Can Students Beyond The Teacher? Distilling Knowledge from Teacher's Bias

Jianhua Zhang[1], **Yi Gao**[1], **Ruyu Liu**[2, 3*], **Xu Cheng**[1, 3*], **Houxiang Zhang**,[4] **Shengyong Chen**[1]

[1]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China
[2]School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China
[3]the Department of Technology, Management and Economics, Technical University of Denmark, Lyngby, Denmark
[4]Norwegian University of Science and Technology
zjh@ieee.org, gaoyi01020304@stud.tjut.edu.cn, lry@hznu.edu.cn, xu.cheng@ieee.org, hozh@ntnu.no, sy@ieee.org

Abstract

Knowledge distillation (KD) is a model compression technique that transfers knowledge from a large teacher model to a smaller student model to enhance its performance. Existing methods often assume that the student model is inherently inferior to the teacher model. However, we identify that the fundamental issue affecting student performance is the bias transferred by the teacher. Current KD frameworks transmit both right and wrong knowledge, introducing bias that misleads the student model. To address this issue, we propose a novel strategy to rectify bias and greatly improve the student model's performance. Our strategy involves three steps: First, we differentiate knowledge and design a bias elimination method to filter out biases, retaining only the right knowledge for the student model to learn. Next, we propose a bias rectification method to rectify the teacher model's wrong predictions, fundamentally addressing bias interference. The student model learns from both the right knowledge and the rectified biases, greatly improving its prediction accuracy. Additionally, we introduce a dynamic learning approach with a loss function that updates weights dynamically, allowing the student model to quickly learn right knowledge-based easy tasks initially and tackle hard tasks corresponding to biases later, greatly enhancing the student model's learning efficiency. To the best of our knowledge, this is the first strategy enabling the student model to surpass the teacher model. Experiments demonstrate that our strategy, as a plug-and-play module, is versatile across various mainstream KD frameworks.

Code — https://github.com/smartyige/BTKD

Introduction

Knowledge Distillation (KD) was proposed by (Hinton, Vinyals, and Dean 2015). This method uses a pre-trained model as a teacher, from which useful knowledge is extracted and transferred to train a smaller student model. This process enables the student model to achieve performance close to that of the teacher model. Consequently, KD effectively compresses models while maintaining their performance, making it widely applicable in scenarios that require both efficiency and accuracy, such as on mobile and embedded devices, real-time image processing, and video analysis. Despite efforts to improve student model performance through various KD methods (Park et al. 2019; Tian, Krishnan, and Isola 2019; Chen et al. 2021; Sarridis et al. 2022; Patel, Mopuri, and Qiu 2023), including those based on logit layers (Müller, Kornblith, and Hinton 2019; Zhao et al. 2022) and intermediate feature layers (Park and No 2022; Sarridis et al. 2022; Liang, Huang, and Liu 2024; Tang et al. 2023; Zhang et al. 2024), these approaches have consistently failed to surpass the teacher model. The fundamental reason is that these methods overlook the fact that the knowledge transmitted by the teacher is not entirely accurate. Bias from the teacher can misguide the student, thereby limiting the student model's performance. Moreover, as data types become more diverse and tasks become more complex, the impact of bias knowledge will further intensify.

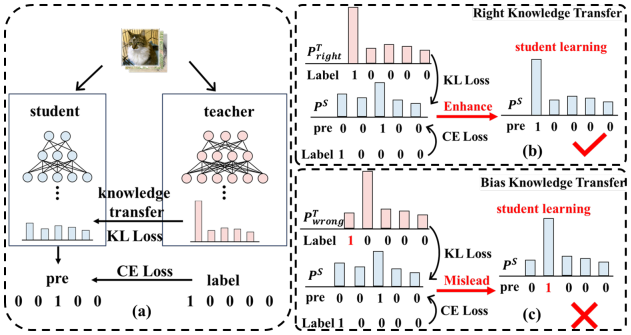

Figure 1: Diagram of the knowledge transfer process in KD. (a) The KD framework based on a teacher-student network model. (b) Right knowledge transfer helps the student model learn effectively. (c) Bias knowledge transfer can mislead the student model with biased knowledge.

We acknowledge the bias present in the KD framework. As shown in Fig. 1, the student model learns from the true labels and the teacher model's predictions through cross-entropy (CE) loss and Kullback-Leibler divergence (KL) loss, respectively. The student's learning is enhanced when the teacher's predictions (P^T_{right}) align with the true labels. Conversely, when the teacher's predictions (P^T_{wrong}) do not match the true labels, they mislead the student, causing its

*Corresponding authors.

learning to deviate and resulting in wrong predictions. Based on these two observations, we categorize the knowledge transferred by the teacher into right knowledge and wrong knowledge, with the latter further defined as bias.

To mitigate the impact of teacher bias on student learning performance, we propose a novel bias rectification strategy. First, based on the categorization of right knowledge and bias, we design a bias elimination method that separates bias from the transferred knowledge, retaining only the right knowledge. This directly eliminates the impact of bias on the student model, thereby greatly enhancing its performance under the guidance of the right knowledge. Second, the student model cannot learn knowledge from the data resulting in teachers' bias, if we directly remove those data when training student model, we further design a bias rectification method that utilizes weighted adjustment to convert bias into the right knowledge. This fundamentally addresses the impact of bias, allowing the student model to learn from both the right knowledge and the rectified bias, thereby greatly improving prediction accuracy. Third, we observe that the learning times for the student model for easy tasks based on right knowledge and for hard tasks corresponding to biases is much different. Therefore, we propose a dynamic learning approach with an improved loss function to optimize the model. Initially, the student focuses on learning easy tasks, and as learning capability increases, more harder tasks are gradually introduced in the later stages of training. This approach significantly reduces the KD time and greatly enhances the student model's prediction efficiency. Finally, we validate the effectiveness of our bias rectification strategy for KD frameworks on complex tasks like image classification and object detection. Extensive experiments demonstrate that our method outperforms state-of-the-art (SOTA) KD methods on benchmark datasets and can be used as a plug-and-play module to enhance existing KD frameworks. Moreover, it is the first time that the student model can surpass the teacher model in KD framework through the proposed strategy.

