor of the student network, (5) Activation Boundaries (Heo, Lee, Yun, & Choi, 2019), which makes the activation boundaries of the neurons in the teacher network layer as close as possible to those of the student network, (6) Relation KD (Park, Kim, Lu, & Cho, 2019), which regards the relationship between different samples in a mini-batch as knowledge, (7) CRD (Tian et al., 2019), which introduces contrastive learning into KD. For the same sample, the representations of the teacher and student networks are as close as possible and vice versa, (8) ICKD-C (Liu et al., 2021), which uses the similarity relationship between different features and different channels as knowledge to help student network, (9) SAKD (Song, Chen, Ye and Song, 2022), which argues that it necessary to set an adaptive distillation position for each sample, and then proposes the spot-adaptive distillation strategy. Following the network settings in Table 1, we conduct experiments on three groups of different network architectures. We compare the results of knowledge transfer on the student network with each teacher network by using different distillation methods alone. The results of the Top-1 accuracy on the CIFAR10 and CIFAR100 datasets are shown in Tables 2 and 3, respectively.

Table 1
Experiments setting with various network architecture on CIFAR10 and CIFAR100 datasets. The table gives the average Top-1 testing accuracy over 3 runs.

Setup	Network	Parameters	Testing accuracy (%)	
			CIFAR10	CIFAR100
(a)	WRN28-4(teacher1)	5.87 MB	95.46	78.64
	WRN40-2(teacher2)	2.26 MB	94.81	76.48
	WRN16-2(student)	1.48 MB	93.88	73.42
(b)	ResNet110(teacher1)	1.74 MB	94.59	73.76
	ResNet56(teacher2)	0.86 MB	94.10	73.08
	ResNet20(student)	0.28 MB	92.42	69.14
(c)	Vgg13(teacher1)	9.46 MB	94.22	74.32
	ResNet18(teacher2)	11.22 MB	95.13	76.89
	Vgg8(student)	3.97 MB	91.92	70.63

Table 2 tabulates the contrastive results of the Top-1 accuracy obtained from three groups of different networks on the CIFAR10 dataset. In each group of network architectures, Teacher1 and Teacher2 denote two different teacher networks, respectively, and Student represents the baseline. As can be seen, our proposed CAG-DAKD method outperforms the state-of-the-art in all the settings of (a), (b), and (c), indicating its remarkable effectiveness in improving the performance of the student network. Specifically, in the setup of (c), our proposed method achieves 95.31% accuracy on the CIFAR10 dataset which is 1.54% higher than the state-of-art method CRD+SAKD (Song, Chen et al., 2022). At the same time, Table 3 compares the Top-1 accuracy on the CIFAR100 dataset. One can find that our proposed method gains the highest accuracy in all three settings. In particular, our method achieves an improvement of at least 0.73% when compared to ICKD-C (Liu et al., 2021) in the setup of (b) and obtains a 0.94% improvement in the setup of (c) compared to the most competitive method.

Besides, we compare several other multi-teacher approaches (Chen et al., 2019; Dvornik et al., 2019; Kwon et al., 2020) on the CIFAR100 dataset to clarify the advantage of the proposed CAG-DAKD. Among them, Dvornik et al. (2019) proposed to directly utilize the average predicted scores of multiple teacher networks as knowledge to supervise the student network. Whereas Kwon et al. (2020) explored the responses of multiple teachers as knowledge to determine each teacher's weight dynamically. Chen et al. (2019) developed two teacher networks for image classification, among which one teacher transfers knowledge from the response layer while the other delivers the knowledge from the feature layer. The comparison results are shown in Table 4. It is clear that CAG-DAKD exceeds the other predecessors with three multi-teacher distillation strategies.

