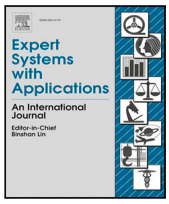
Expert Systems With Applications 262 (2025) 125579

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Why does Knowledge Distillation work? Rethink its attention and fidelity mechanism

Chenqi Guo [a,*], Shiwei Zhong [a], Xiaofeng Liu [a], Qianli Feng [b], Yinglong Ma [a]

[a] Control and Computer Engineering, North China Electric Power University, No. 2 Beinong Road, Beijing, 102206, PR China
[b] Amazon, 300 Boren Ave N, Seattle, 98109, WA, USA

ARTICLE INFO

Keywords:
Knowledge distillation
Ensemble learning
Attention mechanism
Supervised image classification
Data augmentation

ABSTRACT

Does Knowledge Distillation (KD) really work? Conventional wisdom viewed it as a knowledge transfer procedure where a perfect mimicry of the student to its teacher is desired. However, paradoxical studies indicate that closely replicating the teacher's behavior does not consistently improve student generalization, posing questions on its possible causes. Confronted with this gap, we hypothesize that diverse attentions in teachers contribute to better student generalization at the expense of reduced fidelity in ensemble KD setups. Focusing on supervised image classification task, by increasing data augmentation strengths, our key findings reveal a decrease in the Intersection over Union (IoU) of attentions between teacher models, leading to reduced student overfitting and decreased fidelity. We propose this low-fidelity phenomenon as an underlying characteristic rather than a pathology when training KD. This suggests that stronger data augmentation fosters a broader perspective provided by the divergent teacher ensemble and lower student–teacher mutual information, benefiting generalization performance. We further demonstrate that even optimization towards logits-matching between teachers and student can hardly mitigate this low-fidelity effect. These insights clarify the mechanism on low-fidelity phenomenon in KD. Thus, we offer new perspectives on optimizing student model performance, by emphasizing increased diversity in teacher attentions and reduced mimicry behavior between teachers and student. Codes are available at https://github.com/zisci2/RethinkKD

1. Introduction

In the realm of supervised image classification, Knowledge Distillation (KD) (Hinton et al., 2015) is renowned for its effectiveness in deep model compression and enhancement, emerging as a critical technique for knowledge transfer. Previously, this process has been understood and evaluated through model *fidelity* (Stanton et al., 2021), measured by the student model replication degree to its teachers. High fidelity, assessed by metrics like low averaged predictive Kullback–Leibler (KL) divergence and high top-1 logits agreement (Stanton et al., 2021), have conventionally been used to assess the success of KD.

While fidelity has traditionally guided enhancements in model architectures, optimization, and training frameworks, repeated high-fidelity results corresponding to strong student performance seem to indicate that a high degree of mimicry between the student and teachers is desirable (Lao et al., 2023; Li, Li et al., 2022; Wang et al., 2022). Yet this notion was initially challenged in Stanton et al. (2021), which empirically shows that good student accuracy does not imply good distillation fidelity in self and ensemble distillation. However, though (Stanton et al., 2021) underscores their empirical findings on the low-fidelity phenomenon, they still believe that closely matching the teacher is beneficial for KD in terms of knowledge transfer. Further, they identify optimization difficulties as one key reason of student's poor emulation behavior to its teachers during training. Thus, this paradox highlights a need for further exploration on model fidelity and its mechanism in KD.

Among factors in KD analysis, the attention map mechanism serves as a pivotal role in understanding the student-teacher interplay. It is known that in ensemble learning, diverse models improve the overall performance. Tsantekidis et al. (2021) empirically shows that diversifying teachers' learnt policies by training them in different subsets of learning environment, can enhance the distilled student performance in KD. Yet, a theoretical foundation is lack for doing so. Allen-Zhu and Li (2023) proved how the multi-view structure of training images contributes to this improvement. However, beyond examining feature structures within an image from a dataset perspective, assessing model diversity by analyzing the ensemble models' attention maps can provide

* Corresponding author.
E-mail addresses: chenqiguo72@ncepu.edu.cn (C. Guo), zsw@ncepu.edu.cn (S. Zhong), liu_xf@ncepu.edu.cn (X. Liu), fengq@amazon.com (Q. Feng), yinglongma@ncepu.edu.cn (Y. Ma).

https://doi.org/10.1016/j.eswa.2024.125579
Received 14 August 2024; Received in revised form 10 October 2024; Accepted 15 October 2024
Available online 21 October 2024

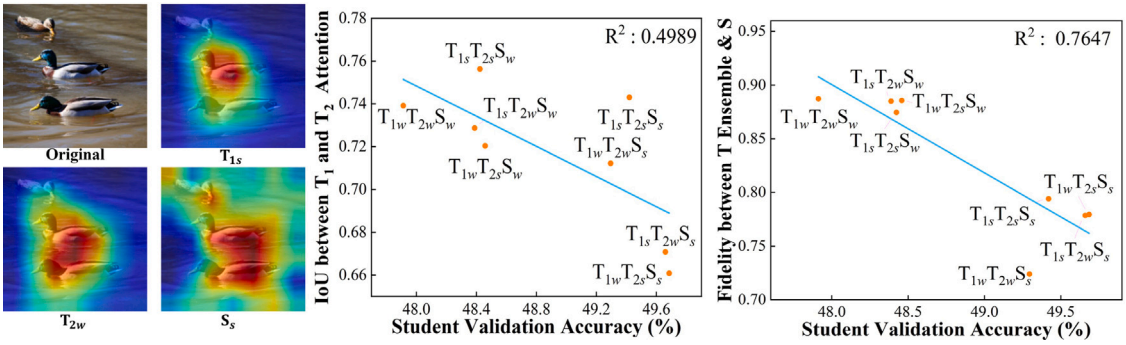


Fig. 1. *Left*: Attention map visualizations for teacher ensembles and student model in Knowledge Distillation (KD) on ImageNet dataset. Stronger data augmentation ($T_{1w}T_{2s}S_s$ and $T_{1s}T_{2w}S_s$ in this case) as measured by *Affinity* improves teachers' attentional divergence, thus providing the student a more comprehensive perspective on the overall characteristics of the target images, leading to a better generalization ability. This is discussed in detail in Section 6.1. *Middle and Right*: Scatter plots of Intersection over Union (IoU) in Attention maps, and Fidelity between teacher ensembles and student during KD training. The decreasing tendency in fidelity challenges the conventional wisdom that higher fidelity consistently correlate with better student performance. Later in Section 6.3 we will demonstrate that the low-fidelity observation is caused by attention map diversification existed within teacher ensembles, and even optimization towards logits-matching can hardly mitigate this low-fidelity effect.

us with more straightforward insights into the learning dynamics, which has been overlooked in previous studies. Besides, it would be intriguing to check the student-teacher fidelity under such circumstance, to see if diversifying teacher models in an ensemble consistently corresponds with low-fidelity as well. If so, one can devote model attention map diversities to explain the existing fidelity paradox. Thus in this paper, we utilize the Intersection over Union (IoU) (Rezatofighi et al., 2019) of attention maps (Zhou et al., 2016) between different teacher models in ensemble KD to help elucidate the existing fidelity paradox.

Following the investigation paradigm in Stanton et al. (2021), where the model fidelity variations were observed with different data augmentations, we adapt this paradigm to our case with a more cautious control over the degree of randomness in augmentation during ensemble KD training. By varying data augmentation strengths, as measured by *Affinity* (Cubuk et al., 2021), which will be introduced later, we modulated the model diversities trained on them. Impacts not only on traditional metrics like student–teacher fidelity, but also on less-explored aspects of attention maps diversity between different teachers, and mutual information between student and teachers are witnessed. Our empirical observations appear to challenge the traditional wisdom on the student–teacher relationship in distillation during training procedure and thus provide further insights on explaining the fidelity paradox.

Specifically, in support and further complement to Stanton et al. (2021), we highlight attention map diversification existed within teacher ensembles as a deeper reason why a student with good generalization performance may be unable to match the teacher during KD training: Stronger data augmentation increases attention divergence in the teacher ensemble, enabling teachers to offer a broader perspective to the student. Consequently, the student surpassing the knowledge of single teacher becomes more independent as measured by lower student-teacher mutual information. And the low-fidelity observed is a demonstration of this phenomenon.

Furthermore, though (Stanton et al., 2021) has demonstrated the low-fidelity observation, they still proposed the difficulties in optimization as the primary reason for it. And recent works including (Sun et al., 2024) remain optimizing in the direction of facilitating the student-teacher emulation procedure. Yet our empirical and theoretically analysis demonstrate that, optimization with logits matching does improve the student generalization ability but is still at the cost of fidelity reduction.

Our primary goal is to explain the fidelity paradox and understand the student learning and knowledge transfer dynamics in ensemble KD, by observing the implications of data augmentation on the student-teacher relationship. By doing so, we seek to provide insights that challenge the traditional or extend preliminary wisdom in KD fidelity by leveraging the attention mechanism in ensemble learning. As shown in Fig. 1, we summarize our contributions as follows:

(1) We demonstrate the correlation between teachers' attention map diversity and student model accuracy in ensemble KD training. Stronger data augmentation improves attentional divergence among teacher models, offering the student a more comprehensive perspective.
(2) We affirm the viewpoint from Stanton et al. (2021) that higher fidelity between teachers and student does not consistently improve student performance. What is more, through analyzing attention maps between teachers in ensemble KD, we highlight this low-fidelity phenomenon as an underlying characteristic rather than a pathology: Student's generalization is enhanced with more diverse teacher models, which causes the reduction in student-teacher fidelity.
(3) We examine data augmentation's effects on learning dynamics in ensemble KD. Through systematically analyzing the impact of modulated data augmentation strengths on the learning dynamics within KD, we offer a novel, simple yet effective perspective on optimizing the ensemble KD learning processes.
(4) We further investigate if optimization towards facilitating the student-teacher logits matching procedure can enhance the KD fidelity. Our empirical and theoretically analysis demonstrate that such optimization improve the student generalization ability but still at the cost of fidelity reduction.

The rest of the paper is structured as follows: Section 2 summarizes the related works, Section 3 clarifies the problem and hypothesis focused in this work, and Section 4 introduces the evaluation metrics used to validate our argues. Section 5 further gives the experimental settings, and the empirical results and theoretical analysis are provided in Section 6. Section 7 compares our method of modulated data augmentations on ensemble KD, with the SOTA KD baselines. Section 8 offers the ablation study. Section 11 finally summarizes the work of this paper.