The main contributions of our paper are as follows:

- Through in-depth theoretical analysis, we have demonstrated the presence of bias in KD and its detrimental impact on student model performance.
- We propose a novel bias rectification strategy through which the student model surpasses the teacher model for the first time. In this strategy, we not only eliminate errors to strengthen the transmission of correct knowledge but also rigorously rectify biases to mitigate the misleading effects of incorrect knowledge.
- We propose a dynamic learning approach that allows the student model to quickly master easy tasks based on right knowledge in the early stages of training, while addressing hard tasks related to biases in the later stages. This approach significantly improves learning efficiency.
- We validated the effectiveness of our strategy by showcasing the student model's superior performance on two types of tasks across three benchmark datasets. Additionally, as a plug-and-play model, the strategy is versatile and can enhance existing KD frameworks.

Related Work

KD methods (Gou et al. 2021; Wang and Yoon 2021; Chen et al. 2022) can be categorized into two categories: logits-based and features-based. The logits-based method primarily involves having the student learn from the teacher's soft labels. Initially proposed by Hinton et al.(Hinton, Vinyals, and Dean 2015), this method is known for its simplicity and ease of implementation, achieving significant results in the early stages of neural network training. Subsequent work by (Kim and Kim 2017) introduced the use of class-distance loss to enhance knowledge transfer to the student model. Zhao et al. (Zhao et al. 2022) improved the effectiveness of logits-based KD by decoupling probabilities in the logits. More recently, CTKD method (Li et al. 2023) has further enhanced student model performance by dynamically adjusting the temperature hyperparameter in KD. However, existing methods generally assume that all transferred knowledge is right, overlooking the bias introduced when the probability distribution predicted by the teacher model does not match the true labels. These biases are directly transferred to the student model through the logit layer via KL loss, which misleads the student learning the biased knowledge. This results in a performance ceiling that the student model cannot surpass.

More researchers (Paulin and Suneson 2012; Huang and Wang 2017; Zhang et al. 2021) have realized that the intermediate layer features of the teacher network contain valuable information. Therefore, features-based KD (Passalis and Tefas 2018; Zhu et al. 2021; Park et al. 2021) has been widely explored to improve student performance by transferring knowledge from the teacher model's intermediate features. FitNet (Romero et al. 2014) is one of the first KD methods based on intermediate layers. It uses a small regression to align the intermediate layer features of the teacher and student model. Subsequent works (Park et al. 2019; Tian, Krishnan, and Isola 2019; Chen et al. 2021; Sarridis et al. 2022) have delved into in-depth research on effectively transmitting the teacher's intermediate layer features to the student. Current methods (Xu, Liu, and Loy 2023; Cho et al. 2023; Zhu et al. 2024; Xie et al. 2024) primarily focus on effectively extracting and transferring knowledge from the intermediate feature layers of the teacher model without considering the rightness of the transferred knowledge. Intermediate features often contain rich semantic information, but biased knowledge is hidden among numerous neurons, making it difficult to detect. This biased knowledge subtly affects the accuracy of semantic information extraction, leading to deviations in final predictions. Therefore, transferring such biased knowledge to the student model can also mislead its final predictions.

While the previous method (Zhou et al. 2020) have attempted to ensure the rightness of feature transfer by simply removing the intermediate features of wrongly predicted samples, this approach of directly associating wrong predictions with intermediate features is unreasonable. Intermediate features of right samples may also contain biased knowledge, Conversely, wrong samples may contain a lot of right feature information, and removing all of them affects the completeness of feature extraction. Although adversarial

defense methods (Li et al. 2024) can effectively mitigate the impact of biased intermediate features, these methods still cannot effectively eliminate the bias in the teacher model.

Methodology

Definitions

We first define and introduce several key concepts and terms that will be used throughout this section. **Right and biased knowledge**: right knowledge comes from cases where the teacher's predicted probability distribution aligns with the true labels. Biased knowledge, or wrong knowledge, arises when the teacher's predicted probability distribution does not match the true labels. **Easy and Hard Tasks**: For the student, tasks where the teacher can make right predictions are considered easy tasks, and the knowledge transferred by the teacher is right knowledge. Conversely, tasks where the teacher makes wrong predictions are considered hard tasks, and the knowledge transferred is biased knowledge.