To further demonstrate the effectiveness of the approach proposed more intuitively, t-SNE visualization is used to visualize ten categories from the CIFAR100 test set, each of which contains 100 test samples. Fig. 4 demonstrates the comparison between the visualization results trained with the baseline and our proposed CAG-DAKD. The "Baseline" denotes the student network trained without any distillation, while "CAG-DAKD" denotes the student network learned by our proposed method. In terms of the visualization results demonstrated in the two figures, one can be find that the result on the right is in more compact clusters of each class because the proposed CAG-DAKD is capable of providing more informative knowledge to the student networks and therefore can improve the discrimination capacity of the feature representations learned from the student network.

Evaluation on ImageNet: We select ResNet34 as the first teacher network, ResNet50 as the other teacher network, and ResNet18 as the student network to verify the superiority of the proposed CAG-DAKD method over other KD approaches. We mainly compare the

Table 2
Comparison of Top-1 test accuracy (%) among various methods conducted on the CIFAR10 dataset. The two columns of the comparison method indicate the student network trained with single teacher network separately. The best results in bold and all results are the average over three runs.

Teachers	WRN28-4	WRN40–2	ResNet110	ResNet56	Vgg13	ResNet18
Student	WRN16-2		ResNet20		Vgg8	
Soft target (Hinton et al., 2015)	94.40	94.51	93.29	92.93	92.85	92.97
FitNet (Romero et al., 2014)	93.96	94.03	93.53	92.95	92.27	92.13
Attention transfer (Zagoruyko & Komodakis, 2016a)	94.23	94.15	93.39	93.23	92.13	92.95
Factor transfer (Kim et al., 2018)	94.36	94.45	93.21	93.47	92.21	91.62
Activate boundary (Heo et al., 2019)	93.51	93.27	92.30	93.29	91.57	92.04
Relation KD (Park et al., 2019)	94.60	94.46	93.30	93.46	93.32	92.51
CRD (Tian et al., 2019)	94.04	93.82	92.97	92.53	93.50	93.56
ICKD-C (Liu et al., 2021)	94.11	94.11	92.89	92.45	93.58	93.14
CRD+SAKD (Song, Chen et al., 2022)	94.23	94.05	93.11	92.80	93.69	93.77
Ours	**94.96**		**93.49**		**95.31**	

Table 3
The Top-1 testing accuracy rate (%) of various methods on the CIFAR100 dataset. The two columns of the comparison method indicate the student network trained with single teacher network separately. The best results in bold and all results are the average over three runs.

Teachers	WRN28-4	WRN40–2	ResNet110	ResNet56	Vgg13	ResNet18
Student	WRN16-2		ResNet20		Vgg8	
Soft Target (Hinton et al., 2015)	74.92	74.77	70.84	70.99	72.73	72.59
FitNet (Romero et al., 2014)	74.97	74.96	70.76	71.53	73.40	70.96
Attention Transfer (Zagoruyko & Komodakis, 2016a)	74.85	75.23	70.90	71.57	73.07	71.73
Factor Transfer (Kim et al., 2018)	75.41	75.10	70.44	71.51	73.14	69.02
Activate Boundary (Heo et al., 2019)	75.29	72.05	71.14	71.26	72.98	70.57
Relation KD (Park et al., 2019)	75.01	74.84	70.92	71.48	73.05	71.13
CRD (Tian et al., 2019)	75.68	75.51	71.50	71.72	74.39	73.29
ICKD-C (Liu et al., 2021)	75.80	75.57	71.91	71.69	73.88	73.32
CRD+SAKD (Song, Chen et al., 2022)	75.77	75.63	71.69	71.80	74.63	73.41
Ours	**76.02**		**72.64**		**75.57**	