2. Related works

Our study contributes to a growing body of research that explores the interactions between data augmentation, model fidelity, attention mechanisms, and their impact on student performance in Knowledge Distillation (KD) with teacher ensembles.

From a dataset perspective, Allen-Zhu and Li (2023) attributes the success of ensemble KD to the multi-view structure commonly found in vision task datasets. For instance, a car image can be identified as a car

by focusing on features like the headlights, wheels, or windows, which are considered positive indicators. However, the headlights might also resemble a cat's eye, which could be a negative feature. These features, along with the soft labels in KD, serve as "dark knowledge" for the student model to learn from the ensemble of teachers. In other words, from a single car image, the student can learn both car and cat features, thereby enhancing its performance. However, beyond analyzing feature structures within an image from a dataset perspective, evaluating model diversity through the attention maps of ensemble models can offer more direct insights into the learning dynamics, a dimension that previous studies have largely overlooked.

In Bai et al. (2023), a KD framework utilizing Masked Autoencoder, one of the primary factors influencing student performance is the randomness introduced by masks in its teacher ensembles. It comes naturally if incorporating randomness into the dataset, through a simple yet effective method like data augmentation, and carefully controlling its strength, will be as effective as integrating it into model architectures.

Yet, theories on the impacts of data augmentation on KD remain diverse and varied. Li, Shao et al. (2022) offers theoretical insights, suggesting that leveraging diverse augmented samples to aid the teacher model's training can enhance its performance but will not extend the same benefit to the student. Shen et al. (2022) emphasizes how data augmentation can alter the relative importance of features, making challenging features more likely to be captured during the learning process. This effect is analogous to the multi-view data setting in ensemble learning, suggesting that data augmentation is likely to be beneficial for ensemble KD.

On the application front, research proposing novel attention-based KD frameworks usually accompanied with intricate designs in model architectures or data augmentation strategies (Lewy & Mańdziuk, 2023; Özdemir & Sönmez, 2022). For instance, studies like Tian and Chen (2022) aim to address the few shot learning in KD with a novel data augmentation strategy based on the attentional response of the teacher model. Gou et al. (2023) proposed a hierarchical multi-attention transfer framework (HMAT), which employs various types of attention to facilitate knowledge transfer at different levels of deep representation learning for KD. Although their concentration is different from ours, the studies nevertheless show the significance of attention mechanism in KD.

In align with the initial "knowledge transfer" definition of KD, as an underlying assumption that a higher degree of emulation between the student and teachers benefits its training, previous studies are devoted to optimizing towards increased student-teacher fidelity or mutual information (Lao et al., 2023; Li, Li et al., 2022; Wang et al., 2022). Recent work Sun et al. (2024) also optimizes in this direction, where a z-score logit standardization process is proposed to mitigate the logits matching difficulties caused by logit shift and variance match between teacher and student. Nevertheless, this idea faced initial challenge in Stanton et al. (2021), indicating that closely replicating the teacher's behavior does not consistently lead to significantly improved student generalization performance during testing, whether in self-distillation or ensemble distillation.

Stanton et al. (2021) first investigates if the low-fidelity is an identifiability problem that can be solved by augmenting the dataset, and the answer is no: experimental results show subtle benefits of this increased distillation dataset. They further explore if the low-fidelity is an optimization problem resulting in a failure of the student to match the teacher even on the original training dataset, and their answer is yes: A shared initialization does make the student slightly more similar to the teacher in activation space, but in function space the results are indistinguishable from randomly initialized students.

Though insightful, it prompts further questions and drives us to think: Is low-fidelity truly undesirable and problematic for KD, especially if it does not harm student performance? Thus, additional exploration into this student fidelity-performance relation is required to elucidate the above paradox. Adopting a similar investigative approach which observes model fidelity variations with different data augmentations, we tailor it to our case, exercising a more cautious control over the data augmentation strength and thus the randomness into the distillation dataset during KD training.

In our work, we applied various data augmentations on KD, aiming to provide a more comprehensive understanding of model fidelity and attention mechanisms. Our empirical results and theoretical analysis challenge conventional wisdom, supporting and extending (Stanton et al., 2021) by demonstrating that student-teacher fidelity or mutual information does decrease with improved student performance during KD training. And, this low-fidelity phenomenon can hardly be mitigated with optimization aimed at improving student generalization. We thus advocate for more cautious practices in future research when designing KD strategies.

3. Problem and hypothesis

We focus on Knowledge Distillation (KD) with teacher ensembles in supervised image classification. In this realm, the efficacy of the process has traditionally been evaluated through the model fidelity and student validation accuracy. However, this conventional approach may not fully capture the complexity and nuances inherent in the knowledge transfer process, especially in light of evolving practices like data augmentation and the growing importance of attention mechanisms in neural networks. This study is driven by a series of interconnected research questions that challenge and extend the traditional understanding of KD as follows.

Impact of Varied Data Augmentation Strengths on Model Diversity in Attention Map Mechanisms. The application of diverse data augmentation strengths during the training of teacher and student models plays a crucial role in shaping KD (Stanton et al., 2021). Consequently, it is natural to inquire whether, across augmentation strategies, stronger data augmentation results in an increase or decrease in model fidelity within teacher ensembles during training. And if so, how does this correlate with the student model's performance. Inspired by the theory in machine learning that diversity among models can enhance ensemble learning performance (Asif et al., 2019; Zhou, 2012), our hypothesis is that varying augmentation strengths in different teachers inject randomness into the data, thereby diversifying teacher models' attention (Zhou et al., 2016) mechanisms trained on them. This diversity promotes heterogeneity in learning features, enables the student to learn diverse solutions to the target problem, and thus enhances the KD process. As a result, the student surpasses the knowledge of a single teacher, leading to a better overall performance, and the observed low-fidelity serves as a demonstration of this phenomenon.

Interplay Between Student Fidelity, Mutual Information and Generalization. Shrivastava et al. (2023) and Stanton et al. (2021) have observed that fidelity or mutual information between teacher and student models interact with varying data augmentation strengths, influencing the overall effectiveness of distilled knowledge. The critical questions then arise: Does lower or higher fidelity and mutual information benefit the KD training and student performance, and why does it happen? We hypothesize that, varied augmentation strengths in different teachers in ensemble KD would provide a broader view for the student to learn. Thus, the student surpassing the knowledge of a specific teacher. Contrary to the traditional perspective, we expect a decreased mimicry behavior of the student to benefit the student generalization ability during training, as it learns more intricate patterns from the diverse set of teachers.

Effect of Optimization towards Student-Teacher Logits Matching on Fidelity. Question also comes on why some works thought a high-fidelity is beneficial, while others thought a low-fidelity is inevitable during training. Our intuition is that the researches devoted to optimizing towards increased student-teacher fidelity or mutual

information do achieve the ultimate goal of improving the overall student performance, but in fact fail at enhancing the mimicry behavior during training. In this paper, we try to answer this question by delving into a logits matching KD case as in Sun et al. (2024). Specifically, we experiment with a z-score standardization method to mitigate the logits magnitudes and variance gap between teacher and student, which facilitates the student-teacher emulation procedure. Our hypothesis is that though such an optimization can relieve the logit shift and variance match problem, in reality its benefit lies in the student generalization rather than the fidelity improvement.

These questions aim to dissect the underlying learning dynamics in KD, moving beyond traditional metrics and exploring how newer facets like data augmentation strength, attention map diversity, fidelity and mutual information interplay to influence the student's learning and generalization abilities. Here, the data augmentation strength is measured by Affinity (Cubuk et al., 2021), the offset in data distribution between the original one and the one after data augmentation as captured by the student model, which we will talk more later. By addressing these questions, this study seeks to provide a more comprehensive understanding of KD.

4. Evaluation metrics

This section introduces evaluation metrics aimed at quantifying the learning dynamics and thus explains the existing fidelity paradox of Knowledge Distillation (KD) with teacher ensemble training, particularly when subject to varied data augmentation strengths.

4.1. IoU in attention maps

To elucidate divergent attentional patterns within teacher ensembles, without losing the generalizability, we examine their attention maps (Zhou et al., 2016) in the most representative model architectures, i.e., ResNet (He et al., 2016) and Transformer (Vaswani et al., 2017), during the training and validation stage. In practice, for ResNet teacher or student models, the features output from the penultimate convolution layer, followed by a new convolution layer constructed using the model's fully-connected layer weights, are selected to compute the attention maps. For the Transformers, the attention maps are obtained directly with their built-in attention modules. Subsequently, the Intersection over Union (IoU) (Rezatofighi et al., 2019) is computed between the attention maps of different teachers to measure their diversities. Take the 2-teacher ensemble KD as an example, for an image sample $\boldsymbol{S}$, to compute the IoU between the teacher models, two attention maps $A_{t1}, A_{t2} \subseteq \boldsymbol{S}$ are obtained associated with each teacher model, with the final metric value computed as in Eq. (1):

$$\text{IoU} = \frac{|A_{t1} \cap A_{t2}|}{|A_{t1} \cup A_{t2}|} \tag{1}$$

4.2. Model dependency in KD

We use *fidelity* metrics, namely the averaged predictive Kullback–Leibler (KL) divergence and top-1 agreement (Stanton et al., 2021), along with mutual information calculated between models' logits. This enables us to showcase the mimicry behavior and dependency between teachers and the student.

Given a classification task with input space $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ and label space $\mathcal{Y} = \{y_c\}_{c=1}^{C}$. Let $f : \mathcal{X} \to \mathbb{R}^C$ be a classifier whose outputs define a categorical predictive distribution over $\mathcal{Y}$, $\hat{p}(y_c|\boldsymbol{x}_i) = \sigma_c(\boldsymbol{z}_i)$, where $\sigma_c(\cdot)$ is the softmax function and $\boldsymbol{z}_i := f(\boldsymbol{x}_i)$ denotes the model *logits* when $\boldsymbol{x}_i$ is feed into f. The formal definition of KL divergence, top-1 agreement (Top-1 A), and mutual information (MI) are formulated as follows:

$$\text{KL}(P_t \| P_s) = \sum_{c=1}^{C} \hat{p}_t(y_c|\boldsymbol{x}) \log \frac{\hat{p}_t(y_c|\boldsymbol{x})}{\hat{p}_s(y_c|\boldsymbol{x})} \tag{2}$$

$$\text{Top-1 A} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\arg\max_c \sigma_c(\boldsymbol{z}^t) = \arg\max_c \sigma_c(\boldsymbol{z}^s)\} \tag{3}$$

$$\text{MI}(\mathcal{Y}^t; \mathcal{Y}^s) = \sum_{\boldsymbol{y}^t \in \mathcal{Y}^t} \sum_{\boldsymbol{y}^s \in \mathcal{Y}^s} P(\boldsymbol{y}^t, \boldsymbol{y}^s) \log \frac{P(\boldsymbol{y}^t, \boldsymbol{y}^s)}{P(\boldsymbol{y}^t)P(\boldsymbol{y}^s)} \tag{4}$$

where $P(\boldsymbol{y}^t, \boldsymbol{y}^s)$ is the joint probability distribution of the teacher and student. $P(\boldsymbol{y}^t)$ and $P(\boldsymbol{y}^s)$ represent the marginal probability distributions of the teacher and student. For metrics calculated between teacher ensemble and student, the logits or outputs of different teachers are first averaged and then computed with the student. This paper uses Top-1 A for fidelity measurement in the main text, and results with KL divergence can be found in Appendix A.2.