Rethinking Knowledge Transmission

In the KD framework, the loss is typically composed of KL loss and CE loss, with the former used to constrain the student's learning of the knowledge transferred by the teacher and the latter to constrain the student's learning of the label. We use T and S to represent the predicted probability values of the teacher and the student, respectively, and Y to represent the label. $t_i \in T$, $s_i \in S$ and $y_i \in Y$ $i \in n$. Generally, we use

$$\begin{aligned} Loss &= \mathcal{L}_{KL}(T,S) + \mathcal{L}_{CE}(Y,S) \\ &= \sum_{i=1}^{n} t_i ln \frac{t_i}{s_i} + \left(- \sum_{i=1}^{n} y_i ln s_i \right) \end{aligned} \tag{1}$$

as the loss function. When minimizing the loss function through gradient backpropagation, minimizing KL will make the s_i approach t_i, and minimizing CE will make the s_i approach y_i. Let a and b represent two categories, where t_a and t_b are the predicted probabilities of the teacher, s_a and s_b are the predicted probabilities of the student, and y_a and y_b are the corresponding class labels. Assuming $y_a = 1$ and $y_b = 0$, when minimizing the loss function, we have

$$min\,(Loss) \Rightarrow \begin{cases} s_a \to 1\,, & s_a \to t_a\,; \\ s_b \to 0\,, & s_b \to t_b\,. \end{cases} \tag{2}$$

If the teacher's predictions match the labels (i.e., $t_a \to 1, t_b \to 0$), then it is beneficial for the student to learn from teacher, where $s_a \to t_a, s_b \to t_b$, and consequently $s_a \to 1$ and $s_b \to 0$. However, when the teacher's predictions do not match the labels (i.e., $t_a \to 0, t_b \to 1$), teaching the student to learn from teacher, where $s_a \to t_a, s_b \to t_b$, actually results in $s_a \to 1$ and $s_b \to 0$, which is misleading for the student.

In fact, since the teacher model's own prediction accuracy is not guaranteed to be perfect, the bias inevitably misleads the student. Therefore, we demonstrate that knowledge can be divided into right knowledge and bias, with the bias having a negative impact on the performance of the student model.

Eliminating Biased Knowledge from Teacher

To further eliminate biased knowledge in KD and ensure the rightness of knowledge transferred from the teacher modelk, we design a eliminate module, as depicted in the blue region in Fig. 2. Firstly, we convert the predicted values output by the teacher model into corresponding predicted probabilities, recorded as P^T. We use the $argmax$ operation to convert the teacher's predicted probabilities P^T into a 0-1 vector pre. This vector is then compared to the true labels using a logical AND operation, marking them as True if they match and False if they do not. We store these True and False values in the mask table in the order corresponding to the teacher's prediction index to record the rightness of the teacher's information. We classify the P^T into right knowledge (recorded as True in the mask table) and biased knowledge (recorded as False in the mask table). Thus far, we have distinguished right and biased knowledge in the teacher model and stored their indexes in the mask table. This enables students to later learn the knowledge transferred from the teacher under these indexes and perform gradient updates only using the corresponding knowledge in each batch of training.

We obtain the right knowledge by multiplying P^T and P^S with the mask. By inverting the mask to get $\sim$mask, we can then multiply P^T and P^S with $\sim$mask to obtain the biased knowledge.

Rectifying Biased Knowledge from Teacher

Eliminating bias can improve the overall predictive performance of the student model to some extent. However, since eliminating bias essentially removes the wrong predictions, it does not enhance the student model's ability to learn these hard tasks. As the data becomes more diverse and complex, the amount of bias will correspondingly increase, leading to a performance bottleneck in enhancing the student model through bias elimination. Therefore, we further propose a bias rectification method (as shown in Module ③ of Fig. 2) aimed at rectifying the teacher model's wrong predictions, thereby equipping the student model with the ability to learn hard tasks and further improving its predictive performance. Suppose t_a represents the predicted probability of the teacher corresponding to the label y_a as 1, t_b represents the largest predicted probability by the teacher, y_b represents the label value corresponding to t_b, $\mathbf{t_o} = \{t_{oi}\}$ represent the others probability prediction values, $\mathbf{y_o} = \{y_{oi}\}$ represents the others label values. y_a, y_b and $\mathbf{y_o}$ sum to 1, as shown in Fig. 3 (a). When we take the bias prior knowledge of the teacher, there is the following relationship:

$$t_a + t_b + \mathbf{t_o} = 1, \quad \begin{cases} y_a = 1, t_a \to 0\,; \\ y_b = 0, t_b \to 1\,; \\ \mathbf{y_o} = \{\mathbf{0}\}, \mathbf{t_o} \to \{\mathbf{0}\}\,. \end{cases} \tag{3}$$

In (3), t_a and t_b are opposite to the corresponding labels y_a and y_b, as we discussed, such biased knowledge from the teacher is harmful. We have considered the following requirements for rectifying this bias. First, we aim to adjust the values of t_a and t_b to be consistent with y_a and y_b . Second,