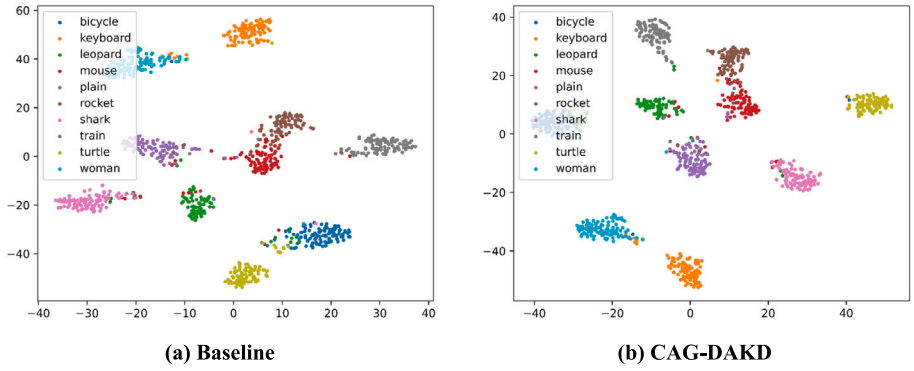


Fig. 4. t-SNE visualization of the student network in setup (a) on the CIFAR100 dataset. The 'Baseline' denotes that training student network without distillation, while 'CAG-DAKD' denotes that training student network with our proposed method.

performance under the following situations, including training the student network with two different teacher networks with original KD (Hinton et al., 2015) and the first teacher network with attention transfer (Zagoruyko & Komodakis, 2016a). The results are represented in Table 5.

As compared the Top-1 and Top-5 accuracies on the ImageNet dataset listed in Table 5, one can find that the proposed CAG-DAKD obtains superior performance than other situations. The following observations are also supported by the results. First, compared with the results obtained from only one teacher network S_T1_KD, the gains of Top-1 and Top-5 accuracies yielded by ours are up to 1.06% and 1.18%, respectively. It can verify that integrating the positive prediction distribution of two teacher networks according to whether the two teacher networks predict correctly and the magnitude of the cross-entropy is conducive to improving the classification performance of the student network. Second, compared with the model S_T1_AT that only uses the attention transfer, our method obtains the gains of Top-1 and Top-5 accuracies up to 0.92% and 0.83%, respectively. The impressive results are attributed to discriminative and complementary knowledge learned by our proposed CAG-DAKD network.

To further verify that our proposed CAG-DAKD method can localize the discriminative regions of objects, we visualize the attention feature maps from the student network trained with the baseline and our method leveraging Score-CAM (Wang et al., 2020), which provides improved visual performance and decision-making impartiality. Fig. 5 depicts several samples from ImageNet. One can see that the student network trained with our method attends to concentrate on the subject of interest compared to the baseline. Taking the first column as an example, we can observe that our proposed method could emphatically focus on the horn regions of the hartebeest, benefiting from learning more discriminative feature representations for classification.

4.4. Ablation study

In this subsection, we conduct a group of ablation studies on the CIFAR100 dataset to investigate the contribution of each component in the proposed CAG-DAKD. Note that the first teacher network here is ResNet110, the second teacher network is ResNet56, and the student network is ResNet20. The comparison results are tabulated in Table 6,

