Adaptive Dual Guidance Knowledge Distillation

Tong Li, Long Liu*, Kang Liu, Xin Wang, Bo Zhou, Hongguang Yang, Kai Lu

Xi'an University of Technology, Xi'an, 710048, China
1230310013@stu.xaut.edu.cn, liulong@xaut.edu.cn, KangLiu@stu.xaut.edu.cn, 1220311023@stu.xaut.edu.cn,
1220311012@stu.xaut.edu.cn, 1230313028@stu.xaut.edu.cn, 1240310013@stu.xaut.edu.cn

Abstract

Knowledge distillation (KD) aims to improve the performance of lightweight student networks under the guidance of pre-trained teachers. However, the large capacity gap between teachers and students limits the distillation gains. Previous methods addressing this problem have two weaknesses. First, most of them decrease the performance of pre-trained teachers, hindering students from achieving comparable performance. Second, these methods fail to dynamically adjust the transferred knowledge to be compatible with the representation ability of students, which is less effective in bridging the capacity gap. In this paper, we propose **A**daptive **D**ual **G**uidance **K**nowledge **D**istillation (ADG-KD), which retains the guidance of the pre-trained teacher and uses the teacher's bidirectional optimization route guiding the student to alleviate the capacity gap problem. Specifically, ADG-KD introduces an initialized teacher, which has an identical structure to the pre-trained teacher and is optimized through the bidirectional supervision from both the pre-trained teacher and student. In this way, we construct the teacher's bidirectional optimization route to provide the students with an easy-to-hard and compatible knowledge sequence. ADG-KD trains the students under the proposed dual guidance approaches and automatically determines their importance weights, making the transferred knowledge better compatible with the representation ability of students. Extensive experiments on CIFAR-100, ImageNet, and MS-COCO demonstrate the effectiveness of our method.

Introduction

Deep Neural Networks (DNNs) have made remarkable achievements in numerous computer vision tasks (He et al. 2016; He and Gkioxari 2017; Long, Shelhamer, and Darrell 2015). However, top-performing DNNs usually contain large numbers of parameters, bringing heavy computation costs at inference time. To tackle this challenge, many model compression methods (Lin et al. 2020; Yamamoto 2021; Cai, Zhu, and Han 2018; Hinton, Vinyals, and Dean 2015) have been proposed. Knowledge Distillation (KD) has been proposed to transfer knowledge from a high-capacity teacher network to a low-capacity student network, attracting increased attention. The concept of knowledge distillation was first proposed by (Hinton, Vinyals, and Dean 2015), which transfers knowledge by minimizing the Kullback-Leibler (KL) divergence between the predicted distributions of the pre-trained teacher and student. Since then, many KD methods have been proposed, which train a lightweight student by mimicking the output logits (Zhang et al. 2018; Jin, Wang, and Lin 2023; Sun et al. 2024) or intermediate features (Romero et al. 2014; Zagoruyko and Komodakis 2016) of a pre-trained teacher, achieving superior performance than training from scratch. However, as the capacity gap between teachers and students increases, existing KD methods may be unable to improve results, which is known as the capacity gap problem (Mirzadeh et al. 2020).

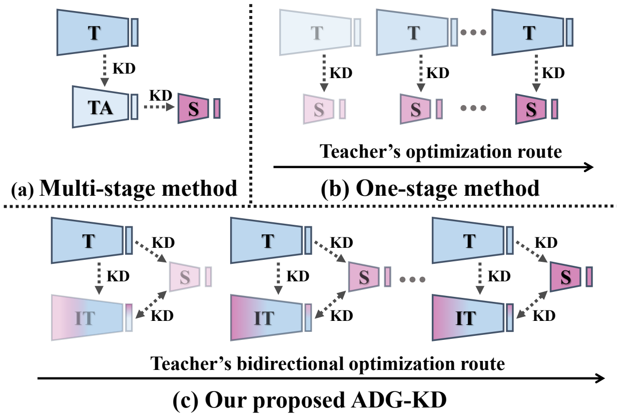

Figure 1: In contrast to multi-stage and one-stage methods, ADG-KD retains the pre-trained teacher and bridges the capacity gap by learning from the BOR.

To alleviate the capacity gap problem, adaptive KD methods have been proposed, such as (Mirzadeh et al. 2020; Son et al. 2021; Xiong et al. 2023; Cho and Hariharan 2019; Jin et al. 2019; Rezagholizadeh et al. 2021). These methods adjust the capacity (Figure 1(a)) or parameter space (Figure 1(b)) of teachers to make the transferred knowledge easier for students to learn. However, this approach inevitably decreases the performance of pre-trained teachers, hindering the student from achieving comparable performance with pre-trained teachers. In addition, these methods fail to dynamically adjust the transferred knowledge to be compati-

*Corresponding author

ble with the varying representation abilities of students at different distillation stages. As a result, they still provide the students with general knowledge, which is less effective in addressing the capacity gap problem.

In this paper, we propose a novel **A**daptive **D**ual **G**uidance **K**nowledge **D**istillation (ADG-KD). As shown in Figure 1(c), when training a student, the knowledge is not distilled only from the pre-trained teacher but also from the teacher's bidirectional optimization route, which is named BOR for simplicity. These two guidance approaches can be adaptively fused concerning a specific training instance. Specifically, ADG-KD introduces an initialized teacher with the same structure as the pre-trained teacher and optimizes it through bidirectional supervision from the pre-trained teacher and student to construct the BOR. Compared with the pre-trained teacher, the BOR provides students with an easy-to-hard and compatible knowledge sequence. By gradually mimicking such sequences, the student can learn from the teacher more effectively, bridging the capacity gap. In ADG-KD, the student receives dual guidance from the pre-trained teacher and BOR. To adaptively fuse these two guidance approaches, we associate the pre-trained teacher and the BOR with latent representations to indicate their characteristics. Based on these latent representations and the instance representation obtained from students, we can automatically determine the importance weights of these two guidance approaches for a specific instance, making the transferred knowledge better compatible with the representation ability of students.