4.3. Quantify data augmentation strength within ensemble KD

In our experiments, we employ various data augmentation techniques on both teacher ensembles and the student model to modulate the level of randomness introduced into the dataset, as detailed in Section 5. To quantify the strength of these applied data augmentations and demonstrate their effects on KD, we leverage *Affinity* measurements (Cubuk et al., 2021), specifically adapted to our KD scenario:

$$\text{Affinity} = \frac{\text{Acc}(D'_{val})}{\text{Acc}(D_{val})} \tag{5}$$

where $\text{Acc}(D'_{val})$ denotes the validation accuracy of the student model trained with augmented distillation dataset and tested on the augmented validation set. $\text{Acc}(D_{val})$ represents the accuracy of the same model tested on clean validation set. It is worth noting that for a specific dataset, the augmented set D'_{val} is shared. And, to avoid introducing possible biases in metrics computation, each time the random seeds are altered for the augmentations.

This metric measures the offset in data distribution between the original one and the augmented one captured by the student model after KD training: Higher Affinity value corresponds to smaller offset between the data distributions. It is a generic metric that is not sensitive to different types of augmentation or the dataset in use. In this paper, Affinity is used as a tool to quantify and thus help on controlling the degree of randomness injected into the distillation dataset. This provides us with a systematic approach to analyze how data augmentation interacts with KD generalization, fidelity, and attention mechanisms. We anticipate that when the data augmentation strength of the student model aligns with that of the teacher model, the Affinity will be higher. And, lower Affinity corresponds to stronger data augmentation, leading to higher student accuracy and better generalization performance.

It is noteworthy that what we mean *low Affinity* is a "moderate low but cannot be as low as 0" notion: An Affinity of 0 presupposes a situation where the augmented data is so drastically different from the original that it no longer retains any of the original data's informative features, or the model has entirely failed to learn from the augmented data. Our claim that models with low Affinity can still exhibit good generalization performance is based on the understanding that these models, through diverse and challenging augmentations, learn to abstract and generalize from complex patterns. This does not necessarily imply that an Affinity of 0, resulting from complete misalignment with the augmented data, is desirable or indicative of strong generalization. Instead, we suggest that moderate to low Affinity, within a range that indicates the model has been challenged but still retains learning efficacy, can foster robustness and generalization. In contrast, an intermediate to high Affinity value is assumed to indicate greater alignment with the original dataset. This will be further clarified in Section 6.1.

5. Experimental setup

In our ensemble Knowledge Distillation (KD), experiments are conducted with two or three teachers. Each teacher model is a ResNet50 classifier pretrained on ImageNet (Deng et al., 2009) and then fine-tuned on their respective target datasets. The student model is ResNet18 trained from scratch using vanilla KD (Hinton et al., 2015). Take the ensemble KD with two teachers as an example, the loss function is defined as:

$$\mathcal{L}_{\mathrm{NLL}}(\boldsymbol{z}^s, \boldsymbol{y}^s) = -\sum_{c=1}^{C} y_c \log \sigma_c(\boldsymbol{z}^s) \tag{6}$$

$$\mathcal{L}_{\mathrm{KD1,2}}(\boldsymbol{z}^s, \boldsymbol{z}^{t1,2}) = -\tau^2 \sum_{c=1}^{C} \sigma_c(\frac{\boldsymbol{z}^{t1,2}}{\tau}) \log \sigma_c(\frac{\boldsymbol{z}^s}{\tau}) \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{NLL}} + \frac{1}{2}(\mathcal{L}_{\mathrm{KD1}} + \mathcal{L}_{\mathrm{KD2}}) \tag{8}$$

where $\mathcal{L}_{\mathrm{NLL}}$ is the usual supervised cross-entropy between the student logits $\boldsymbol{z}^s$ and the one-hot labels $\boldsymbol{y}^s$. $\mathcal{L}_{\mathrm{KD1,2}}$ is the added knowledge distillation term that encourages the student to match the teacher ensembles.

In this paper, we are focusing on ensemble KD with 2 teachers T_1 and T_2. Results with 3 teachers are discussed in Appendix A.5. We also provide experiments with Vision Transformers (ViTs) (Dosovitskiy et al., 2021) where the attention map can be obtained directly with the built-in attention module in Appendix A.7.

Experiments are conducted on well-recognized long-tailed datasets ImageNet-LT (Liu et al., 2019), CIFAR100 (Krizhevsky, 2009) with an imbalanced factor of 100, and their balanced counterparts. Hyperparameters remain consistent across experiments for each dataset. More detailed settings, including learning rates and temperatures, are provided in Appendix A.1.

In this paper, we distinguish between two types of data augmentation: (1) Weak data augmentation, encompassing conventional methods such as random resized crop, random horizontal flip, and color jitters. (2) Strong data augmentation includes RandAugment (RA) (Cubuk et al., 2020) applied to ImageNet-based datasets and AutoAugment (AA) (Cubuk et al., 2019) applied to CIFAR-based datasets. Further technical details on these augmentation methods can be found in Appendix A.1. For denotation purposes, we use $\mathrm{T}_s, \mathrm{S}_s$ to represent teacher or student models trained with strong augmentation, while $\mathrm{T}_w, \mathrm{S}_w$ denote those trained with weak augmentation.

It is essential to highlight that technically, the strong data augmentation applied to both teacher ensemble and student model in KD does not necessarily result in the highest data augmentation strength, as measured by our Affinity metric (defined in Eq. (5)). This will be shown and clarified further in Section 6.1 Table 1. Therefore, in this study, we varied the data augmentation strengths in ensemble KD. Specifically, in the series of experiments conducted on each dataset, we utilized the entire permutation set of $\mathrm{T}_w, \mathrm{T}_s, \mathrm{S}_w, \mathrm{S}_s$ to construct trials (for example, $\mathrm{T}_{1s}\mathrm{T}_{2w}\mathrm{S}_s$ is one trial denotation), and then computed their Affinity to quantify their data augmentation strength. In practice, for evaluation, we computed our metrics introduced in Section 4 on both the training set and validation set, considering each trial's corresponding data augmentation strength.

6. Results and analysis

Our comprehensive set of experiments has yielded several intriguing insights into the learning dynamics of Knowledge Distillation (KD) and explains the fidelity paradox through various data augmentation strengths. We particularly emphasize the roles of attention map diversity, model fidelity, and mutual information, as they interact with student performance in terms of top-1 accuracy and overfitting during both the training and validation procedures.

6.1. Impact on attention map diversity

Fig. 2 *Top* shows that during training, a consistent decrease is observed in the Intersection over Union (IoU) of attention maps between different teacher models with stronger data augmentation. This decrease is correlated with an increase in the student model's accuracy. Trial denotations are also marked as data labels in these scatter plots, together with Table 1 demonstrating their data augmentation strengths.

These Affinity values aid in understanding the data augmentation strengths and the decreasing tendencies in the scatter plots: Recall that Affinity measures the offset in data distribution between the original one and the one after data augmentation captured by the student, and lower Affinity corresponds to higher augmentation strength, leading to higher student accuracy. As evidence, for those trials with strong data augmentation and low Affinity, e.g., $\mathrm{T}_{1s}\mathrm{T}_{2w}\mathrm{S}_s$ in CIFAR-100, $\mathrm{T}_{1w}\mathrm{T}_{2s}\mathrm{S}_s$ in CIFAR-100 imb100, $\mathrm{T}_{1s}\mathrm{T}_{2w}\mathrm{S}_s$ in ImageNet, and $\mathrm{T}_{1s}\mathrm{T}_{2w}\mathrm{S}_s$ in ImageNet-LT, a relatively high validation accuracy is observed for each dataset. It is important to emphasize that the application of strong data augmentation to both teacher ensemble and student model in KD does not lead to the highest level of data augmentation strength, as quantified by our Affinity metric defined in Eq. (5). That is, it is the diversity of teachers' augmentation strength but not the strong data augmentation for a single teacher or student model matters: $\mathrm{T}_{1s}\mathrm{T}_{2w}\mathrm{S}_s$ is stronger than $\mathrm{T}_{1s}\mathrm{T}_{2s}\mathrm{S}_s$. Appendix A.4 also offers scatter plots of IoU between T_1 and T_2 attention maps versus Affinity during KD training.

Significantly, this observation suggests that as the ensemble of teachers focuses on increasingly diverse aspects of the input data, the student model benefits from a richer, more varied set of learned representations, leading to enhanced performance, as visualized in Fig. 2 *Bottom*. This finding aligns with and extends ensemble learning theories in KD, where diversity among models enhances overall student performance even by simply manipulating the data augmentation strength. It introduces a new dimension to Knowledge Distillation theory, emphasizing the value of diverse learning stimuli.

Section 7 also compares our method with the SOTA baselines, which suggests that our approach, achieved solely by injecting varied levels of randomness into the dataset through modulated data augmentation strength, can attain comparable student performance on both balanced and imbalanced datasets with SOTA methods featuring intricate designs on architectures, optimization, or distillation procedures. Additionally, to demonstrate the effectiveness of the proposed data augmentation trials in ensemble KD, Section 8 presents an ablation study. This includes results from directly training student models with varying augmentation strengths, as well as KD training results without any data augmentation as a control.

6.2. Revisiting the role of fidelity and mutual information

As in Fig. 3, during training, we observed a decrease in both fidelity and mutual information between teacher ensembles and the student model with stronger data augmentation. Intriguingly, this decrease was accompanied by improved validation accuracy in the student model. This indicates that a lower level of direct mimicry, in terms of output logits distribution, between teacher ensembles and the student is conducive to more effective learning in KD, possibly due to student learning from more divergent teachers' attentions.