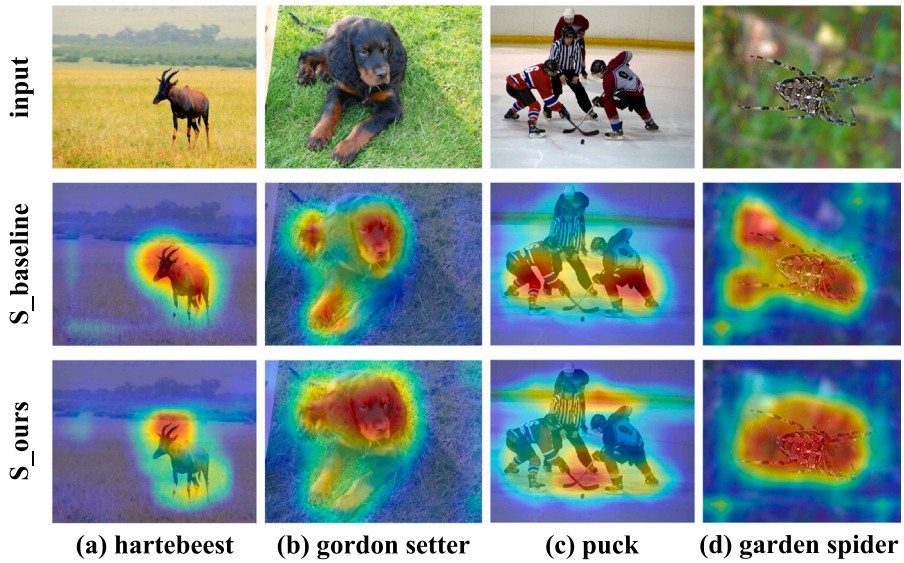


Fig. 5. Visualizations of class activation mapping of various samples from ImageNet. Note that "S_baseline" in the second row and "S_ours" in the third row denote the student network trained with baseline and our method, respectively.

Table 4
Experimental results on CIFAR100 dataset compared with other multi-teacher methods. The response-based knowledge and feature-based knowledge are abbreviated as 'ResK' and 'FeaK' respectively. Better performance is shown in bold.

Methods	Types of knowledge		Top-1 accuracy	Top-5 accuracy
	ResK	FeaK		
Dvornik et al. (2019)	✓	✗	71.26	93.02
Kwon et al. (2020)	✓	✗	71.70	93.14
Chen et al. (2019)	✓	✓	71.58	93.11
Ours	✓	✓	**72.64**	**93.43**

Table 5
Experimental results on ImageNet dataset when training the student network with different teachers of original KD and attention transfer. Better performance is shown in bold.

Model	Top-1 accuracy	Top-5 accuracy
Teacher1(T1)	73.79	91.62
Teacher2(T2)	75.85	92.96
S_baseline	69.52	89.07
S_T1_KD	70.38	89.24
S_T1_AT	70.52	89.43
S_T2_KD	70.26	89.16
S_T1+T2_Ours	**71.44**	**90.26**

Table 6
Ablation experiment on each components and loss function of the proposed method. The best results in bold are the average over three runs.

Model	Top-1 accuracy	Top-5 accuracy
CAG-DAKD	**72.64**	**93.43**
w/o attention distillation	72.11(↓ 0.53)	93.17(↓ 0.26)
w/o teacher network2	71.17(↓ 0.47)	92.98(↓ 0.45)
w/o adaptive distillation	71.94(↓ 0.70)	93.11(↓ 0.32)

where "w/o attention distillation" denotes the student network trained without the coordinate attention distillation, "w/o teacher network2" denotes the student network trained without the second teacher network which only acquires knowledge from the first teacher network with the attention distillation and the soft label, and "w/o adaptive distillation" denotes the student network replaces the adaptive distillation with the average soft labels predicted by the two teacher networks as the final soft label.

As illustrated in Table 6, when we remove the coordinate attention module and other experimental variables being constant, the Top-1 testing accuracy of student network decreases by 0.53%. It confirms that the coordinate attention module provides more discriminative supervision information to promote the student network learning. Additionally, when the second teacher network is removed, the performance of the student network sharply declines, highlighting the necessity of the proposed dual-teacher distillation architecture. Furthermore, when the adaptive distillation module is replaced by the average predictions of the two teachers, the performance of the student network decreases by 0.7%, indicating the effectiveness of the proposed adaptive distillation module. It helps the student network learn complementary knowledge from the two teacher networks.

Table 7
Ablation experiment on the number of the coordinate attention modules. The best results in bold are the average over three runs.

Attention position			Top-1 accuracy	Top-5 accuracy
Block1	Block2	Block3		
✓	✓	✓	**72.64**	**93.43**
✗	✓	✓	72.29(↓ 0.35)	93.13(↓ 0.30)
✓	✗	✓	72.51(↓ 0.13)	93.15(↓ 0.28)
✓	✓	✗	72.17(↓ 0.47)	93.07(↓ 0.36)

Besides, we implement another ablation experiment to investigate how the number of applied coordinate attention modules affects the performance. The teacher and student networks are chosen in the same way as above. Table 7 lists the compared results, and it is evident that the performance drops slightly when removing the coordinate attention module of the second block. Specifically, the performance of our proposed CAG-DAKD method without considering the knowledge from the first block and the final block degrades by 0.35% and 0.47%, respectively. The results imply that both the deep and the shallow features in the teacher are conducive to transferring more discriminative features to the student.