Extensive experiments on CIFAR-100, ImageNet, and MS-COCO demonstrate the effectiveness of the proposed method. Moreover, our method can be integrated with other KD methods, boosting their performance. To sum up, our major contributions are as follows:

- We propose a novel **A**daptive **D**ual **G**uidance **K**nowledge **D**istillation (ADG-KD) that uses the adaptive dual guidance to train the student, bridging the capacity gap and promoting the student achieves comparable performance with the pre-trained teacher.
- ADG-KD introduces an initialized teacher, which is optimized through bidirectional supervision to construct the BOR, and designs a computational method to adaptively fuse the guidance of the pre-trained teacher and BOR, making the transferred knowledge better compatible with the representation ability of students.
- We conduct extensive experiments on CIFAR-100, ImageNet, and MS-COCO to verify the effectiveness of ADG-KD. Additionally, integrating our approach with other KD methods can improve their performance, further illustrating the superiority of ADG-KD.

Related Work

The concept of knowledge distillation was proposed by (Hinton, Vinyals, and Dean 2015), where a lightweight student tries to mimic the predicted distribution of a pre-trained teacher by minimizing the KL divergence. Since then, various KD methods have been proposed to improve distillation performance, which can be categorized into two types, distillation from logits (Yang et al. 2021; Wu and Gong 2021; Zhao et al. 2022; Li et al. 2022; Li and Jin 2022; Huang et al. 2022; Li et al. 2023; Jin, Wang, and Lin 2023; Gong et al. 2023; Sun et al. 2024) and intermediate features (Park et al. 2019; Heo et al. 2019a,b; Tung and Mori 2019; Ahn et al. 2019; Wang et al. 2019; Tian, Krishnan, and Isola 2019; Chen et al. 2021a,b; Song et al. 2022; Chen et al. 2022; Lin et al. 2022; Guo et al. 2023; Yang et al. 2024).

These methods train a lightweight student with the knowledge distilled from a pre-trained teacher. However, as the capacity gap between teachers and students increases, the distillation gains will be limited. To alleviate this problem, adaptive KD methods have been proposed. Most of them are multi-stage approaches. TAKD (Mirzadeh et al. 2020) bridged this gap by training intermediate-sized teacher assistants (TAs). DGKD (Son et al. 2021) used a densely guiding manner to train each TA with higher TAs and the teacher to alleviate error avalanche problems in TAKD. ResKD (Li et al. 2021) used additional residual networks as TAs, bridging the capacity gap. RKD (Gao, Wang, and Wan 2021) introduced a TA to mimic the residual error between the feature maps of the student and teacher, complementing the student with missing information. CES-KD (Amara et al. 2022) used grouping data samples based on their difficulty level and assigned them to the corresponding teacher or TA with appropriate capacity. AAKD (Xiong et al. 2023) introduced a knowledge sample selection strategy and an adaptive teacher strategy to automatically select suitable samples and teachers. Another type is the one-stage approach. ESKD (Cho and Hariharan 2019) proposed an early stopping strategy for the teacher, facilitating a more favorable solution. RCO (Jin et al. 2019) constructed a gradually mimicking sequence by selecting some checkpoints from the training footpath to guide the student. Pro-KD (Rezagholizadeh et al. 2021) provides a smoother training path for the student by following the teacher training routes. Unlike these methods, ADG-KD uses the dual guidance of the pre-trained teacher and BOR to train the student, effectively bridging the capacity gap and promoting the student achieves comparable performance with the pre-trained teacher.

Method

This section provides a detailed introduction of ADG-KD. The overall framework is illustrated in Figure 2.

Revisit of Vanilla KD

We first review the formulation of vanilla KD. For a C-way classification task, we denote network output logits on a training sample (x, y) as $z = [z_1, z_2, \ldots, z_t, \ldots z_c] \in \mathbb{R}^{1\times c}$, then each element in soften classification probability $p = [p_1, p_2, \ldots, p_t, \ldots p_c] \in \mathbb{R}^{1\times c}$ can be calculated by a softmax function:

$$p_j = \frac{e^{z_j/\tau}}{\sum_{i=1}^{c} e^{z_i/\tau}}, \tag{1}$$

where p_j and z_j are the soften probability and output logit on the j class, and τ is a temperature factor to smooth output logit. Following the discussion in DKD (Zhao et al.