To further demonstrate the causality between teachers' attention divergence and low student-teacher fidelity, i.e., a more diverse attention maps within teacher ensemble causes a lower fidelity, an A/B test is conducted in the setup of ensemble KD with two teachers. Specifically, the control group is the vanilla KD (denoted as vKD) with different data augmentation strengths we used in all previous experiments, and the experimental group (denoted as hKD) is designed as follows: Each training image is first cropped into two parts, left and right, as input to teacher model T_1 and T_2 respectively. This allows us to

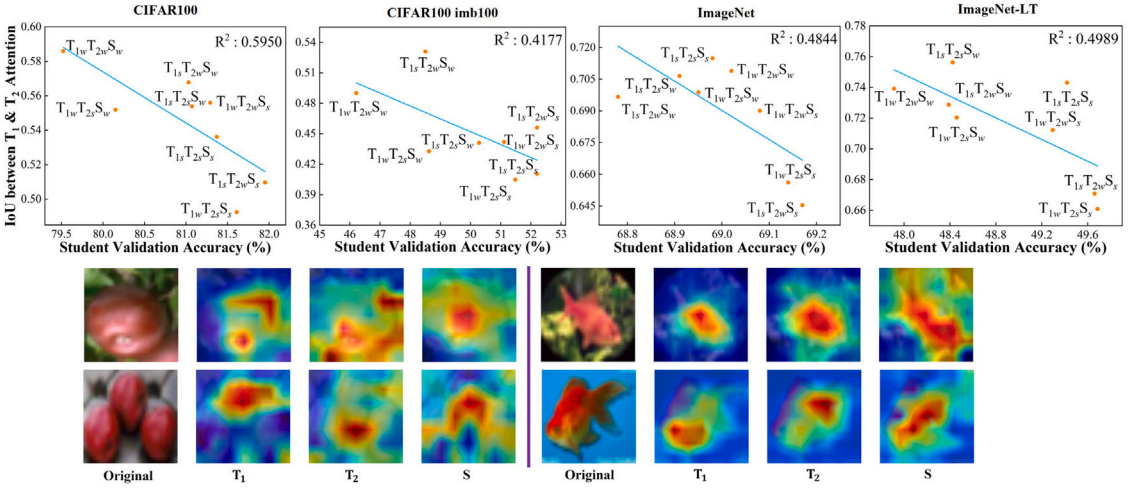


Fig. 2. For ResNet models. *Top*: Scatter plots of IoU between T_1 and T_2 attention maps during KD training. *Bottom*: Exampled attention maps of T_1, T_2 and S. This attention divergence among teacher ensembles, attributed to the randomness injected by data augmentation, gives the student distilled on them a more comprehensive perspective.

Table 1
Affinity, and Validation Accuracy (Val-Acc) of models with various data augmentation strengths.

Dataset	Metric	Model							
		$T_{1w}T_{2w}S_w$	$T_{1w}T_{2w}S_s$	$T_{1s}T_{2w}S_w$	$T_{1s}T_{2w}S_s$	$T_{1w}T_{2s}S_w$	$T_{1w}T_{2s}S_s$	$T_{1s}T_{2s}S_w$	$T_{1s}T_{2s}S_s$
Cifar100	Affinity	0.9807	**0.8611**	0.9805	0.9083	0.9858	0.9143	0.9729	0.9310
	Val-Acc	0.7952	0.8129	0.8103	**0.8195**	0.8015	0.8161	0.8107	0.8137
Cifar100 imb100	Affinity	0.9763	**0.8132**	0.9810	0.8637	0.9751	0.8635	0.9723	0.8955
	Val-Acc	0.4621	0.5111	0.4850	**0.5220**	0.4862	0.5148	0.5028	0.5210
ImageNet	Affinity	0.9901	**0.8767**	0.9930	0.8988	0.9845	0.9131	0.9871	0.9122
	Val-Acc	0.6902	0.6908	0.6878	**0.6917**	0.6895	0.6914	0.6891	0.6898
ImageNet long-tail	Affinity	0.9850	**0.8311**	0.9755	0.8704	0.9782	0.8751	0.9903	0.8971
	Val-Acc	0.4791	0.4929	0.4839	0.4966	0.4846	**0.4968**	0.4842	0.4942

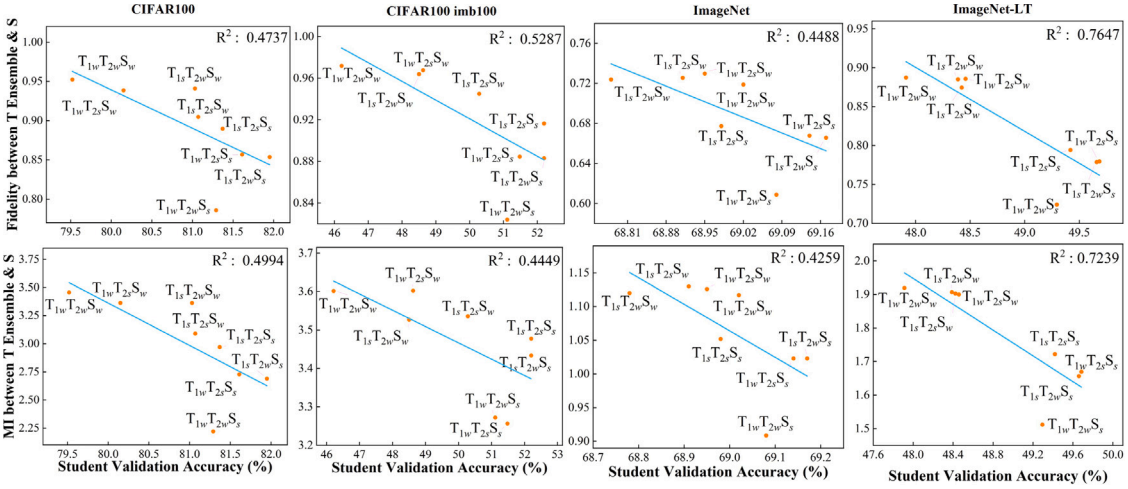


Fig. 3. For ResNet models. Scatter plots of *Top*: Fidelity (measured by top-1 A) and *Bottom*: Mutual Information (MI) between teacher ensembles and student during KD training. These decreasing tendencies along with the improved student validation accuracy are in contrast to the traditional viewpoint that higher fidelity consistently benefits student performance, indicating that some extent of student independency may be desired during KD training.

proactively diversify the attention maps of each teacher model, rather than passively altering it in the case of varying data augmentation strengths. Then in average, we can expect the experimental group to have far less attention IoU values than the control group, while keeping comparable generalization performance, because in the former each teacher's attention is constrained to one half of each image. The null hypothesis H_0 is that from control (vKD) to experimental (hKD) group, as the teachers' attention maps IoU decrease, an increase in student-teacher fidelity is observed. Denoting the total number of trials as Num, the corresponding p-value is calculated as:

$$p\text{-value} = \frac{\#|\text{fidelity(hKD)} > \text{fidelity(vKD)}|}{Num} \tag{9}$$

Experiments reveal a p-value less than 0.05, suggesting that we should reject this null hypothesis. Detailed experimental results are provided in Appendix A.3. In summary, more divergent teacher attentions (i.e., lower IoU values) does cause the decrease in student-teacher fidelity.

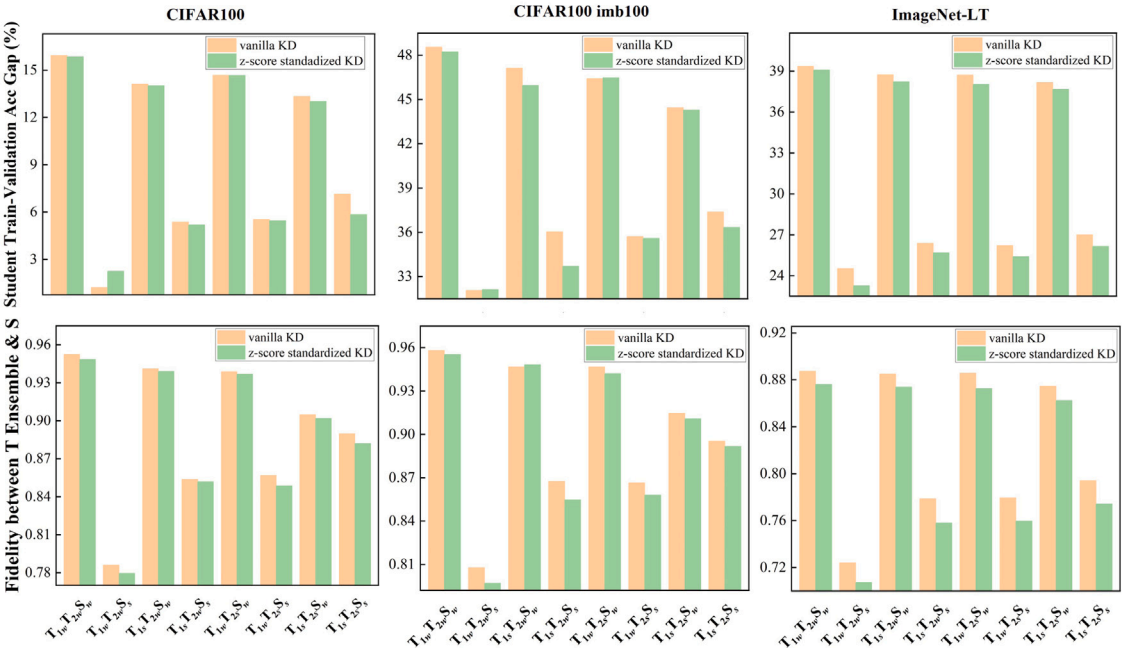


Fig. 4. For ResNet models. Bar plots comparing between vanilla KD and z-score standardization KD. *Top*: Generalization performance in terms of train-validation accuracy gap. *bottom*: Student–teacher fidelity. The z-score standardization, aimed at facilitating the student–teacher logits matching procedure, does improve student generalization performance (indicated by a lower accuracy gap) in most cases. However, it also leads to a decrease in student–teacher fidelity during training, suggesting that the benefit lies more in student generalization than in fidelity improvement.