4.5. Robustness study

It is well known that practical images often undergo various degradations, noise effects, or variabilities in the process of imaging. To verify the robustness of the proposed method, we follow the similar data augmentation method proposed in Hong, Yokoya, Chanussot, and Zhu (2019) to simulate the degradation of images by applying a random Gaussian blurring on the CIFAR100 dataset. To this end, we apply a

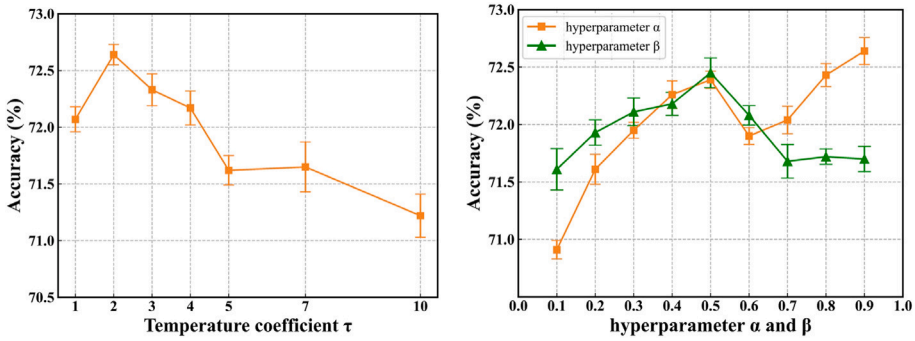


Fig. 6. Hyperparameter analysis of the student network in setup (b) on the CIFAR100 dataset. Average over 3 runs.

Table 8
Robustness experiments on the degraded CIFAR100 image dataset by applying 3 × 3 Gaussian blurring with a randomly determined standard deviation within the range of 0.1 to 2.0. The best results in bold are the average over three runs.

Model	Top-1 accuracy (%)
ResNet110_Blur(T1)	72.19(↓ 1.37)
ResNet56_Blur(T2)	71.85(↓ 1.23)
ResNet20_Blur(S)	67.34(↓ 1.80)
S_T1_KD	69.71(↑ 2.37)
S_T2_KD	69.57(↑ 2.23)
S_T1+T2_Ours	**71.73**(↑ **4.39**)

3 × 3 Gaussian kernel with a randomly determined standard deviation falling within the range of 0.1 to 2.0 to the training images. For simplicity, we denote the first teacher as ResNet110_Blur(T1), the second teacher as ResNet56_Blur(T2), and the student as ResNet20_Blur(S), respectively. Correspondingly, we use S_T1_KD and S_T2_KD to represent the models transferring the knowledge from the first teacher and the second teacher with the original KD strategy, respectively. S_T1+T2_Ours is the dual-teacher model trained by the proposed CAG-DAKD. As the compared results tabulated in Table 8, both the teacher network and the student network degrade to a certain degree when applying random Gaussian blur on the database. When distilling the student network independently with one teacher network with original KD, the performance of the student network yields the improvements by 2.37% and 2.23%, respectively. Noticeably, the performance of the student network trained by the proposed CAG-DAKD scheme improves by 4.39%, which almost approximates to the performance obtained by the two teacher networks.

4.6. Hyperparameter analysis

The selection of hyperparameters is also a key factor affecting the performance of the student network. Several hyperparameters seem worthwhile of further investigation in the proposed CAG-DAKD method: (1) The temperature coefficient τ used to soften the probability prediction in Eq. (5), and (2) The hyperparameter α and β to weight the influence of the loss. We adopt the setup (b) on CIFAR100 for analysis.