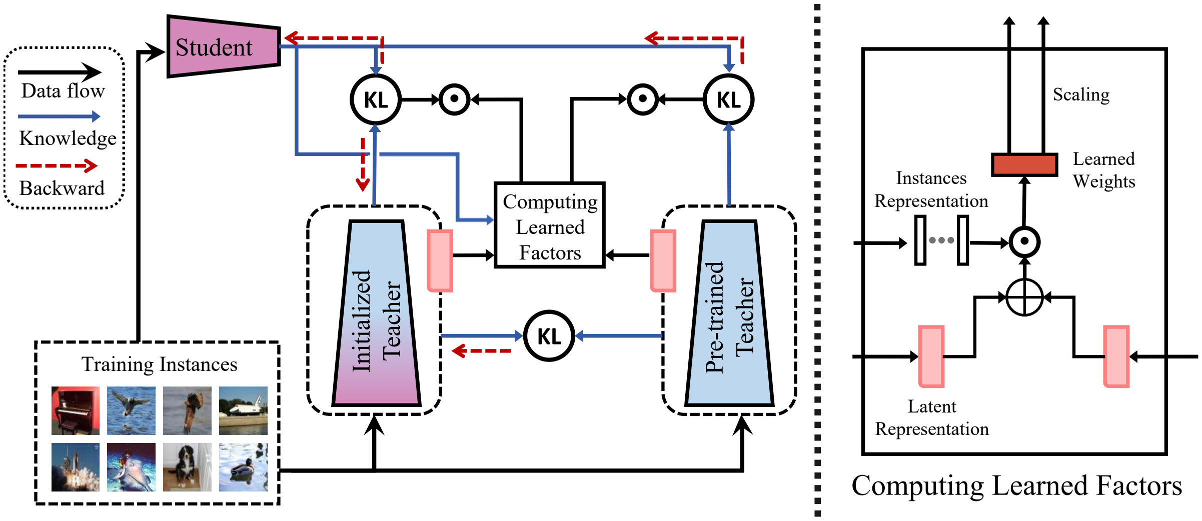

Figure 2: ADG-KD introduces an initialized teacher, which is optimized under the bidirectional supervision of the pre-trained teacher and student to construct the teacher's bidirectional optimization route. During distillation, the student receives dual guidance from the pre-trained teacher and the teacher's bidirectional optimization route. These two guidance approaches are adaptively fused by leveraging their latent representations and the instance representation obtained from the student, making the transferred knowledge better compatible with the representation ability of students.

2022), we divide original predictions p into predictions relevant and irrelevant to the target class. Specifically, we define $b = [p_t, p_{\backslash t}] \in \mathbb{R}^{1\times 2}$ to represent binary probabilities of the target class and all other non-target classes:

$$p_t = \frac{e^{z_t/\tau}}{\sum_{i=1}^{c} e^{z_i/\tau}}, \quad p_{\backslash t} = \frac{\sum_{i=1,i\neq t}^{c} e^{z_i/\tau}}{\sum_{i=1}^{c} e^{z_i/\tau}}. \tag{2}$$

We let $q = [q_1, q_2, \ldots, q_{t-1}, q_{t+1}, \ldots q_c] \in \mathbb{R}^{1\times(c-1)}$ to represent probabilities among non-target classes, each element in q is calculated by:

$$q_j = \frac{e^{z_j/\tau}}{\sum_{i=1,i\neq t}^{c} e^{z_i/\tau}}. \tag{3}$$

The student mimics the predicted distributions of pre-trained teacher via minimizing:

$$\mathcal{L}_{\mathtt{TS}} = KL(p^{\mathtt{T}} \parallel p^{\mathtt{S}}) = p_t^{\mathtt{T}} \log \frac{p_t^{\mathtt{T}}}{p_t^{\mathtt{S}}} + \sum_{i=1,i\neq t}^{c} p_i^{\mathtt{T}} \log \frac{p_i^{\mathtt{T}}}{p_i^{\mathtt{S}}}, \tag{4}$$

where KL is KL divergence, $p^{\mathtt{T}}$ and $p^{\mathtt{S}}$ represent the predicted distributions of the pre-trained teacher and student, respectively. $\mathcal{L}_{\mathtt{TS}}$ can be rewritten as:

$$\mathcal{L}_{\mathtt{TS}} = KL(b^{\mathtt{T}} \parallel b^{\mathtt{S}}) + (1 - p_t^{\mathtt{T}}) KL(q^{\mathtt{T}} \parallel q^{\mathtt{S}}), \tag{5}$$

where $b^{\mathtt{T}}$, $b^{\mathtt{S}}$ denote the predicted distributions for the target class and $q^{\mathtt{T}}$ and $q^{\mathtt{S}}$ represent the predicted distributions for the non-target classes, by the pre-trained teacher and student, respectively.

The pre-trained teacher has high confidence in the target class and limited confidence in the non-target ones. In contrast, due to fewer parameters and simpler architecture, the student lacks confidence in the target class and spreads its confidence across non-target classes in most distillation stages. This disparity results in high values of $KL(b^{\mathtt{T}} \parallel b^{\mathtt{S}})$ and $KL(q^{\mathtt{T}} \parallel q^{\mathtt{S}})$, as well as a suppression for $1 - p_t^{\mathtt{T}}$, resulting in the capacity gap problem. Therefore, despite the high performance of the pre-trained teacher, their knowledge is not conducive to the student's learning.

Formulation of ADG-KD

Teacher's Bidirectional Optimization Route. ADG-KD introduces an initialized teacher and optimizes it through bidirectional supervision of the pre-trained teacher and student to construct the teacher's bidirectional optimization route. During distillation, the student possesses limited representation ability in the early stages, gradually developing more appropriate representation ability in the later stages. Consequently, the predicted distributions of the initialized teacher should be closer to that of the student in the early distillation stages, while in the later stages, they should be closer to that of the pre-trained teacher. We develop a conditional triplet loss to control the distance among them:

$$\mathcal{L}_{tri} = \begin{cases} Max(\mathcal{D}(p^{\mathtt{I}}, p^{\mathtt{S}}) - \mathcal{D}(p^{\mathtt{I}}, p^{\mathtt{T}}) + a, 0) & \mathtt{E} < \eta \\ Max(\mathcal{D}(p^{\mathtt{I}}, p^{\mathtt{T}}) - \mathcal{D}(p^{\mathtt{I}}, p^{\mathtt{S}}) + a, 0) & \mathtt{E} \geq \eta \end{cases}, \tag{6}$$

where $p^{\mathtt{I}}$ is the predicted distributions of the initialized teacher, $\mathcal{D}$ is the distance function, defined as KL divergence in our method. a is a margin value to ensure the non-negativity of $\mathcal{L}_{tri}$ and $\mathtt{E}$ denotes the training epoch. In the early distillation stages, it is desirable for $p^{\mathtt{I}}$ to be close to $p^{\mathtt{S}}$ to ensure the transferred knowledge is easy and compatible with the student. However, in the later distillation stages,