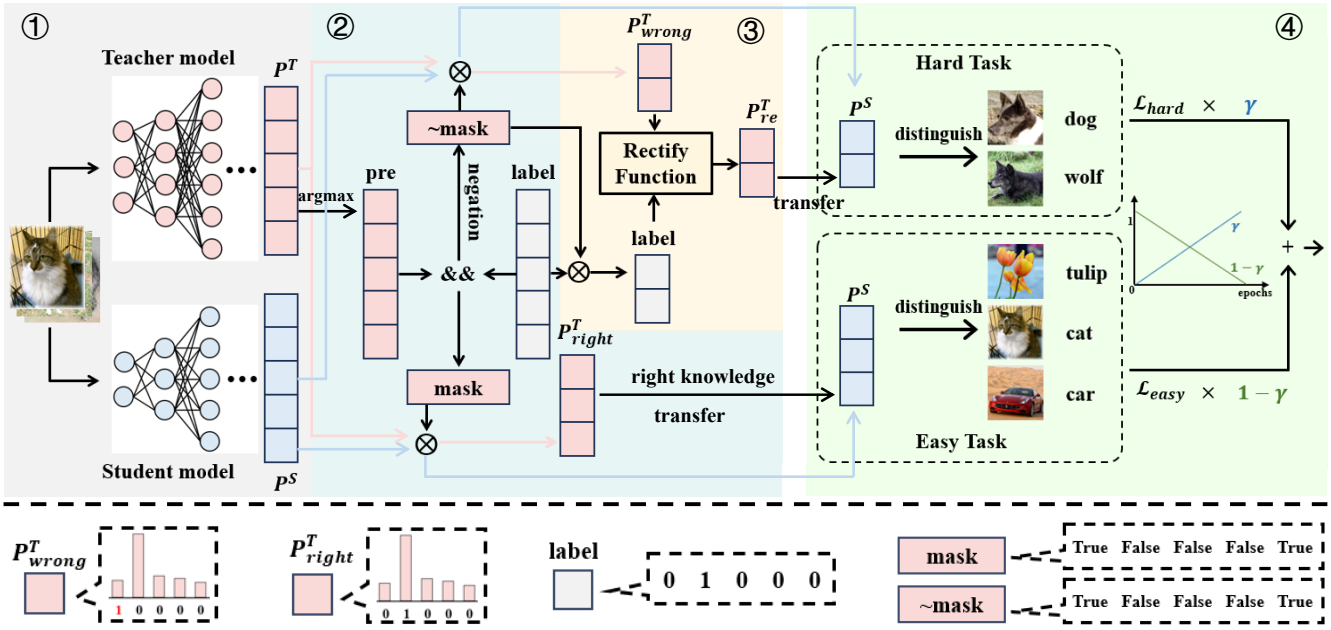

Figure 2: The overall framework of our framework. Module ① is a KD framework based on the teacher-student model. Module ② is the bias elimination method. Module ③ is the bias rectification method. Module ④ is the dynamic learning approach.

we ensure that the sum of all adjusted t_i values is 1 (satisfying the principle of probability distribution). Finally, since the $\mathbf{t_o}$ represents other classes unrelated to the two classes that need rectification, we cannot affect $\mathbf{t_o}$ when adjusting t_a and t_b. Based on these considerations, we propose the following strategy.

First, we weight t_a and t_b with the corresponding labels y_a and y_b. Then, we take the average to obtain the new probabilities $t^{'}$, as shown in the following equations:

$$\begin{cases} t_a^{'} = \frac{t_a + y_a}{2} = \frac{t_a + 1}{2} > 0.5, \\ t_b^{'} = \frac{t_b + y_b}{2} = \frac{t_b}{2} < 0.5, \\ \mathbf{t_o^{'}} = \mathbf{t_o} \end{cases} \tag{4}$$

The rectified probabilities obtained are shown in Fig. 3(b), where the largest probability value corresponds to a label value of 1, indicating that the predicted value correctly corresponds to the label value. This transformation from biased knowledge to right knowledge is illustrated.

Because

$$t_a^{'} + t_b^{'} = \frac{t_a + t_b + y_a + y_b}{2} = \frac{t_a + t_b + 1}{2}, \\ t_a + t_b < 1, \tag{5}$$

so

$$\frac{t_a + t_b + 1}{2} > t_a + t_b \Rightarrow t_a^{'} + t_b^{'} + \mathbf{t_o^{'}} > 1 \tag{6}$$

Considering that the sum of the probabilities for the new $t_a^{'}$, $t_b^{'}$ and $\mathbf{t_o^{'}}$ is not equal to 1, disrupting the principle of probability distribution. Therefore, we need to readjust the value of $t_a^{'}$ and $t_b^{'}$ to make $t_a^{'} + t_b^{'} + \mathbf{t_o^{'}} = 1$ without changing $\mathbf{t_o^{'}}$. Let t^{new} represent the adjusted value, then there is the following formula:

$$t_a^{new} = t_a^{'} \times \frac{t_a + t_b}{t_a^{'} + t_b^{'}}, \quad t_b^{new} = t_b^{'} \times \frac{t_a + t_b}{t_a^{'} + t_b^{'}} \tag{7}$$

At this point, we have obtained the final rectification probability values (shown in Fig. 3(c)), ensuring that the probabilities sum to one without altering the value of $\mathbf{t_o}$, while also ensuring that the new predicted results align with the label.

Dynamic Learning Approach

Since the student's learning efficiency differs between easy and hard tasks, we consider focusing on easy tasks in the early stages of training. As the student's capability improves, we increase the emphasis on tackling hard tasks in the later stages of training (as shown in Module ④ of Fig. 2). Both right knowledge and rectified knowledge play important roles in the student learning process and contribute to the student's eventual surpassing of the teacher. Consequently, we have defined two constraint functions to respectively constrain the student's learning of these two types of knowledge.

$$\mathcal{L}_{easy} = \mathcal{L}_{KL}\left(P^{S_r} \| P^{T_r}\right), \\ \mathcal{L}_{hard} = \mathcal{L}_{KL}\left(P^{S_{re}} \| P^{T_{re}}\right) \tag{8}$$

$\mathcal{L}_{easy}$ represents the loss function for transmitting easy task knowledge, where P^{T_r} denotes the probability distribution of the right knowledge in the teacher and P^{S_r} represents the probability distribution of the corresponding student. We use the KL divergence to constrain the approximation between the two distributions. $\mathcal{L}_{hard}$ represents the loss function for transmitting hard task knowledge, where $P^{T_{re}}$ represents the probability distribution of knowledge after rectification in the teacher and $P^{S_{re}}$ represents the probability