This counterintuitive result aligns with and complements the paradoxical observation in Stanton et al. (2021). It implies that while the student model develops a certain level of independence from the teachers (evidenced by lower fidelity and mutual information), it still effectively captures and generalizes the core knowledge of the teachers. Combining with the observation on how varying data augmentation strengths influence the teachers' attention divergence in Section 6.1, we highlight attention diversification in teacher ensembles as a deeper reason why a student with good generalization may be unable to match the teacher during KD training: Stronger data augmentation increases attention divergence, enabling teachers to offer a broader perspective to the student. Consequently, the student surpasses the knowledge of a single teacher, becoming more independent, and the observed low-fidelity is a demonstration of this phenomenon rather than a pathology.

6.3. Effects of logits matching optimization on KD

Although Stanton et al. (2021) has shown the phenomenon of low-fidelity, they attributed the challenges in optimization as the key factor for the student's inability to match the teacher. Recent studies, such as Sun et al. (2024), continue to focus on optimizing the student-teacher logits matching process. Yet in Section 3 the 3rd hypothesis, we suggested that the optimization towards increasing student-teacher mimicry behavior in fact benefits generalization performance rather than the fidelity.

To illustrate, here we compared the aforementioned vanilla KD with a logits-matching optimization method in KD (Sun et al., 2024) under different data augmentation strengths, for dataset CIFAR100, CIFAR100-imb100, and ImageNet-LT. Specifically, we experiment with a z-score standardization method applied on logits before the softmax. This mitigates the logits magnitudes and variance gap between teacher and student, which facilitates the student-teacher emulation procedure.

Theoretically, denote the logits of teacher model and student model as $\boldsymbol{z}^t$ and $\boldsymbol{z}^s$ respectively, and the softmax function as $\sigma(\cdot)$. Then for a finally well-distilled student with predicted probability density perfectly matching the teacher, i.e., $\sigma(\boldsymbol{z}^s) = \sigma(\boldsymbol{z}^t)$, we have the following two properties proved in (Sun et al., 2024):

$$\text{Logit shift: } \boldsymbol{z}^s = \boldsymbol{z}^t + \Delta \tag{10}$$

$$\text{Variance match: } \frac{\text{Var}(\boldsymbol{z}^s)}{\text{Var}(\boldsymbol{z}^t)} = \frac{\tau_s}{\tau_t} \tag{11}$$

where Δ can be considered constant for each sample image, and τ_s, τ_t are temperatures for the student and teacher respectively during training. That is, even for the student with highest fidelity to its teacher such that $\sigma_c(\boldsymbol{z}^s) = \sigma_c(\boldsymbol{z}^t)$ for any class c in the dataset, still we have $\boldsymbol{z}^s = \sqrt{\frac{\tau_s}{\tau_t}} \cdot \boldsymbol{z}^t + \Delta$ which means the student logits cannot match the teacher logtis. A z-score normalization applied on both the student and teacher logits during KD training can soothe this mismatch by making their logtis distribution equal mean and variance, and thus improve generalization performance. However, from the fidelity definition in Eq. (3), since the softmax function is monotonic, what we are looking for is the agreed index c of maximum logits between the teacher and student $\arg\max_c(\boldsymbol{z}^t) = \arg\max_c(\boldsymbol{z}^s)$, which unfortunately cannot be directly affected by such optimization method.

In conclusion, though an optimization towards student-teacher logits matching can relieve the logit shift and variance match problem, in reality its benefit lies in the student generalization rather than the fidelity improvement. As shown in Fig. 4, the z-score standardization does improve the student train-validation accuracy gap in most cases, but a decrease in the student-teacher fidelity is still witnessed.

7. Quantitative evaluation

Table 2 compares our method with SOTA KD baselines: LFME (Xiang & Ding, 2020), DMAE (Bai et al., 2023), FFKD (Gou et al., 2024), and z-score logit standardization (Sun et al., 2024), focusing on the top-1 validation accuracy. LFME is specifically tailored for long-tailed datasets, so we only present its results on those. DMAE, originally designed for balanced datasets, performs less effectively on long-tailed ones. The z-score logit standardization process is introduced to alleviate the challenges of logits matching due to logit shifts and variance

Table 2
Validation accuracies for our method, LFME, and DMAE on four data sets.

Method	Cifar100	ImageNet	Cifar100 imb100	ImageNet long-tail
LFME	–	–	0.4380	0.3880
DMAE	**0.8820**	**0.8198**	0.3725	0.4395
FFKD	0.7980	0.7217	0.5001	0.4684
Z-Score	0.8044	0.6781	0.5219	0.4805
Ours(1T)	0.8133	0.6802	0.5152	0.4910
Ours(2T)	0.8195	0.6917	0.5220	**0.4968**
Ours(3T)	0.8204	0.7032	**0.5302**	0.4965

mismatches between teacher and student, and it ultimately reduces the overfitting and enhances the student's generalization ability. FFKD is a two-stage forward and feedback KD method. In the feedback stage, the teacher evaluates the student model's knowledge mastery, allowing the teacher to adjust its teaching strategy to enhance overall performance.

For our method shown in this table: Ours(1T) is refferred to the KD with one ResNet50 teacher model distilled to one ResNet18 student model, with T_wS_s. Ours(2T) is refferred to the KD with two ResNet50 teacher models distilled to one ResNet18 student model, with $T_{1s}T_{2w}S_s$. Ours(3T) is refferred to the KD with three ResNet50 teacher models distilled to one ResNet18 student model, with $T_{1s}T_{2w}T_{3w}S_s$. For the z-score standardization KD method shown in this table, the same models as in Ours (2T) are used.

This table demonstrates that our approach, achieved solely by injecting varied levels of randomness into the dataset through controlled data augmentation strength, can attain comparable student performance on both balanced and imbalanced datasets with methods featuring intricate designs on architectures, optimization, or distillation procedures.

8. Ablation study

Our augmented KD approach focuses on two key components: (1) an ensemble KD framework with multiple teachers, and (2) the use of varying data augmentation strengths across models. We conducted ablation studies to assess the impact of these components on generalization ability, fidelity, attention mechanisms, and student validation performance. Specifically, we compared the results of our ensemble KD with two teachers, T_1 and T_2, as previously described, against KD training without data augmentation (denoted as S_n and T_n for student and teacher, respectively) and direct student training with different augmentation strengths (denoted as S_w and S_s for weak and strong data augmentation, respectively). Additionally, we discuss the results of ensemble KD with three teachers in Appendix A.5 as a supplementary analysis.

As shown in Table 3, across all datasets, our ensemble KD framework under the data augmentation trial with a low Affinity value (i.e., $T_{1s}T_{2w}S_s$) achieves the highest validation accuracy, coupled with the lowest fidelity, mutual information, T_1-T_2 attention maps IoU, and the second-lowest train-val accuracy gap. These results not only align with our previous analysis but also underscore the effectiveness of our approach in enhancing supervised image classification tasks.

Moreover, regarding the generalization ability indicated by the train-val accuracy gap, we observe that for a specific dataset, as we progress from S_n to $T_{1n}T_{2n}S_n$, and then from S_w to S_s, ultimately reaching $T_{1s}T_{2w}S_s$, an increase in data augmentation strength correlates with decreasing fidelity and a diminishing accuracy gap. This demonstrates that the relationship between low fidelity and enhanced generalization ability is linked to stronger data augmentations, such as S_s or $T_{1s}T_{2w}S_s$.

9. Generalization to other network architectures

To demonstrate the generalizability of our methods and conclusions to other model architectures, we present experimental results using the CIFAR-100 dataset with VGG network architectures. Specifically, the teachers are VGG19 models, while the students are VGG16 models. All other settings remain consistent with those used in the ResNet experiments.

Fig. 5 presents the results from distilling two VGG19 teachers onto one VGG16 student for the CIFAR-100 dataset. The decreasing trends in the top row and the increasing trends in the bottom row confirm that the conclusions drawn from the ResNet cases also apply to these VGG trials. Specifically, greater teacher attention diversity correlates with higher student validation accuracy, as do lower student-teacher fidelity and mutual information. Additionally, increased attention diversity, reduced fidelity, and lower mutual information also correspond to lower Affinity, which indicates stronger data augmentation.

For visualization and comparison, Table 4 further summarizes the results of train-val accuracy gap, as an indicator to the model generalization ability, alongside student-teacher fidelity for both ResNet and VGG architectures, under various data augmentation strengths on the CIFAR-100 dataset. As shown, for a specific data augmentation strength (i.e., within each column), a consistent decrease in both fidelity and Acc-Gap is observed when moving from ResNet to VGG, indicating that low fidelity and better generalization ability correlate with a simpler model architecture.

Table 3
Ablation study: Comparison of our two-teacher ensemble KD against KD training without any data augmentation (denoted as T_n and S_n), and against the direct student training with different augmentation strengths. Evaluation metrics involve the validation accuracy (Val-Acc), the train-val accuracy gap (Acc-Gap) as an indicator to the model generalization ability, fidelity between teacher ensemble and student (Fidelity), mutual information between teacher ensemble and student (MI), and IoU between T_1 and T_2 attention maps (IoU).

Dataset	Metric	Model				
		S_n	S_w	S_s	$T_{1n}T_{2n}S_n$	$T_{1s}T_{2w}S_s$
Cifar100	Val-Acc	0.640	0.714	0.726	0.680	**0.819**
	Acc-GAP	0.360	0.126	**−0.014**	0.320	0.054
	Fidelity	N/A	N/A	N/A	0.999	**0.854**
	MI	N/A	N/A	N/A	2.779	**2.690**
	IoU	N/A	N/A	N/A	0.601	**0.510**
Cifar100 imb100	Val-Acc	0.311	0.343	0.340	0.351	**0.522**
	Acc-GAP	0.689	0.391	**0.221**	0.649	0.360
	Fidelity	N/A	N/A	N/A	1.000	**0.868**
	MI	N/A	N/A	N/A	**1.462**	2.900
	IoU	N/A	N/A	N/A	0.491	**0.456**
ImageNet	Val-Acc	0.516	0.536	0.545	0.568	**0.692**
	Acc-GAP	0.166	−0.101	**−0.011**	0.426	−0.093
	Fidelity	N/A	N/A	N/A	0.996	**0.666**
	MI	N/A	N/A	N/A	1.610	**1.023**
	IoU	N/A	N/A	N/A	0.746	**0.557**
ImageNet long-tail	Val-Acc	0.245	0.372	0.388	0.312	**0.497**
	Acc-GAP	0.754	0.363	**0.184**	0.688	0.264
	Fidelity	N/A	N/A	N/A	0.999	**0.779**
	MI	N/A	N/A	N/A	2.180	**1.657**
	IoU	N/A	N/A	N/A	0.793	**0.671**

10. Discussion

In this study on ensemble KD, we explore the underlying mechanisms of how teacher models' attentions interact with student model performance and influence student–teacher fidelity. We also propose a simple yet effective teacher-ensemble KD framework that achieves performance comparable to SOTA methods. However, beyond our findings, there is still room for further exploration: our research is limited to data augmentation techniques that control teachers' attention divergence. While simple, this approach may not fully capitalize on the potential benefits of attention divergence and low-fidelity characteristics.