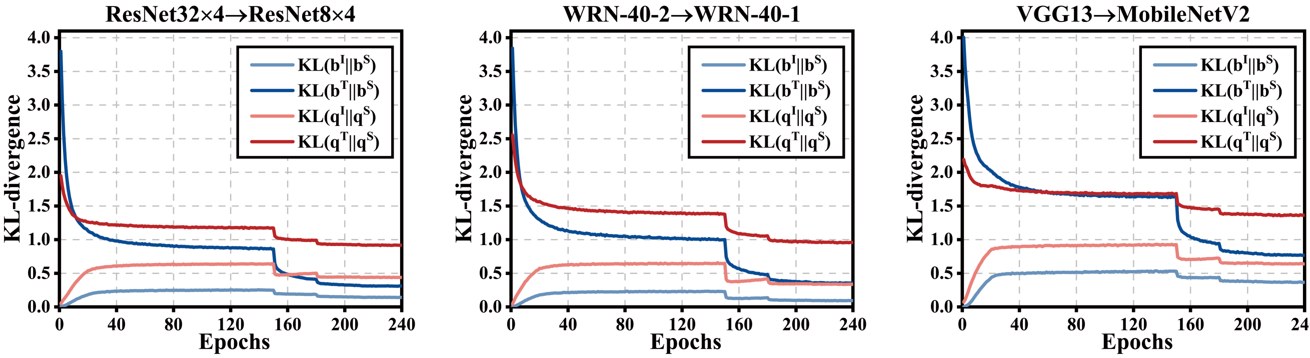

Figure 3: Comparison of KL-divergence among the outputs of the student, BOR, and pre-trained teacher. We take teacher-student pairs are ResNet32 ×4 → ResNet8×4 (left), WRN-40-2 → WRN-40-1 (middle), and VGG13 → MobileNetV2 (right).

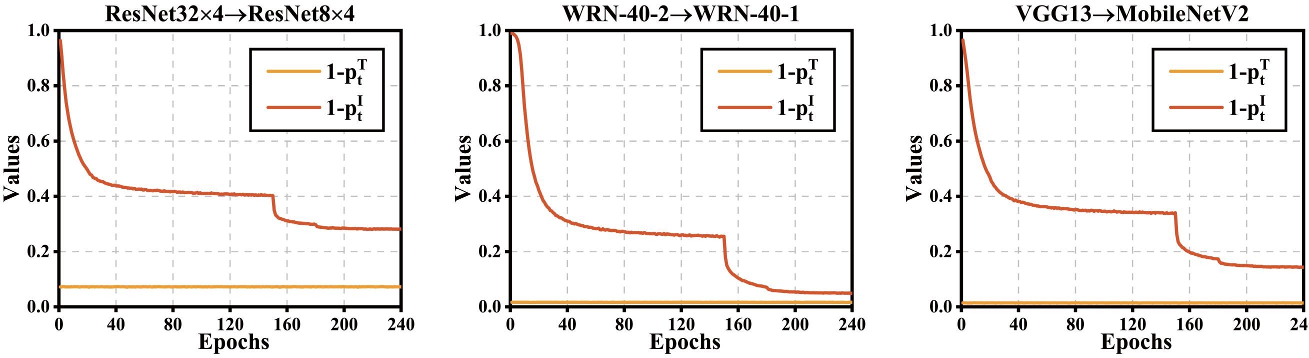

Figure 4: Comparison of $1 - p_t^T$ and $1 - p_t^I$, we take teacher-student pairs are ResNet32 ×4 → ResNet8×4 (left), WRN-40-2 → WRN-40-1 (middle), and VGG13 → MobileNetV2 (right).

the distance between p^{I} and p^{S} should be expanded while the distance between p^{I} and p^{T} should be reduced, better compatible with the increasing representation ability of the student. We use a hyper-parameter η to control the shift of these two objective functions.

Compared with the pre-trained teacher, the knowledge of BOR is more conducive to the student's learning. As shown in Figure 3 and Figure 4, we have:

$$\begin{cases} KL(b^{\mathrm{I}} \parallel b^{\mathrm{S}}) < KL(b^{\mathrm{T}} \parallel b^{\mathrm{S}}), \\ KL(q^{\mathrm{I}} \parallel q^{\mathrm{S}}) < KL(q^{\mathrm{T}} \parallel q^{\mathrm{S}}), \\ 1 - p_t^{\mathrm{T}} < 1 - p_t^{\mathrm{I}}, \end{cases} \tag{7}$$

where p_t^{T} and p_t^{I} are the predicted distributions of the pre-trained teacher and BOR for the target class, respectively. It can be seen from Eq. (7) that learning from the BOR is more conducive for the student than a pre-trained teacher.