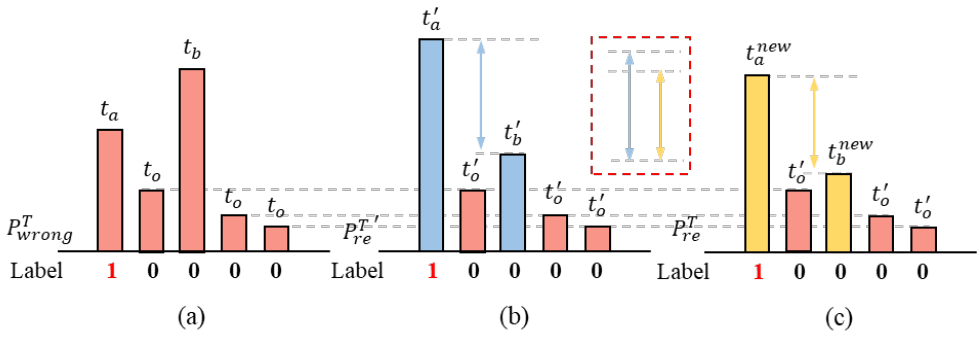

Figure 3: Diagram of the rectification method. Each column corresponds to a category, $Label$ corresponds to the one-hot vector of ground truth, P^T_{wrong} corresponds to the probability value predicted by the teacher on each category, P^T_{re} corresponds to the probability value after rectification. (a) shows the teacher's wrong prediction, (b) shows the predicted probability is consistent with $Label$ after rectification, (c) represents further adjustment of the rectification predicted probabilities.

distribution of the corresponding student. Similarly, we use the KL divergence to constrain the approximation between the two distributions.

The overall loss function is defined as follows:

$$\mathcal{L}_{all} = (1-\gamma)\left(\mathcal{L}_{CE} + \mathcal{L}_{easy}\right) + \gamma\mathcal{L}_{hard} \tag{9}$$

In order to shorten the training cycles, we employed a method for dynamically adjusting the student's learning focus. We have the following dynamic adjustment coefficient as $\gamma = \frac{e}{E}$. e represents the current training iteration, and E denotes the total training iterations.

By adjusting γ throughout the entire training process, the student is initially encouraged to prioritize learning easy tasks as the training iterations progress. This approach allows the student to quickly attain a foundational level of proficiency in basic knowledge. Subsequently, with the deepening of training, the focus gradually shifts towards the learning of more challenging knowledge. Effective fine-tuning on the established foundational proficiency leads to better performance enhancements. Ultimately, this strategy aims to achieve the goal of reducing the training time cost and improving the convergence effectiveness.

Experiment

We conducted experiments on three classification datasets, **CIFAR-10**(Krizhevsky, Hinton et al. 2009), **CIFAR-100**(Krizhevsky, Hinton et al. 2009), and **ImageNet 1K**(Russakovsky et al. 2015), as well as on an object detection dataset **MS-COCO**(Lin et al. 2014).

Comparison with SOTA KD methods

Classification on CIFAR dataset. In Table 1, we present the results of KD in two datasets. Our method demonstrates superior performance over the teacher on both the CIFAR-10 and CIFAR-100 datasets. Due to the increased difficulty of the 100-class classification task compared to the 10-class task, the teacher model makes more wrong predictions, resulting in more bias being transmitted. As a result, our method demonstrates superior performance.

For cross-model KD, we employed MobileNet (Sandler et al. 2018) and ShuffleNet (Zhang et al. 2018) as student models, both exhibiting significant structural and parameter differences compared to the teacher models (ResNet-50 (He et al. 2016), VGG (Simonyan and Zisserman 2014) and WRN (Zagoruyko and Komodakis 2016b)). This is sufficient to validate the capability of cross-model knowledge distillation. In this scenario, our method still achieves the best performance, enabling the student model to surpass the teacher model. Particularly in the challenging CIFAR-100 task, our method shows the most significant improvement compared to the SOTA methods.

Classification on ImageNet 1K dataset. As shown in Table 2, we validated the capability of our method for KD where the teacher and student have similar models. Using ResNet-34 (81.39MB) as the teacher model and ResNet-18 (42.83MB) as the student model, our method demonstrates the ability to surpass the teacher model on challenging classification tasks. Moreover, compared to the best KD method, our approach achieved an additional 1.75% improvement in classification accuracy. In Table 3, we validated the capability of our method in a cross-model knowledge distillation. Employing ResNet-50 (90.46MB) as the teacher model and MobileNet-V1 (9.01MB) as the student model, our method continues to enable the student network to surpass the teacher model in cross-model knowledge transfer. Furthermore, compared to other methods, our method achieves a substantial 3.68% improvement in classification accuracy over the best-performing method. This improvement is particularly significant in challenging classification tasks.

Object detection on MS-COCO dataset. To validate the effectiveness of our method not only in classification tasks but also in object detection tasks, we applied our method on the MS-COCO dataset. The ultimate performance of object detection in this task is heavily dependent on the quality of feature extraction, particularly when dealing with the detection of numerous small objects, which significantly challenges the detector's feature extraction capability (Li, Jin, and Yan 2017; Wang et al. 2019). As shown in Table 4,