To address this issue, there are two potential paths: one is to construct frameworks that specifically diversify teachers' attentions in ensemble KD or combine different types of attention at various

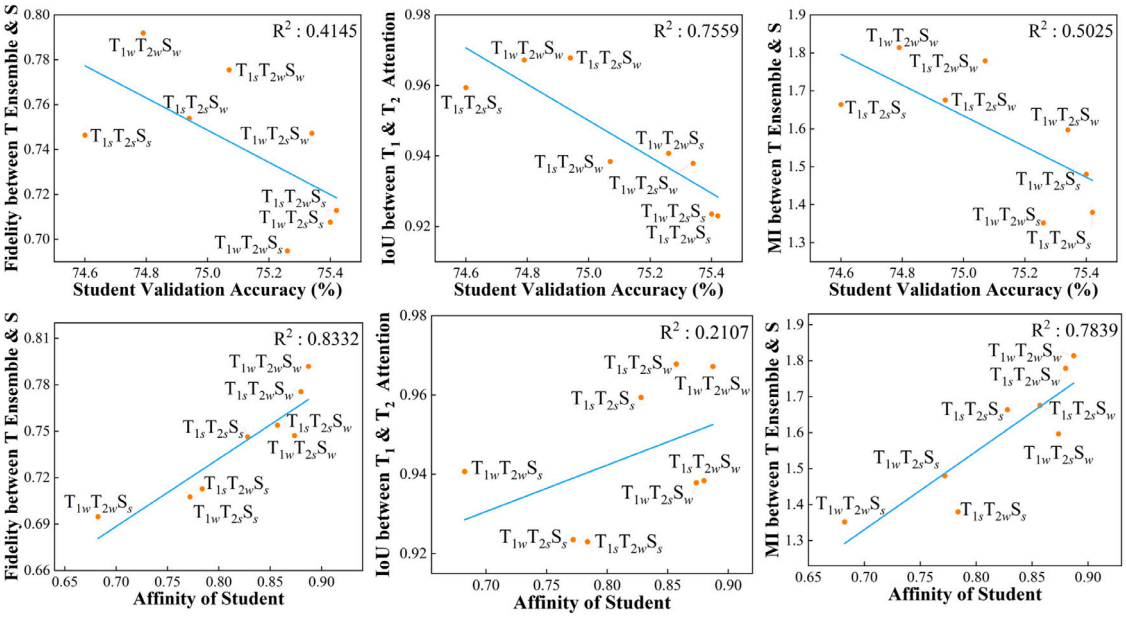


Fig. 5. For VGG models on CIFAR-100 dataset. Scatter plots of Fidelity (measured by top-1 A), IoU between T_1 and T_2 attention maps, and Mutual Information (MI) between teacher ensembles and student during KD training. *Top*: Versus student validation accuracy. *Bottom*: Versus Affinity. The decreasing tendencies in the top row, and the increasing trends in the bottom row, demonstrate that the conclusions drawn from the previous ResNet cases still holds for these VGG trials.

Table 4
Train-val accuracy gap (Acc-Gap) and student-teacher fidelity for different model architectures under various data augmentation strengths, on CIFAR-100 dataset.

Arch	Metric	Model							
		$T_{1w}T_{2w}S_w$	$T_{1w}T_{2w}S_s$	$T_{1s}T_{2w}S_w$	$T_{1s}T_{2w}S_s$	$T_{1w}T_{2s}S_w$	$T_{1w}T_{2s}S_s$	$T_{1s}T_{2s}S_w$	$T_{1s}T_{2s}S_s$
ResNet	Acc-Gap	0.1593	0.0122	0.1411	0.0537	0.1468	0.0553	0.1333	0.0714
	Fidelity	0.9523	0.7859	0.9411	0.8536	0.9387	0.8568	0.9048	0.8897
VGG	Acc-Gap	−0.0093	−0.2008	−0.0195	−0.1162	−0.0204	−0.1201	−0.0501	−0.0719
	Fidelity	0.7919	0.6948	0.7755	0.7128	0.7472	0.7676	0.7539	0.7464

levels (Gou et al., 2023). The other, which has been less explored, is to introduce external guidance or multi-modal features to enrich the knowledge derived from diverse attentions. For instance, one could leverage CLIP (Radford et al., 2021) features for KD, using CLIP image features for one teacher and CLIP text features for another. The challenge with this approach lies in effectively combining and distilling knowledge from the two teachers, as well as optimizing the CLIP model during KD training to enhance feature representations. This will be the focus of our future work.

11. Conclusion

Our research, aiming to explain the fidelity paradox, intersects with and expands upon existing theories for ensemble Knowledge Distillation (KD) in several ways. **(1)** It introduces a novel perspective on the learning and knowledge transfer process by investigating the impact of attention map diversity on fidelity in KD with various data augmentation strength. **(2)** It reevaluates the student-teacher fidelity and mutual information challenge, providing insights into the ongoing debates about the relation between student's ability to mimic its teachers and its generalization performance in KD. **(3)** It highlights that for optimization towards facilitating student-teacher logits matching which relieves the logit shift and variance match problem, its benefit lies in the student generalization rather than the fidelity improvement. These insights have the potential to catalyze further theoretical advancements in the pursuit of robust KD.

CRediT authorship contribution statement

Chenqi Guo: Conceptualization, Investigation, Software, Supervision, Writing – original draft, Funding acquisition. **Shiwei Zhong:** Methodology, Investigation, Software, Data curation, Validation, Writing – review & editing. **Xiaofeng Liu:** Methodology, Investigation, Validation, Writing – review & editing. **Qianli Feng:** Methodology, Writing – review & editing. **Yinglong Ma:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The corresponding author previously worked at the Ohio State University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China (No. JB2024020).

Appendix

A.1. Detailed experimental settings

The experiments are run on a GPU machine with RTX 4090 GPU, AMD 5995WX CPU and 128 GB memory. In each trial, the teacher model of ResNet50 is trained for 30 epochs for ImageNet-LT dataset, and 60 epochs for all the others. The student model of ResNet18 is distilled for: 200 epochs for CIFAR-100; 175 epochs for CIFAR-100 imb100; 60 epochs for ImageNet; and 165 epochs for ImageNet-LT dataset, when their validation accuracy converges.

In the main text, we apply strong data augmentations using RandAugment (RA) for the ImageNet-based datasets and AutoAugment (AA) for the CIFAR-based datasets. We selected these methods because, among the various augmentation techniques and policies we tested, AA and RA yielded the best validation accuracy. Below are some technical details regarding the implementation of these methods in our work:

(1) For weak data augmentation, we employ a transformation sequence that includes random resized cropping, random horizontal flipping, random rotation, and color jittering.
(2) For strong data augmentation on the CIFAR-100 and CIFAR-100 imb100 datasets, we utilize a transformation sequence that includes random resized cropping and random horizontal flipping, followed by the AA policy and the image cutout technique. Notably, within the AA policy, one of 25 sub-policies is randomly selected each time, with each sub-policy consisting of a combination of two transformations such as image shearing, translation, rotation, contrast adjustment, inversion, sharpness enhancement, brightness adjustment, color enhancement, posterization, and solarization. The image cutout technique improves model robustness by randomly selecting rectangular regions in an image and setting their pixel values to zero.
(3) For strong data augmentation on the ImageNet and ImageNet-LT datasets, we utilize a transformation sequence that includes random resized cropping, random horizontal flipping, and color jittering, followed by the RA policy. In the RA policy, a random number of transformation operations (up to 2) are selected from a comprehensive list, which includes image shearing, translation, rotation, contrast adjustment, inversion, sharpening, brightness adjustment, color enhancement, posterization, and solarization. These operations are executed sequentially, each with a randomly determined magnitude (up to 10).

Hyper-parameters, including temperatures of $\tau = 10$, hard label weight of $\alpha = 0.2$, initial learning rate of 0.1, momentum of 0.9, and batch size of 128, remain the same throughout the entire procedure in each case, ensuring consistent and reliable results for evaluation.

For training with balanced ImageNet dataset, we use a cosine annealing learning rate scheduler, with $T_{\max} = 30, \text{eta}_{\min} = 0$ for teacher training, and $T_{\max} = 60, \text{eta}_{\min} = 0$ for student distillation. For other datasets, a lambda learning rate scheduler is used. Specifically, during teacher training, with the following hyperparameters: $\text{step}_1 = 25, \text{step}_2 = 40, \text{step}_3 = 60$ for CIFAR-100; $\text{step}_1 = 25, \text{step}_2 = 40, \text{step}_3 = 60$ for CIFAR-100 imb100; and $\text{step}_1 = 35, \text{step}_2 = 50$ for ImageNet-LT. During student distillation, with the following hyperparameters: $\text{step}_1 = 190, \text{step}_2 = 195$ for CIFAR-100; $\text{step}_1 = 160, \text{step}_2 = 165, \text{step}_3 = 170$ for CIFAR-100 imb100; and $\text{step}_1 = 150, \text{step}_2 = 155, \text{step}_3 = 160$ for ImageNet-LT.

A.2. Fidelity with KL divergence measurement

In the main text Section 6.2, Top-1 A is used for the fidelity metric. Here we also provide results with Kullback–Leibler (KL) divergence between teacher ensembles and student during KD training, as in Fig. A.1. Note that for KL divergence, a higher value implies lower fidelity.

Table A.1
Results for the A/B test on CIFAR100 dataset.

Model	Acc Gap		IoU		Fidelity	
	vKD	hKD	vKD	hKD	vKD	hKD
$T_{1w}T_{2w}S_w$	0.1593	0.1631	0.5860	0.3188	0.9523	0.7564
$T_{1w}T_{2w}S_s$	0.0122	0.0171	0.5560	0.3062	0.7859	0.5921
$T_{1s}T_{2w}S_w$	0.1411	0.1560	0.5678	0.3033	0.9411	0.7295
$T_{1s}T_{2w}S_s$	0.0537	0.0654	0.5097	0.2970	0.8536	0.6520
$T_{1w}T_{2s}S_w$	0.1468	0.1784	0.5519	0.2619	0.9387	0.7248
$T_{1w}T_{2s}S_s$	0.0553	0.0759	0.4925	0.2549	0.8568	0.6513
$T_{1s}T_{2s}S_w$	0.1333	0.1541	0.5539	0.2738	0.9048	0.6621
$T_{1s}T_{2s}S_s$	0.0714	0.0657	0.5361	0.2747	0.8897	0.6801

Table A.2
Results for the A/B test on CIFAR100 IMB100 dataset.