Adaptive Dual Guidance Approaches. During distillation, different training instances present varying degrees of learning difficulty for the student. For complicated or error-prone instances, the guidance of BOR can ease the learning difficulty, bridging the capacity gap. Conversely, for more accessible samples, the guidance of the pre-trained teacher can further refine the student's learning. Therefore, we use two automatically learned factors ξ_t and ξ_i to fuse these two guidance approaches adaptively. Inspired by latent factor models in the recommender system (Koren 2008), we introduce two latent representations to indicate the characteristics of the pre-trained teacher and BOR. Concretely, the pre-trained teacher and BOR are associated with factors $\theta_t \in \mathbb{R}^d$ and $\theta_i \in \mathbb{R}^d$ where d is the dimension of the factor. We take the student's output logit as the representation of instances. As such, we get $Z \in R^{d\times c}$ for a batch of training data where d and c correspond to the number of instances and categories of the student's output logit, respectively. Then, we calculate the importance weights of the pre-trained teacher and BOR:

$$\begin{aligned} \phi_t &= \omega^T(\theta_t \odot Z), \\ \phi_i &= \omega^T(\theta_i \odot Z), \end{aligned} \tag{8}$$

where ω is a learned weight parameter that determines whether or not each logit has a positive effect on the score. $\odot$ denotes the element-wise product, which can capture the interaction between the representations of corresponding terms and Z. ϕ_t and ϕ_i are the importance weights of the pre-trained teacher and BOR, respectively.

To keep the value of the importance weights within a proper range and ensure its non-negativity, we scale the ϕ_t and ϕ_i with the following equation:

$$\begin{aligned} \xi_t &= \xi_{init} + \xi_{range}(\delta(\phi_t)), \\ \xi_i &= \xi_{init} + \xi_{range}(\delta(\phi_i)), \end{aligned} \tag{9}$$

Item		Homogeneous architecture				Heterogeneous architecture					
Method	Teacher	R32×4	R56	W40-2	W40-2	VGG13	R32×4	W40-2	VGG13	R50	R32×4
	Acc	79.42	72.34	75.61	75.61	74.64	79.42	75.61	74.64	79.34	79.42
	Student	R8×4	R20	W40-1	W16-2	VGG8	SV1	SV1	MV2	MV2	SV2
	Acc	72.50	69.06	71.98	73.26	70.36	70.50	70.50	64.60	64.60	71.82
Feature	FitNet	73.50	69.21	72.24	73.58	71.02	73.59	73.73	64.14	63.16	73.54
	CRD	75.51	71.16	74.14	75.48	73.94	75.11	76.05	69.73	69.11	75.65
	WCoRD	75.95	71.56	74.73	75.88	74.55	75.40	76.32	69.47	70.45	75.96
	KR	75.63	71.89	75.09	76.12	74.84	77.45	77.14	70.37	69.89	77.78
	CAT-KD	76.91	71.62	74.82	75.60	74.65	**78.26**	77.35	69.13	**71.36**	78.41
Logit	KD	73.33	70.66	73.54	74.92	72.98	74.07	74.83	67.37	67.35	74.45
	DKD	76.32	71.97	74.81	76.24	74.68	76.45	76.70	69.71	70.35	77.07
	CTKD	N/A	71.19	73.93	75.45	73.52	74.48	75.78	68.46	68.47	75.31
	MKD	77.08	**72.19**	75.35	76.63	75.18	77.18	77.44	**70.57**	71.04	78.44
	LSKD	76.62	71.43	74.37	76.11	74.36	N/A	N/A	68.61	69.02	75.56
Ours	ADG-KD	**77.44**	**72.46**	**75.84**	**76.98**	**75.49**	**77.82**	**77.65**	70.35	70.63	**78.72**
	KR†	**77.33**	72.07	75.56	76.74	75.03	77.65	77.39	**70.79**	70.48	78.34
	DKD†	76.78	72.35	**75.98**	**76.87**	**75.41**	77.34	76.93	69.85	70.54	77.36
	MKD†	**77.51**	**72.67**	**76.61**	**77.09**	**75.73**	77.43	**77.95**	**71.63**	**71.67**	**78.83**
	CAT-KD†	77.23	72.05	75.62	76.24	75.18	**78.65**	**77.89**	69.68	**71.72**	**78.96**

Table 1: Comparison of Top-1 accuracy (%) with powerful distillation methods on CIFAR-100. R32×4, R8×4, R56, R50, R20, W40-2, W40-1, W16-2, MV2, SV1 and SV2 stand for ResNet32×4, ResNet8×4, ResNet56, ResNet50, ResNet20, WRN-40-2, WRN-40-1, WRN-16-2, MobileNetV2, ShuffleNetV1 and ShuffleNetV2. All results are the average of five trials. We use red, blue, and green to indicate the results of the top three methods.

where ξ_{init} represents the initial value, ξ_{range} represents the range for ξ_t and ξ_i, δ is the sigmoid function. We default ξ_{init} and ξ_{range} as 1 and 3, ensuring all reasonable values can be included.

We use these two learned factors ξ_t and ξ_i to adaptively fuse these two guidance approaches for a specific instance, which is achieved by the element-wise product operation. The overall loss for the student is presented as:

$$\begin{cases} \mathcal{L}_{\text{S}} = \lambda CE(p^{\text{S}}, y) + (1-\lambda)(\mathcal{L}_{\text{TS}} + \mathcal{L}_{\text{IS}}), \\ \mathcal{L}_{\text{TS}} = \xi_t \odot KL(p^{\text{T}} \parallel p^{\text{S}}), \\ \mathcal{L}_{\text{IS}} = \xi_i \odot KL(p^{\text{I}} \parallel p^{\text{S}}). \end{cases} \tag{10}$$

By fusing these two guidance approaches, the transferred knowledge can be better compatible with the representation ability of students, boosting distillation performance.