CIFAR-10 dataset	teacher	ResNet-50 96.08	ResNet-101 96.83	WRN-40-2 93.52	VGG-13 93.28	ResNet-50 96.08	VGG-13 93.28	ResNet-50 96.08	WRN-40-2 93.52
	student	ResNet-18 93.30	ResNet-34 94.59	WRN-16-2 90.50	VGG-8 88.78	MobileNet-V2 90.08	MobileNet-V2 90.08	ShuffleNet-V1 90.70	ShuffleNet-V1 90.70
	KD(Hinton, Vinyals, and Dean 2015)	93.81	94.87	92.73	89.30	93.79	91.73	91.21	91.09
	FitNet(Romero et al. 2014)	93.87	94.80	92.01	89.70	93.70	91.79	91.30	90.97
	RKD(Park et al. 2019)	94.37	95.01	92.79	89.93	94.01	92.07	92.47	91.21
	CRD(Tian, Krishnan, and Isola 2019)	94.90	95.39	93.01	90.74	94.21	92.37	92.80	91.30
	OFD(Heo et al. 2019)	94.87	95.20	92.97	90.51	94.22	92.20	92.59	91.24
	ReviewKD(Chen et al. 2021)	95.07	95.57	93.17	91.99	95.08	92.90	93.97	92.68
	Indistill(Sarridis et al. 2022)	94.99	95.19	93.20	91.63	94.38	92.43	93.50	92.06
	DKD(Zhao et al. 2022)	95.02	95.49	93.21	91.75	94.79	92.81	93.89	92.60
	CTKD(Li et al. 2023)	94.73	95.24	93.01	91.10	94.20	92.05	93.41	92.19
	Ours	**96.89**	**96.99**	**94.18**	**93.60**	**96.28**	**93.76**	**96.20**	**93.70**
	Δ	+1.82	+1.42	+0.97	+1.61	+1.20	+0.86	+2.23	+1.02
CIFAR-100 dataset	teacher	ResNet-50 79.34	ResNet-101 81.87	WRN-40-2 73.57	VGG-13 74.64	ResNet-50 79.34	VGG-13 74.64	ResNet-50 79.34	WRN-40-2 73.57
	student	ResNet-18 69.75	ResNet-34 72.50	WRN-16-2 70.26	VGG-8 70.36	MobileNet-V2 65.40	MobileNet-V2 65.40	ShuffleNet-V1 70.50	ShuffleNet-V1 70.50
	KD(Hinton, Vinyals, and Dean 2015)	71.56	73.17	70.92	72.98	69.35	67.37	71.97	70.83
	FitNet(Romero et al. 2014)	70.21	73.08	70.98	71.02	65.56	64.14	71.03	70.73
	RKD(Park et al. 2019)	71.67	73.87	71.32	71.48	66.73	64.52	72.84	71.21
	CRD(Tian, Krishnan, and Isola 2019)	72.16	74.60	71.37	73.94	71.11	69.73	73.10	72.05
	OFD(Heo et al. 2019)	71.98	73.91	71.10	73.95	71.04	69.48	72.78	71.85
	ReviewKD(Chen et al. 2021)	73.19	75.80	71.59	74.84	72.89	70.37	76.10	73.14
	Indistill(Sarridis et al. 2022)	73.17	75.17	71.09	74.65	72.36	70.01	75.48	72.10
	DKD(Zhao et al. 2022)	73.97	75.67	71.54	74.68	72.35	69.71	75.88	73.10
	CTKD(Li et al. 2023)	72.29	74.58	71.45	73.52	70.46	68.46	75.34	71.78
	Ours	**81.50**	**84.77**	**74.30**	**78.03**	**80.45**	**75.71**	**81.07**	**74.78**
	Δ	+7.53	+8.97	+2.71	+3.19	+8.10	+5.34	+4.97	+1.64

Table 1: The results on the CIFAR-10 and CIFAR-100 dataset. Δ represents the classification accuracy improvement over the best result of the current SOTA methods in knowledge distillation.

		top-1 acc(%)	top-5 acc(%)
	Teacher: ResNet-34	73.31	91.42
	Student: ResNet-18	69.75	89.07
features	AT(Zagoruyko and Komodakis 2016a)	70.69	90.01
	OFD(Heo et al. 2019)	70.81	89.98
	CRD(Tian, Krishnan, and Isola 2019)	71.17	90.13
	ReviewKD(Chen et al. 2021)	71.61	90.51
	InDistill(Sarridis et al. 2022)	71.63	90.37
logits	KD(Hinton, Vinyals, and Dean 2015)	71.03	90.05
	DKD(Zhao et al. 2022)	71.70	90.41
	CTKD(Li et al. 2023)	71.32	90.27
	Ours	**73.38**	**91.71**
	Δ	+1.75	+1.2

Table 2: acc refers to classification accuracy(%) on the ImageNet 1K dataset, top-1 and top-5 are standards for calculating classification accuracy on the validation set. Δ represents the classification accuracy improvement over the best result of the current SOTA methods in knowledge distillation.

		top-1 acc(%)	top-5 acc(%)
	Teacher: ResNet-50	76.16	92.86
	Student: MobileNet-V1	68.87	88.76
features	AT(Zagoruyko and Komodakis 2016a)	69.56	89.33
	OFD(Heo et al. 2019)	71.25	90.34
	CRD(Tian, Krishnan, and Isola 2019)	71.37	90.41
	ReviewKD(Chen et al. 2021)	72.56	91.00
	InDistill(Sarridis et al. 2022)	72.52	91.53
logits	KD(Hinton, Vinyals, and Dean 2015)	70.50	89.80
	DKD(Zhao et al. 2022)	72.50	91.50
	CTKD(Li et al. 2023)	71.47	90.65
	Ours	**76.24**	**93.38**
	Δ	+3.68	+1.85

Table 3: acc refers to classification accuracy(%) on the ImageNet 1K dataset, top-1 and top-5 are standards for calculating classification accuracy on the validation set. Δ represents the classification accuracy improvement over the best result of the current SOTA methods in knowledge distillation.