Model	Acc Gap		IoU		Fidelity	
	vKD	hKD	vKD	hKD	vKD	hKD
$T_{1w}T_{2w}S_w$	0.4854	0.4836	0.4900	0.3195	0.9580	0.7114
$T_{1w}T_{2w}S_s$	0.3206	0.3742	0.4419	0.3094	0.8078	0.5406
$T_{1s}T_{2w}S_w$	0.4712	0.4995	0.5309	0.3041	0.9467	0.6892
$T_{1s}T_{2w}S_s$	0.3604	0.3994	0.4560	0.2992	0.8675	0.6040
$T_{1w}T_{2s}S_w$	0.4641	0.4860	0.4329	0.2643	0.9467	0.6892
$T_{1w}T_{2s}S_s$	0.3570	0.3827	0.4084	0.2558	0.8664	0.5997
$T_{1s}T_{2s}S_w$	0.4444	0.4790	0.4410	0.2717	0.9145	0.6192
$T_{1s}T_{2s}S_s$	0.3738	0.3225	0.4107	0.2721	0.8953	0.6242

Table A.3
Results for the A/B test on ImageNet Long-tail dataset.

Model	Acc Gap		IoU		Fidelity	
	vKD	hKD	vKD	hKD	vKD	hKD
$T_{1w}T_{2w}S_w$	0.3937	0.4104	0.7391	0.6245	0.8873	0.5657
$T_{1w}T_{2w}S_s$	0.2453	0.2426	0.7122	0.6311	0.7240	0.4542
$T_{1s}T_{2w}S_w$	0.3873	0.4152	0.7287	0.5948	0.8850	0.5554
$T_{1s}T_{2w}S_s$	0.2639	0.2713	0.6708	0.5607	0.7786	0.4901
$T_{1w}T_{2s}S_w$	0.3871	0.4161	0.7204	0.5798	0.8856	0.5559
$T_{1w}T_{2s}S_s$	0.2622	0.2680	0.6608	0.5537	0.7795	0.4916
$T_{1s}T_{2s}S_w$	0.3816	0.4133	0.7563	0.6244	0.8745	0.5308
$T_{1s}T_{2s}S_s$	0.2700	0.2663	0.7431	0.6490	0.7941	0.5138

A.3. In-depth results for the A/B test

In the main text, to demonstrate the causality between teachers' attention divergence and low student-teacher fidelity, an A/B test is conducted for ensemble KD with two teachers. Experiments reveal a p-value less than 0.05, suggesting that more divergent teacher attentions (i.e., lower IoU values) does cause the decrease in student-teacher fidelity. In this section, we further provides the detailed experimental results of the A/B test, as shown in Tables A.1–A.3. Here, vKD denotes the control group of vanilla KD experiments, and hKD denotes the control group of half-image inputs experiments. From these results, it can be seen that in average, hKD has far less attention IoU values than vKD, while keeping comparable generalization performance (indicated by a lower accuracy gap).

A.4. Evaluation metrics versus affinity

In the main text, we show that during training, a consistent decrease is observed in the student-teacher fidelity, mutual information (MI), and Intersection over Union (IoU) of attention maps between different teacher models, which correlates with higher student validation accuracy. Here, we also provide scatter plots illustrating the relationship between fidelity, MI, and IoU of attention maps between T_1 and T_2 versus Affinity during KD training, as shown in Fig. A.2. These increasing trends demonstrate that stronger data augmentation (reflected by smaller Affinity) is associated with lower fidelity, lower MI, and greater divergence in teacher attentions (indicated by lower IoU).

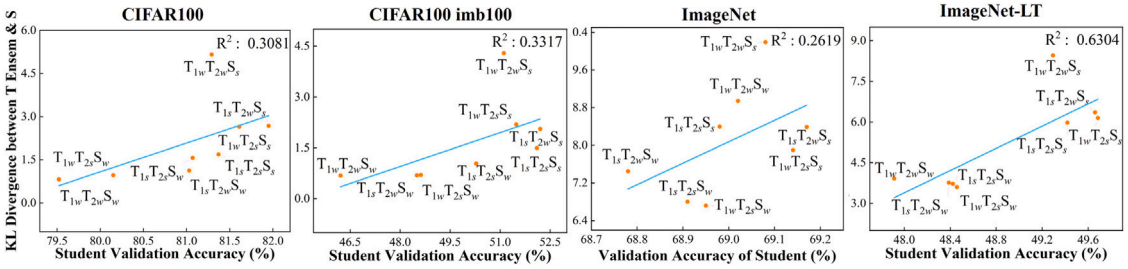


Fig. A.1. For ResNet models. Scatter plots of fidelity (measured by KL divergence) between teacher ensembles and student during KD training. For KL divergence, a higher value implies lower fidelity. Thus, these increasing tendencies align with the decreasing ones with Top-1 A in the main text.

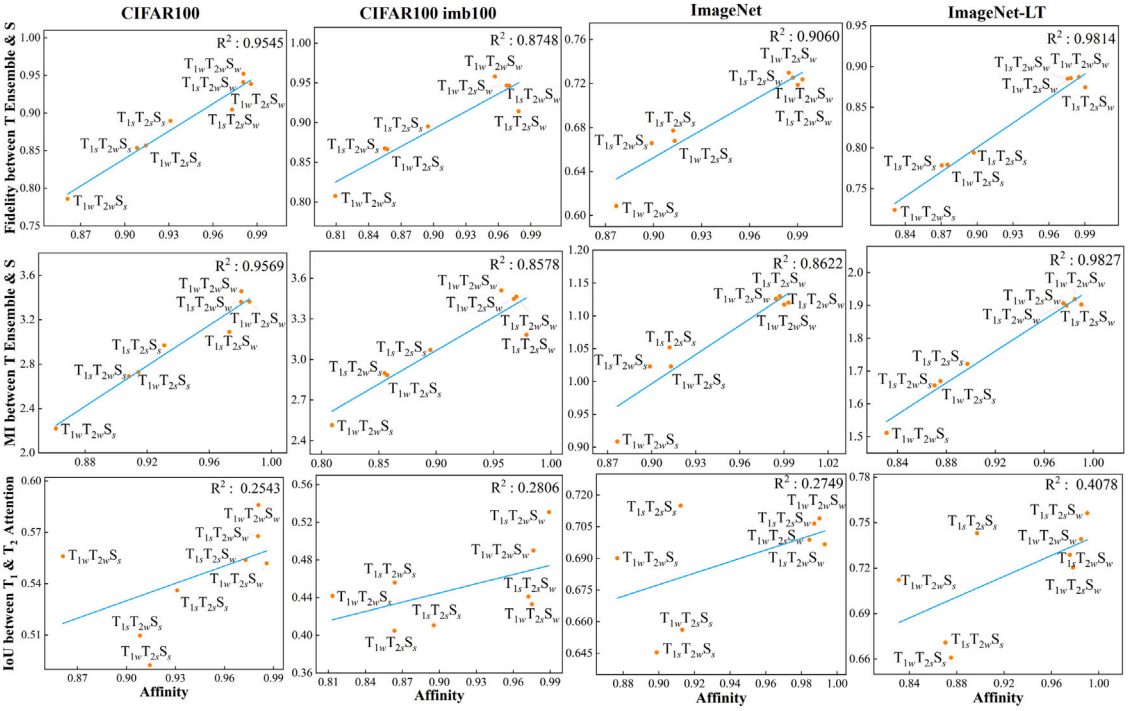


Fig. A.2. For ResNet models. Scatter plots of student-teacher fidelity, mutual information (MI), and IoU between T_1 and T_2 attention maps, versus Affinity during KD training. These increasing tendencies demonstrate that stronger data augmentation (reflected by smaller Affinity) is associated with lower fidelity, lower MI, and greater divergence in teacher attentions (indicated by lower IoU).

A.5. Results with more teacher numbers in ensemble knowledge distillation

In the main text, we focused on Knowledge Distillation (KD) with 2 teachers in the ensemble. Results with 3 teachers are discussed here. Fig. A.3 provides scatter plots of teacher attention IoU, fidelity, mutual information, and student entropy in 3-teacher ensemble KD cases, for CIFAR100 and CIFAR100 imb100 datasets. These plots align with the tendencies observed in 2-teacher cases in the main text.

A.6. Model calibration and overfitting effects in our experiments

As a supplementary study, in this section we further investigate the model calibration effects in ensemble KD. Empirically, the student model can be better calibrated by simply enhancing data augmentation strength. And, as the augmentation strength (measured by Affinity) and/or teacher numbers increased, the calibration effects become more pronounced.

While Guo et al. (2017) has revealed the calibration effects of temperature scaling, a common technique in KD that does not influence the student's accuracy, the impact of data augmentation on the student's prediction confidence and model calibration in KD remains unexplored. This impact is typically gauged by entropy and Expected Calibration Error (ECE) in predictions and is crucial in understanding how they relate to the student's ability to generalize and perform on unseen data, as measured by overfitting tendencies. Our hypothesis is that, beyond the inherent calibration effects of KD, the student model can be effectively calibrated by elevating data augmentation strengths as well.