Experiments

In this section, we evaluate our method on image classification and object detection tasks, including:

CIFAR-100 (Krizhevsky, Hinton et al. 2009) is a medium-scale image classification dataset consisting of 60,000 images (50,000 training samples and 10,000 testing samples) from 100 categories and its resolution is 32 × 32 pixels.

ImageNet (Deng et al. 2009) is one of the most important benchmark datasets for image classification, with a total of 1.28 million training samples and 50,000 testing samples from 1000 categories. The resolution of input samples is fixed to 224 × 224.

MS-COCO (Lin et al. 2014) is a fundamental object detection dataset, with 118k images to train and 5k images to test from 80 categories.

Main Results

CIFAR-100 classification. Table 1 evaluates our method on CIFAR-100. For the homogeneous teacher-student pairs, ADG-KD achieves 3.40% - 5.13% absolute gains than baseline and outperforms vanilla KD with 1.80% - 4.11% margins. Besides, ADG-KD achieves more significant gains on heterogeneous teacher-student pairs with 5.55% - 7.12% margins than baseline and surpasses vanilla KD with 2.62% - 4.07% margins. Compared to state-of-the-art (SOTA) distillation methods, ADG-KD achieves comparable or even better performance. We also integrate ADG-KD with other KD methods. As shown in Table 1, our method brings comprehensive improvements. For SOTA methods MKD and CAT-KD, our approach brings 0.25% - 1.06% and 0.32% - 0.80% accuracy gains, respectively. These experimental results verify the effectiveness of our approach.

ImageNet classification. Table 2 shows the performance of our method on ImageNet. The proposed method consistently improves Top-1 and Top-5 accuracies over vanilla KD. Specifically, ADG-KD obtains 1.56% Top-1 and 0.94% Top-5 absolute gains over vanilla KD within the same network structure. In addition, it brings a 4.85% improvement

Network			Base Training		Feature				Logit				**Ours**	
Teacher	Student		Teacher	Student	AT	SRRL	KR	CAT-KD	KD	DKD	MKD	LSKD	ADG-KD	DKD†
R34	R18	Top-1	73.31	69.75	70.69	71.73	71.61	71.26	70.66	71.70	**71.90**	71.42	**72.22**	**71.98**
		Top-5	91.42	89.07	90.01	**90.60**	90.51	90.45	89.88	90.41	90.55	90.29	**90.82**	**90.56**
R50	MV2	Top-1	76.16	68.87	69.56	72.49	72.56	72.24	68.58	72.05	**73.01**	72.18	**73.43**	**72.87**
		Top-5	92.86	88.76	89.33	90.92	91.00	91.13	88.98	91.05	**91.42**	90.80	**91.49**	**91.36**

Table 2: Comparison of Top-1 and Top-5 accuracies (%) with powerful distillation methods on ImageNet. R34, R18, R50 and MV2 stand for ResNet34, ResNet18, ResNet50 and MobileNetV2. All results are the average of three trials. We use red, blue, and green to indicate the performance of the top three methods.

		AP	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$	AP	AP$_{50}$	AP$_{75}$
Method			R101			R101			R50	
	Teacher	42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
			R18			R50			MV2	
	Student	33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
Feature	FitNet	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
	FGFI	35.44	55.51	38.17	39.44	60.27	43.04	31.16	50.68	32.92
	KR	**36.75**	56.72	34.00	**40.36**	60.97	44.08	33.71	53.15	**36.13**
Logit	KD	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
	TAKD	34.59	55.35	37.12	39.01	60.32	43.10	31.26	51.03	33.46
	DKD	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
	MKD	36.03	57.28	**38.51**	40.15	**61.67**	44.57	33.83	54.01	35.22
Ours	ADG-KD	**36.34**	**57.43**	**38.65**	**40.45**	**61.84**	**44.76**	**34.01**	**54.14**	35.42
	KR†	**37.22**	**57.62**	**38.94**	**40.63**	**61.73**	**44.62**	**34.38**	**54.26**	**36.45**
	DKD†	36.21	**57.35**	38.47	40.32	61.59	**44.59**	**34.07**	**54.21**	**35.53**

Table 3: Experimental results on MS-COCO. We use open-source report Detectron2 (Wu et al. 2019) as our baseline, Faster-RCNN (Ren et al. 2015)-FPN (Lin et al. 2017) as backbone, and AP, AP$_{50}$, and AP$_{75}$ as evaluation metrics. R101, R50, R18 and MV2 stand for ResNet101, ResNet50, ResNet18 and MobileNetV2. All results are the average of three trials. We used red, blue, and green to indicate the performance of the top three methods.

in Top-1 accuracy and 2.51% improvement in Top-5 accuracy compared to vanilla KD under different network structures. It is worth mentioning that the performance of ADG-KD is better than that of the MKD and CAT-KD. We also integrate our method with DKD, bringing significant improvements. These experimental results verify the superiority of ADG-KD on the large-scale dataset.

MS-COCO object detection. Apart from image classification, we also apply our method to the object detection. Table 3 shows the experimental results. ADG-KD achieves consistent improvement over vanilla KD, verifying the effectiveness of our method. For example, in the distillation experiment ResNet101→ResNet50, ADG-KD surpasses vanilla KD with 2.10, 2.43, and 3.05 points on AP, AP$_{50}$, and AP$_{75}$ metrics, respectively. By combining our approach with KR and DKD, we boost their performance to a higher level. These results validate that our method is still effective in the object detection task.