Effects of Temperature Coefficient τ. To validate how the temperature coefficient τ affects the performance, we perform an empirical study by fixing the parameter to 1, 2, 3, 4, 5, 7, and 10 to seek an optimal configuration. From the left panel in Fig. 6, it can be concluded that either extremely large or small temperature value will reduce the classification accuracy. Particularly the student network achieves the best performance when $\tau = 2$. Based on the validation experiment, we suggest setting τ to 2 for all our experiments.

Effects of Hyperparameter α and β. To investigate how the hyperparameter α and β influence the student network's performance, we test the Top-1 accuracy of our proposed model by varying the values of α and β from 0 to 1 at an interval of 0.1, respectively. As shown in the right panel of Fig. 6, it can be seen that our method is susceptible to the values of hyperparameter α and β. Specifically, a higher value of the parameter α tends to harvest better performance while the best accuracy is achieved at $\beta = 0.5$.

5. Conclusion

In this paper, we have presented a novel KD method called CAG-DAKD to deliver more discriminative and complementary knowledge from two teacher networks to a lightweight student network. In the intermediate layers of the first teacher network and the student network with comparable structures, we utilized the coordinate attention mechanism to explore coordinate attention features at different positions of the first teacher network as knowledge to supervise the learning of the student network. Furthermore, we integrated the positive prediction distribution of two teacher networks via an adaptive weighting strategy to transfer better output distribution guidance to the student network. Thorough experiments on three common datasets have confirmed that the proposed CAG-DAKD have shown the impressive superiority over those compared approaches.

Albeit excellent performance and promising applicability achieved by our proposed CAG-DAKD, there are still several nonnegligible limitations. First, the dual-teacher network framework may require more considerable computational cost in the training stage and more convergence time. Second, the student's performance is fundamentally affected by the two designated teacher networks. If chosen inappropriately, they may not be able to provide complementary information. In future work, we will explore more advanced feature distillation methods, such as adversarial contrastive distillation (Bai, Zhao, & Wen, 2023) and knowledge pre-training knowledge distillation (Song, Yang, Wang, & Xu, 2023) to further boost up the performance of our dual-teacher KD framework.

CRediT authorship contribution statement

Dongtong Ma: Writing – original draft, Revised version preparation. **Kaibing Zhang:** Methodology - Proponents of major academic ideas and supervision. **Qizhi Cao:** Writing – review & editing. **Jie Li:** Writing - Polishing the English presentation. **Xinbo Gao:** Conceptualization, Resources.

Declaration of competing interest

All co-authors have seen and agree with the contents of the manuscript and confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We hope that the editorial board will agree on the interest of this study.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62441601, 62050175, 61971339, in part by the Textile Intelligent Equipment Information and Control Innovation Team of Shaanxi Innovation Ability Support Program under Grant 2021TD-29, and in part by the Textile Intelligent Equipment Information and Control Innovation Team of Shaanxi Innovation Team of Universities.

References

Bai, Tao, Zhao, Jun, & Wen, Bihan (2023). Guided adversarial contrastive distillation for robust students. *IEEE Transactions on Information Forensics and Security*.

Blalock, Davis, Gonzalez Ortiz, Jose Javier, Frankle, Jonathan, & Guttag, John (2020). What is the state of neural network pruning? *Proceedings of Machine Learning and Systems*, *2*, 129–146.

Chen, Lin, Jiang, Xue, Liu, Xingzhao, & Zhou, Zhixin (2021). Logarithmic norm regularized low-rank factorization for matrix and tensor completion. *IEEE Transactions on Image Processing*, *30*, 3434–3449.

Chen, Xingjian, Su, Jianbo, & Zhang, Jun (2019). A two-teacher framework for knowledge distillation. In *International symposium on neural networks* (pp. 58–66). Springer.

Dai, Xing, Jiang, Zeren, Wu, Zhao, Bao, Yiping, Wang, Zhicheng, Liu, Si, et al. (2021). General instance distillation for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7842–7851).