	AP	AP_{50}	AP_{70}
Teacher: ResNet-101	42.04	62.48	45.88
Student: ResNet-18	33.26	53.61	35.26
KD(Hinton, Vinyals, and Dean 2015)	33.97	54.66	36.62
FitNet(Romero et al. 2014)	34.13	54.16	36.71
ReviewKD(Chen et al. 2021)	36.75	56.72	34.00
InDistill(Sarridis et al. 2022)	34.93	56.56	37.46
DKD(Zhao et al. 2022)	35.05	56.60	37.54
CTKD(Li et al. 2023)	34.56	55.43	36.91
Ours	**42.10**	**62.59**	**45.91**
Teacher: ResNet-50	40.22	61.02	43.81
Student: MobileNet-V2	29.47	48.87	30.90
KD(Hinton, Vinyals, and Dean 2015)	30.13	50.28	31.35
FitNet(Romero et al. 2014)	30.20	49.80	31.69
ReviewKD(Chen et al. 2021)	33.71	53.15	36.13
InDistill(Sarridis et al. 2022)	32.17	53.49	34.56
DKD(Zhao et al. 2022)	32.34	53.77	34.01
CTKD(Li et al. 2023)	31.39	52.34	33.10
Ours	**41.02**	**62.13**	**44.08**

Table 4: Results on MS-COCO based on Faster-RCNN (Ren et al. 2015)-FPN(Lin et al. 2017): AP evaluated on val2017.

our method outperforms the teacher on three metrics: AP, AP_{50}, and AP_{70}. This accomplishment is often challenging to achieve with conventional knowledge distillation methods.

Ablation Analysis

We have validated the effectiveness of two modules for the elimination and rectification of biased knowledge, as shown in Table 5. EBK represents KD after the elimination of biased knowledge, while RBK represents KD after the rectification of biased knowledge. Validation was performed on the CIFAR-100 dataset, with the teacher model selected

Student	EBK	RBK	acc(%)	Δ
ResNet-50 as the teacher(79.34)				
ResNet-18			71.56	-
	√		79.70	+8.26
		√	73.86	+2.30
	√	√	81.50	+9.94
MobileNet-V2			69.35	-
	√		79.85	+10.50
		√	71.27	+1.92
	√	√	80.45	+11.10

Table 5: The acc denotes the classification accuracy on the CIFAR-100 validation dataset, and the Δ represents the improvement in classification accuracy compared to the baseline KD method. The √ symbol indicates the usage of the corresponding module, while the absence of √ indicates the absence of both modules, implying the use of the basic KD method.

as ResNet-50 (79.34%). We chose ResNet-18 (69.75%) and MobileNet-V2 (65.40%) with the same model structures and different as the student models, respectively. It can be observed that when using either of the two modules alone, the performance surpasses that of regular knowledge distillation. The best results are achieved when both modules are applied together in knowledge distillation. This effect is evident across both identical and different model structures.

As shown in Fig. 4, our method is compared with several representative methods. It can be seen that within the same training period, our method (pink line) achieves the highest final accuracy. Our method requires fewer training epochs to reach the same accuracy level of the student model compared to others, excelling in both final accuracy and training speed. By using the approach, we achieved dynamic adjustments throughout the overall training process. This approach helps accelerate the learning process of the student network during knowledge distillation. Compared to methods that do not adopt this approach, it reduces the training time cost by 25%.

Wide Applicability. Our method serves as a knowledge distillation strategy to enhance the effectiveness of existing knowledge distillation methods. As shown in Table 6, our method consistently improves both feature-based and logit-based knowledge distillation methods. The combination with the DKD(Zhao et al. 2022) method has achieved the best results, which is attributed to the ability of eliminating the influence of biased knowledge by our method, strengthening the transmission of right knowledge, and addressing the confusion caused by biased knowledge for students learning with the DKD method. Therefore, the combination of these two methods has yielded the most outstanding learning outcomes. In feature-based methods, the student network may be influenced by certain misleading information present in the teacher network's intermediate layer knowledge, which is challenging to effectively eliminate. Consequently, the performance improvement of such methods is not as significant as that of logit-based methods. However, feature-based methods still offer notable performance gains.

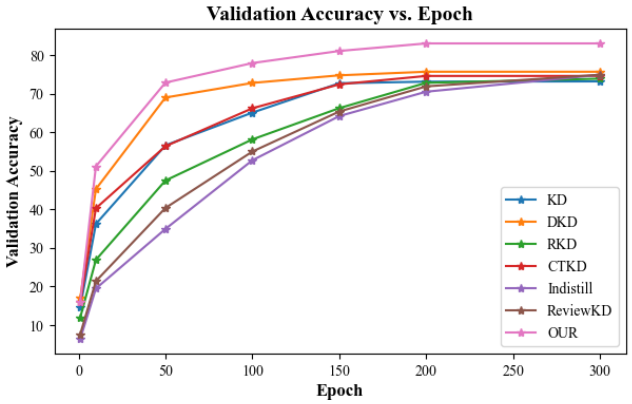

Figure 4: The accuracy variation curves during the training process for several KD methods. The x-axis is training epochs, and the y-axis is the accuracy of the ResNet-34 model on the CIFAR-100 validation set.

		acc(%)	Δ
	teacher(ResNet-50)	79.34	-
	student(ResNet-18)	69.75	-
features	FitNet(Romero et al. 2014)	70.21	-
	FitNet+Ours	72.96	+2.75
	RKD(Park et al. 2019)	71.67	-
	RKD+Ours	74.07	+2.40
	CRD(Tian, Krishnan, and Isola 2019)	72.16	-
	CRD+Ours	74.21	+2.05
	InDistill(Sarridis et al. 2022)	73.17	-
	InDistill+Ours	74.83	+1.66
logits	KD(Hinton, Vinyals, and Dean 2015)	71.56	-
	KD+Ours	78.81	+7.25
	DKD(Zhao et al. 2022)	73.97	-
	DKD+Ours	80.17	+6.20
	CTKD(Li et al. 2023)	72.29	-
	CTKD+Ours	79.00	+6.71

Table 6: The table compares the performance of our method applied to other knowledge distillation mathod. Accuracy represents the classification accuracy on the CIFAR-100 dataset, Δ indicates the difference from the original method results, and the + indicates improvement.