Ablation Study

The dual guidance approach. To investigate the effect of the guidance of the pre-trained teacher and BOR, we design a variant of ADG-KD, ADG-KD-NT. ADG-KD-NT removes the guidance of the pre-trained teacher. Therefore, the student exclusively learns knowledge from the BOR. The experiments are conducted on CIFAR-100, and corresponding results are presented in Table 4.

Compared to the pre-trained teacher, the BOR provides students with an easy-to-hard and compatible knowledge sequence, effectively addressing the capacity gap problem. Comparing ADG-KD-NT with KD, DML, and TAKD, we find that ADG-KD-NT still performs better, validating its effectiveness in bridging the capacity gap. Moreover, ADG-KD performs better than ADG-KD-NT, which is consistent with our initial viewpoint that the proposed dual guidance can effectively bridge the capacity gap and promote the student to achieve more comparable performance with the pre-trained teacher.

The learned factors ξ_t and ξ_i. In ADG-KD, we use two learned factors ξ_t and ξ_i to adaptively fuse the guidance of the pre-trained teacher and BOR. To illustrate the benefit of this approach, we take an experiment on CIFAR-100. As shown in Figure 5, compared to base setting ($\xi_t = \xi_i = 1$), using learned factors ξ_t and ξ_i can achieve better validating accuracy (↑ 0.29%), improving training efficiency and attaining satisfactory distillation performance.

Teacher	R32×4	R56	W40-2	W40-2	VGG13	R32×4	W40-2	VGG13	R50	R32×4
Acc	79.42	72.34	75.61	75.61	74.64	79.42	75.61	74.64	79.34	79.42
Student	R8×4	R20	W40-1	W16-2	VGG8	SV1	SV1	MV2	MV2	SV2
Acc	72.50	69.06	71.98	73.26	70.36	70.50	70.50	64.60	64.60	71.82
KD	73.33	70.66	73.54	74.92	72.98	74.07	74.83	67.37	67.35	74.45
DML	72.12	69.52	72.68	73.58	71.79	72.89	72.76	65.63	65.71	73.45
TAKD	74.91	71.37	73.99	75.62	74.12	74.53	75.34	67.91	68.02	74.82
ADG-KD-NT	76.45	71.81	75.49	76.58	74.71	76.21	76.65	69.83	70.13	77.18
ADG-KD	77.44	72.46	75.84	76.98	75.49	77.82	77.65	70.35	70.63	78.72

Table 4: Top-1 accuracy(%) of ADG-KD-NT on CIFAR-100. All results are the average of five trials.

Teacher	R32×4	R56	W40-2	W40-2	VGG13	R32×4	W40-2	VGG13	R50	R32×4	Avg
Acc	79.42	72.34	75.61	75.61	74.64	79.42	75.61	74.64	79.34	79.42	
Student	R8×4	R20	W40-1	W16-2	VGG8	SV1	SV1	MV2	MV2	SV2	
Acc	72.50	69.06	71.98	73.26	70.36	70.50	70.50	64.60	64.60	71.82	
Gap-Base	6.92	3.28	3.63	2.35	4.28	8.92	5.11	10.04	14.74	7.60	6.68
Gap-KD	6.09	1.68	2.07	0.69	1.66	5.35	0.78	7.27	11.99	4.97	4.25
Gap-ADG-KD	1.98	-0.12	-0.23	-1.37	-0.85	1.60	-2.04	4.29	8.71	0.70	1.26

Table 5: We conducted experiments on CIFAR-100 and used Top-1 accuracy as the evaluation metric to show the performance gap between teachers and students. Note that when the student outperforms the teacher, the gap is negative.

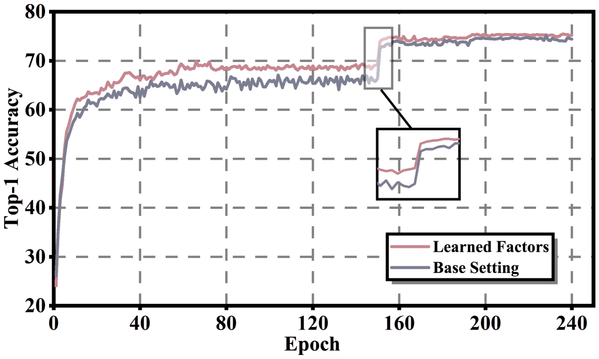

Figure 5: Comparison of testing curves of base setting (75.24%) and using learned factors ξ_t and ξ_i (75.53%) for VGG8 with VGG13 as teacher on CIFAR-100.

The capacity gap problem. It is interesting to show the performance gap between teachers and students. In Table 5, we calculate the gap between the performance of the teacher and student. When the student outperforms the teacher, the corresponding gap is negative. It can be seen that with ADG-KD, the student performance is highly close to that of the pre-trained teacher. Another surprising phenomenon is that sometimes, the student performs even better than the pre-trained teacher. We conjecture that the cause of this may be that the proposed adaptive dual guidance works well in the distillation process.

To further demonstrate the strength of our method in bridging the capacity gap, we compare it against some adaptive KD methods and SRRL. We conducted experiments on CIFAR-100 and present the results in Table 6. The experimental results indicate the superiority of ADG-KD.