Dvornik, Nikita, Schmid, Cordelia, & Mairal, Julien (2019). Diversity with cooperation: Ensemble methods for few-shot classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3723–3731).

El-Dahshan, El-Sayed A, Bassiouni, Mahmoud M, Khare, Smith K, Tan, Ru-San, & Acharya, U Rajendra (2024). ExHyptNet: An explainable diagnosis of hypertension using EfficientNet with PPG signals. *Expert Systems with Applications*, *239*, Article 122388.

Fei, Wen, Dai, Wenrui, Li, Chenglin, Zou, Junni, & Xiong, Hongkai (2022). General bitwidth assignment for efficient deep convolutional neural network quantization. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(10), 5253–5267.

Feng, Hao, Wang, Nian, & Tang, Jun (2021). Deep Weibull hashing with maximum mean discrepancy quantization for image retrieval. *Neurocomputing*, *464*, 95–106.

Fu, Shipeng, Li, Zhen, Liu, Zitao, & Yang, Xiaomin (2021). Interactive knowledge distillation for image classification. *Neurocomputing*, *449*, 411–421.

Gou, Jianping, Yu, Baosheng, Maybank, Stephen J, & Tao, Dacheng (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789–1819.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heo, Byeongho, Lee, Minsik, Yun, Sangdoo, & Choi, Jin Young (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3779–3787).

Hinton, Geoffrey, Vinyals, Oriol, & Dean, Jeff (2015). Distilling the knowledge in a neural network. *Computer Science*, *14*(7), 38–39.

Hong, Danfeng, Yokoya, Naoto, Chanussot, Jocelyn, & Zhu, Xiao Xiang (2019). An augmented linear mixing model to address spectral variability for hyperspectral unmixing. *IEEE Transactions on Image Processing*, *28*(4), 1923–1938.

Hong, Danfeng, Zhang, Bing, Li, Xuyang, Li, Yuxuan, Li, Chenyu, Yao, Jing, et al. (2023). SpectralGPT: Spectral foundation model. arXiv:2311.07113.

Hou, Qibin, Zhou, Daquan, & Feng, Jiashi (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13713–13722).

Hu, Jie, Shen, Li, & Sun, Gang (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Huang, Kun, Guo, Xin, & Wang, Meng (2024). Towards efficient pre-trained language model via feature correlation distillation. *Advances in Neural Information Processing Systems*, *36*.

Huang, Zhihao, Su, Lumei, Wu, Jiajun, & Chen, Yuhan (2023). Rock image classification based on EfficientNet and triplet attention mechanism. *Applied Sciences*, *13*(5), 3180.

Ji, Mingi, Heo, Byeongho, & Park, Sungrae (2021). Show, attend and distill: Knowledge distillation via attention-based feature matching. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7945–7952).

Kim, Jangho, Park, SeongUk, & Kwak, Nojun (2018). Paraphrasing complex network: Network compression via factor transfer. *Advances in Neural Information Processing Systems*, *31*, 2765–2774.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, *1*(4).

Kwon, Kisoo, Na, Hwidong, Lee, Hoshik, & Kim, Nam Soo (2020). Adaptive knowledge distillation based on entropy. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7409–7413). IEEE.

Liu, Li, Huang, Qingle, Lin, Sihao, Xie, Hongwei, Wang, Bing, Chang, Xiaojun, et al. (2021). Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8271–8280).

Liu, Ze, Ning, Jia, Cao, Yue, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, et al. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3202–3211).

Long, Mingsheng, Cao, Zhangjie, Wang, Jianmin, & Jordan, Michael I (2018). Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, *31*, 1647–1657.

Luo, Weihao, Zeng, Zezhen, & Zhong, Yueqi (2024). A progressive distillation network for practical image-based virtual try-on. *Expert Systems with Applications*, *246*, Article 123213.

Ma, Ye, Jiang, Xu, Guan, Nan, & Yi, Wang (2023). Anomaly detection based on multi-teacher knowledge distillation. *Journal of Systems Architecture*, *138*, Article 102861.