Conclusion

We introduce a novel KD framework that tackles the challenge of biased knowledge transferring from teacher to student. This novel framework can completely eliminate the influence of biased knowledge of the teacher and substantially enhance the student by correcting knowledge through an elaborately rectifying strategy. We conduct extensive experiments on four widely used datasets and eight sets of teacher-student models, achieving SOTA results. Experimental results validate that our framework effectively boosts the performance of student, even surpassing the teacher in most scenarios. Furthermore, our framework is compatible with various existing KD methods and enhances their effectiveness.

Acknowledgements

This work is supported by the National Natural Science Foundation of China Excellent Young Scientists Fund (Grant No. T2422015), the National Natural Science Foundation of China Youth Fund (Grant No. 62202137 and 62306212) and the Marie Skłodowska-Curie Postdoctoral Individual Fellowship under Grant No. 101154277.

References

Chen, K.; Zhang, J.; Liu, J.; Tong, Q.; Liu, R.; and Chen, S. 2022. Semantic visual simultaneous localization and mapping: A survey. *arXiv preprint arXiv:2209.06428*.

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

Cho, Y.; Ham, G.; Lee, J.-H.; and Kim, D. 2023. Ambiguity-aware robust teacher (ART): Enhanced self-knowledge distillation framework with pruned teacher network. *Pattern Recognition*, 140: 109541.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129: 1789–1819.

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. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1921–1930.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network (2015). *arXiv preprint arXiv:1503.02531*, 2.

Huang, Z.; and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Kim, S.; and Kim, H. 2017. Transferring Knowledge to Smaller Network with Class-Distance Loss. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

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

Li, Q.; Jin, S.; and Yan, J. 2017. Mimicking very efficient network for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 6356–6364.

Li, S.; Cheng, X.; Shi, F.; Zhang, H.; Dai, H.; Zhang, H.; and Chen, S. 2024. A Novel Robustness-Enhancing Adversarial Defense Approach to AI-Powered Sea State Estimation for Autonomous Marine Vessels. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1–15.

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.

Liang, M.; Huang, S.; and Liu, W. 2024. Dynamic semantic structure distillation for low-resolution fine-grained recognition. *Pattern Recognition*, 148: 110216.

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.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Park, D. Y.; Cha, M.-H.; Kim, D.; Han, B.; et al. 2021. Learning student-friendly teacher networks for knowledge distillation. *Advances in Neural Information Processing Systems*, 34: 13292–13303.

Park, J.; and No, A. 2022. Prune Your Model Before Distill It. In *European Conference on Computer Vision*, 120–136. Springer.

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.

Passalis, N.; and Tefas, A. 2018. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 268–284.

Patel, G.; Mopuri, K. R.; and Qiu, Q. 2023. Learning to Retain while Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7786–7794.

Paulin, D.; and Suneson, K. 2012. Knowledge transfer, knowledge sharing and knowledge barriers–three blurry terms in KM. *Electronic Journal of Knowledge Management*, 10(1): pp82–92.

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.

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*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Sarridis, I.; Koutlis, C.; Papadopoulos, S.; and Kompatsiaris, I. 2022. InDistill: Transferring Knowledge From Pruned Intermediate Layers. *arXiv preprint arXiv:2205.10003*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tang, R.; Liu, Z.; Li, Y.; Song, Y.; Liu, H.; Wang, Q.; Shao, J.; Duan, G.; and Tan, J. 2023. Task-balanced distillation for object detection. *Pattern Recognition*, 137: 109320.

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

Wang, L.; and Yoon, K.-J. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3048–3068.

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.

Xie, Y.; Wu, H.; Lin, Y.; Zhu, J.; and Zeng, H. 2024. Pairwise difference relational distillation for object reidentification. *Pattern Recognition*, 152: 110455.

Xu, G.; Liu, Z.; and Loy, C. C. 2023. Computation-Efficient Knowledge Distillation via Uncertainty-Aware Mixup. *Pattern Recognition*, 138: 109338.

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

Zagoruyko, S.; and Komodakis, N. 2016b. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, F.; Qu, S.; Shi, F.; and Xu, C. 2024. Overcoming the Pitfalls of Vision-Language Model for Image-Text Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 2350–2359. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zhang, H.; Hu, Z.; Qin, W.; Xu, M.; and Wang, M. 2021. Adversarial co-distillation learning for image recognition. *Pattern Recognition*, 111: 107659.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856.

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.

Zhou, Z.; Zhuge, C.; Guan, X.; and Liu, W. 2020. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*.

Zhu, J.; Tang, S.; Chen, D.; Yu, S.; Liu, Y.; Rong, M.; Yang, A.; and Wang, X. 2021. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9260–9269.

Zhu, S.; Shang, R.; Yuan, B.; Zhang, W.; Li, W.; Li, Y.; and Jiao, L. 2024. DynamicKD: An effective knowledge distillation via dynamic entropy correction-based distillation for gap optimizing. *Pattern Recognition*, 153: 110545.