Teacher	R56	VGG13	W40-2	R32×4	R32×4
Student	R20	VGG8	W16-2	SV1	SV2
TA	R32	VGG11	W22-2	R14×4	R14×4
RCO	71.52	74.67	75.28	75.31	76.12
TAKD	71.37	74.12	75.62	74.93	75.88
DGKD	71.49	74.31	76.10	76.13	76.41
SHAKE	72.04	74.84	76.62	77.38	78.25
AAKD	71.55	74.17	75.66	75.20	75.98
SRRL	71.40	74.40	75.96	75.66	76.40
ADG-KD	**72.46**	**75.49**	**76.98**	**77.82**	**78.72**

Table 6: We conduct experiments on CIFAR-100 to compare ADG-KD with some adaptive KD methods and SRRL. The best results are indicated in boldface.

Conclusion

In this paper, we propose **A**daptive **D**ual **G**uidance **K**nowledge **D**istillation (ADG-KD), which retains the guidance of the pre-trained teacher and uses the teacher's bidirectional optimization route to guide the student to alleviate the capacity gap problem. To construct the teacher's bidirectional optimization route, we introduce an initialized teacher and optimize it under the bidirectional supervision of the pre-trained teacher and student. During distillation, the student receives the dual guidance of the pre-trained teacher and the teacher's bidirectional optimization route. These two guidance approaches are adaptively fused, making the transferred knowledge better compatible with the representation ability of students. Extensive experiments on image classification and object detection demonstrate the effectiveness of our method. In possible future work, we plan to extend ADG-KD to other distillation approaches, such as self-distillation, and apply it to additional computer vision tasks.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant No.61673318, Xi'an Science and Technology Project under Grant NO. 22GXFW0096, and Key Industry Chain Project of Shaanxi Province under Grant NO. 2020ZDLGY04-04.

References

Ahn, S.; Hu, S. X.; Damianou, A.; Lawrence, N. D.; and Dai, Z. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9163–9171.

Amara, I.; Ziaeefard, M.; Meyer, B. H.; Gross, W.; and Clark, J. J. 2022. CES-KD: curriculum-based expert selection for guided knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1901–1907. IEEE.

Cai, H.; Zhu, L.; and Han, S. 2018. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*.

Chen, D.; Mei, J.-P.; Zhang, H.; Wang, C.; Feng, Y.; and Chen, C. 2022. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11933–11942.

Chen, L.; Wang, D.; Gan, Z.; Liu, J.; Henao, R.; and Carin, L. 2021a. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16296–16305.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021b. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5008–5017.

Cho, J. H.; and Hariharan, B. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4794–4802.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gao, M.; Wang, Y.; and Wan, L. 2021. Residual error based knowledge distillation. *Neurocomputing*, 433: 154–161.

Gong, L.; Lin, S.; Zhang, B.; Shen, Y.; Li, K.; Qiao, R.; Ren, B.; Li, M.; Yu, Z.; and Ma, L. 2023. Adaptive hierarchy-branch fusion for online knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7731–7739.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class Attention Transfer Based Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11868–11877.

He, K.; and Gkioxari, G. 2017. Piotr Dollá r, and Ross B. *Girshick,"Mask r-cnn," CoRR, vol. abs/1703.06870.*

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019a. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1921–1930.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019b. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3779–3787.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Jin, X.; Peng, B.; Wu, Y.; Liu, Y.; Liu, J.; Liang, D.; Yan, J.; and Hu, X. 2019. Knowledge distillation via route constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1345–1354.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-Level Logit Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24276–24285.

Koren, Y. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 426–434.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, C.; Lin, M.; Ding, Z.; Lin, N.; Zhuang, Y.; Huang, Y.; Ding, X.; and Cao, L. 2022. Knowledge condensation distillation. In *European Conference on Computer Vision*, 19–35. Springer.

Li, L.; and Jin, Z. 2022. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems*, 35: 635–649.

Li, X.; Li, S.; Omar, B.; Wu, F.; and Li, X. 2021. Reskd: Residual-guided knowledge distillation. *IEEE Transactions on Image Processing*, 30: 4735–4746.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. Hrank: Filter pruning using high-rank feature map. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1529–1538.

Lin, S.; Xie, H.; Wang, B.; Yu, K.; Chang, X.; Liang, X.; and Wang, G. 2022. Knowledge distillation via the target-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10915–10924.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Rezagholizadeh, M.; Jafari, A.; Salad, P.; Sharma, P.; Pasand, A. S.; and Ghodsi, A. 2021. Pro-kd: Progressive distillation by following the footsteps of the teacher. *arXiv preprint arXiv:2110.08532*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Son, W.; Na, J.; Choi, J.; and Hwang, W. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9395–9404.

Song, J.; Chen, Y.; Ye, J.; and Song, M. 2022. Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, 31: 3359–3370.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1365–1374.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4933–4942.

Wu, G.; and Gong, S. 2021. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, 10302–10310.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2.

Xiong, Y.; Zhai, W.; Xu, X.; Wang, J.; Zhu, Z.; Ji, C.; and Cao, J. 2023. Ability-aware knowledge distillation for resource-constrained embedded devices. *Journal of Systems Architecture*, 102912.

Yamamoto, K. 2021. Learnable companding quantization for accurate low-bit neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5029–5038.

Yang, J.; Martinez, B.; Bulat, A.; Tzimiropoulos, G.; et al. 2021. Knowledge distillation via softmax regression representation learning. International Conference on Learning Representations (ICLR).

Yang, S.; Yang, J.; Zhou, M.; Huang, Z.; Zheng, W.-S.; Yang, X.; and Ren, J. 2024. Learning from Human Educational Wisdom: A Student-Centered Knowledge Distillation Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.