Manzari, Omid Nejati, Ahmadabadi, Hamid, Kashiani, Hossein, Shokouhi, Shahriar B, & Ayatollahi, Ahmad (2023). MedViT: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, *157*, Article 106791.

Niu, Zhaoyang, Zhong, Guoqiang, & Yu, Hui (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62.

Park, Wonpyo, Kim, Dongju, Lu, Yan, & Cho, Minsu (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3967–3976).

Pham, Cuong, Nguyen, Van-Anh, Le, Trung, Phung, Dinh, Carneiro, Gustavo, & Do, Thanh-Toan (2024). Frequency attention for knowledge distillation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2277–2286).

Romero, Adriana, Ballas, Nicolas, Kahou, Samira Ebrahimi, Chassang, Antoine, Gatta, Carlo, & Bengio, Yoshua (2014). Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Simonyan, Karen, & Zisserman, Andrew (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Song, Jie, Chen, Ying, Ye, Jingwen, & Song, Mingli (2022). Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, *31*, 3359–3370.

Song, Yaozhe, Tang, Hongying, Meng, Fangzhou, Wang, Chaoyi, Wu, Mengmeng, Shu, Ziting, et al. (2022). A transformer-based low-resolution face recognition method via on-and-offline knowledge distillation. *Neurocomputing*, *509*, 193–205.

Song, Yaguang, Yang, Xiaoshan, Wang, Yaowei, & Xu, Changsheng (2023). Recovering generalization via pre-training-like knowledge distillation for out-of-distribution visual question answering. *IEEE Transactions on Multimedia*.

Tian, Yonglong, Krishnan, Dilip, & Isola, Phillip (2019). Contrastive representation distillation. arXiv preprint arXiv:1910.10699.

Tzelepi, M., Passalis, N., & Tefas, A. (2021). Online subclass knowledge distillation. *Expert Systems with Applications*, *181*, Article 115132.

Wang, Yu, Gui, Guan, Gacanin, Haris, Ohtsuki, Tomoaki, Dobre, Octavia A, & Poor, H Vincent (2021). An efficient specific emitter identification method based on complex-valued neural networks and network compression. *IEEE Journal on Selected Areas in Communications*, *39*(8), 2305–2317.

Wang, Haofan, Wang, Zifan, Du, Mengnan, Yang, Fan, Zhang, Zijian, Ding, Sirui, et al. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 24–25).

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, & Kweon, In So (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision* (pp. 3–19).

Yang, Ling, Zhang, Zhilong, Song, Yang, Hong, Shenda, Xu, Runsheng, Zhao, Yue, et al. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 1–39.

Yim, Junho, Joo, Donggyu, Bae, Jihoon, & Kim, Junmo (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4133–4141).

You, Shan, Xu, Chang, Xu, Chao, & Tao, Dacheng (2017). Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1285–1294).

Yuan, Fei, Shou, Linjun, Pei, Jian, Lin, Wutao, Gong, Ming, Fu, Yan, et al. (2021). Reinforced multi-teacher selection for knowledge distillation. *Vol. 35*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 14284–14291).

Zagoruyko, Sergey, & Komodakis, Nikos (2016a). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Zagoruyko, Sergey, & Komodakis, Nikos (2016b). Wide residual networks. arXiv preprint arXiv:1605.07146.

Zhou, Peiyong, Aysa, Alimjan, Ubul, Kurban, et al. (2024). Research on knowledge distillation algorithm based on Yolov5 attention mechanism. *Expert Systems with Applications*, *240*, Article 122553.

Zhou, Zaida, Zhuge, Chaoran, Guan, Xinwei, & Liu, Wen (2020). Channel distillation: Channel-wise attention for knowledge distillation. arXiv preprint arXiv:2006.01683.

Zhu, Jieming, Liu, Jinyang, Li, Weiqi, Lai, Jincai, He, Xiuqiang, Chen, Liang, et al. (2020). Ensembled CTR prediction via knowledge distillation. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 2941–2958).