In this study, we leverage logits entropy and Expected Calibration Error (ECE), along with calibration reliability diagrams (Guo et al., 2017) for visualization, to assess the calibration properties for teachers and student under varied data augmentation strengths. Specifically, the model logits entropy is computed as:

$$H(\boldsymbol{x}) = -\sum_{c=1}^{C} \hat{p}(y_c|\boldsymbol{x}) \log \hat{p}(y_c|\boldsymbol{x}) \tag{12}$$

For ECE calculation, we first group all the validation samples into M interval bins, which are defined based on the prediction confidence of the model for each sample. The ECE thus can be formulated as follows:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{Acc}(B_m) - \mathrm{Conf}(B_m)\right| \tag{13}$$

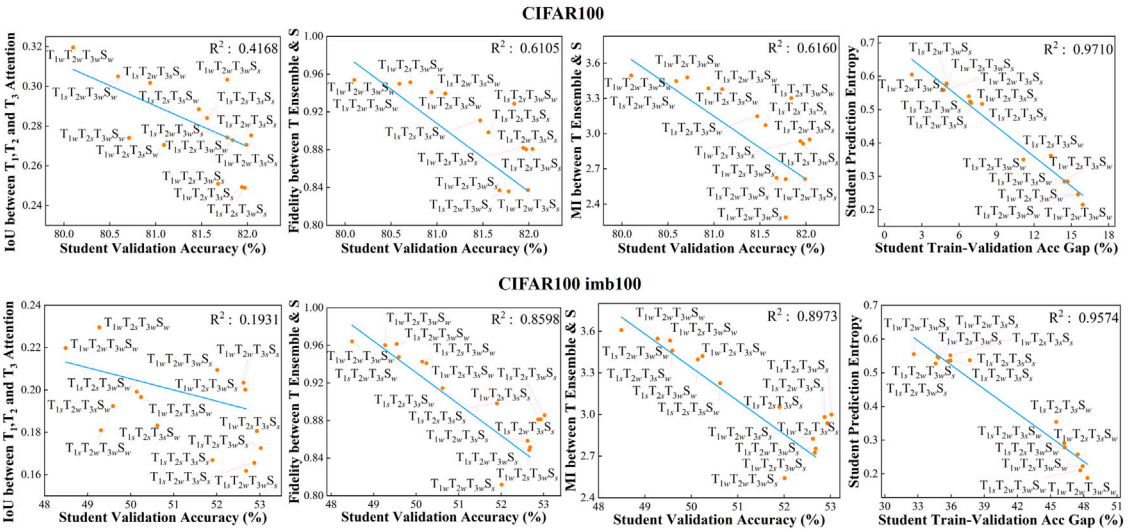


Fig. A.3. For ResNet models. Scatter plots of teacher attention IoU, fidelity, mutual information, and student entropy in 3-teacher ensemble KD cases. These results, aligning with the tendencies observed in 2-teacher cases, further support our conclusions in the main text.

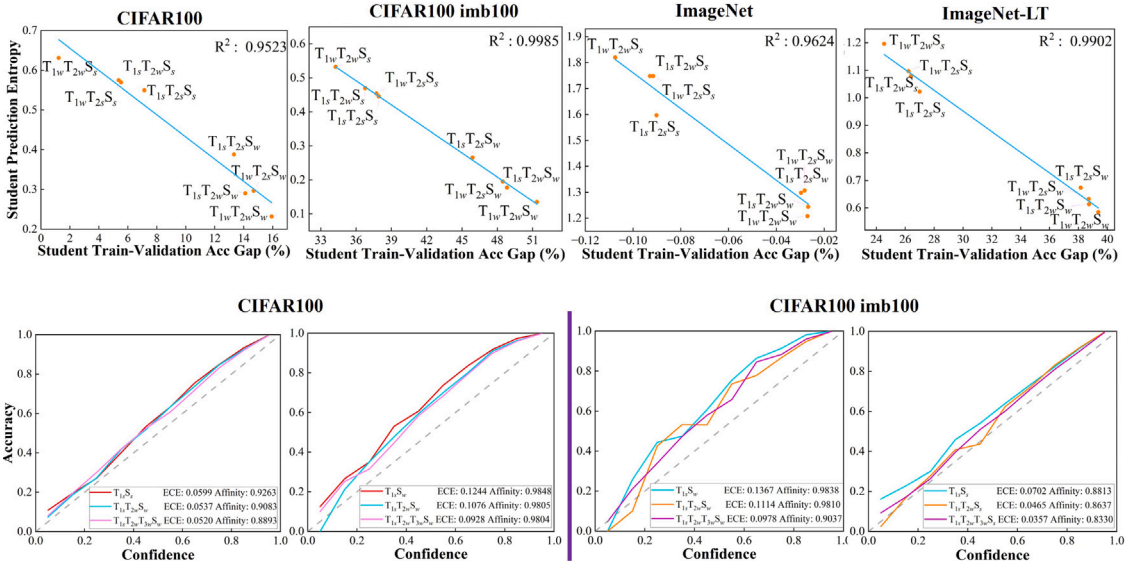


Fig. A.4. For ResNet models. *Top*: Scatter plots of student entropy versus overfitting (gap between top-1 validation and training accuracy) during KD training. *Bottom*: Calibration reliability diagrams with varied teacher numbers (1 to 3) for CIFAR100 imb100 and its balanced counterpart. Stronger augmentation (indicated by decreased Affinity) and more teachers in the ensemble contributes to improved model calibrations and mitigate overfitting effects.

where B_m denotes the set of samples in the mth bin. The function $\mathrm{Acc}(B_m)$ calculates the accuracy within bin B_m, while $\mathrm{conf}(B_m)$ computes the average predicted confidence of samples in the same bin.

In Fig. A.4 *Top*, a notable inverse relationship was observed between the entropy of the student model's predictions and overfitting. While stronger data augmentation leading to increased entropy (indicative of lower confidence), there was a concurrent decrease in the tendency of the student model to overfit the training data, as evidenced by the reduction in the train-validation accuracy gap. Fig. A.4 *Bottom* further compares the model calibration reliability diagrams for KD with varied teacher numbers (from 1 to 3) and data augmentation strengths. It can be observed that as the number of teachers increased or the augmentation strength increased (indicated by decreased Affinity), the student models exhibited better calibration.

Table A.4 further provides the Expected Calibration Error (ECE) with corresponding Affinity values for all the trials with 2-teacher ensemble KD. This aids in understanding the data augmentation strengths and the decreasing tendencies in all the previous scatter plots: Recall that Affinity measures the offset in data distribution between the original one and the one after data augmentation captured by the student, and lower Affinity corresponds to higher augmentation strength, leading to higher student accuracy. Thus, for the trials with strong data augmentation (e.g., $T_{1w}T_{2w}S_s$ in CIFAR-100, CIFAR-100 imb100, and ImageNet; $T_{1s}T_{2w}S_s$ in ImageNet-LT), they not only correspond to a relatively small ECE but also a high validation accuracy.

A.7. Experiments with vision transformers

In this section, we also provide experiments with Vision Transformers (ViTs) (Dosovitskiy et al., 2021) on CIFAR100 imb100 dataset where the attention map can be obtained directly with the built-in attention module. As shown in Fig. A.5, our analysis method can be applied to attention-based methods such as ViT. The only difference is that when calculating IoU, we can directly use the built-in attention module of ViT to obtain the attention maps. In this experiment, two ViT-b32 teachers are distilled on one ViT-b16 student for CIFAR100

Table A.4
ECE and Affinity of models with various data augmentation strengths.

Dataset	Metric	Model							
		$T_{1w}T_{2w}S_w$	$T_{1w}T_{2w}S_s$	$T_{1s}T_{2w}S_w$	$T_{1s}T_{2w}S_s$	$T_{1w}T_{2s}S_w$	$T_{1w}T_{2s}S_s$	$T_{1s}T_{2s}S_w$	$T_{1s}T_{2s}S_s$
Cifar100	ECE	0.0776	**0.0124**	0.1076	0.0537	0.0994	0.0568	0.1397	0.0745
	Affinity	0.9807	**0.8611**	0.9805	0.9083	0.9858	0.9143	0.9729	0.9310
Cifar100 imb100	ECE	0.0979	0.0103	0.1114	0.0465	0.0711	0.0482	0.1303	0.0651
	Affinity	0.9763	**0.8132**	0.9810	0.8637	0.9751	0.8635	0.9723	0.8955
ImageNet	ECE	0.0275	**0.0095**	0.0233	0.0118	0.0126	0.0107	0.0122	0.0193
	Affinity	0.9901	**0.8767**	0.9930	0.8988	0.9845	0.9131	0.9871	0.9122
ImageNet long-tail	ECE	0.0322	0.0226	0.0357	0.0224	0.0494	0.0307	0.0499	**0.0178**
	Affinity	0.9850	**0.8311**	0.9755	0.8704	0.9782	0.8751	0.9903	0.8971

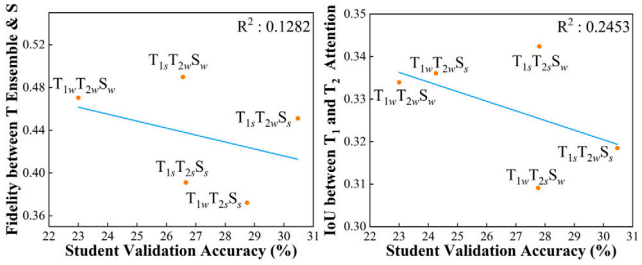


Fig. A.5. Scatter plots for experiments with Vision Transformer (ViT) on CIFAR100 imb100 dataset. *Left*: Fidelity (measured by top-1 A) and *Right*: IoU between T_1 and T_2 during KD training. These decreasing tendencies align with our conclusions drawn from ResNet experiments, suggesting the applicability of our analysis method to attention-based methods like ViT. The main distinction is in calculating IoU, where we can directly use ViT's built-in attention module to obtain the attention maps.

imb100 dataset. And the conclusions in our manuscript still holds for these two cases. That is, lower student–teacher fidelity and larger teachers' attention diversity correlate with higher student validation accuracy.

A.8. Few-shot knowledge distillation scenario

Previously, our problem setting required internal access to the teachers and their complete original training set with labels. However, in real-world applications, these resources are not always available. Teachers may withhold their parameters or logits due to security and privacy concerns, or the distillation process might occur on an external party's side where access to data is limited. In light of these practical considerations, this section focuses on KD in few-shot, data-free, or black-box scenarios.

Specifically, this section conducts experiments with the black-box few-shot unsupervised KD method FS-BBT (Nguyen et al., 2022), and the data-free KD method MAD (Do et al., 2022) on the CIFAR-100 and ImageNet datasets. For comparison, we also evaluate our proposed KD method, both in white-box supervised settings and black-box unsupervised scenarios within few-shot learning. Notably, in the unsupervised case, we adhere to the settings outlined in Nguyen et al. (2022), where teacher models are trained using the full dataset in a supervised manner, while the student model accesses only a small subset of training images without class labels, receiving target information from the teachers' outputs. To ensure fairness, we use the same number of original subset images N as in Nguyen et al. (2022), specifically $N = 10\,000$ for CIFAR-100 and $N = 50\,000$ for ImageNet.

The experimental results are presented in Table A.5. For reference, we include the results of our proposed method, Ours(2T), which utilizes fully supervised datasets, and these results are identical to those in Table 2 in Section 7. In this context, Ours(2T) refers to the same setup as before, specifically the KD process where two ResNet50 teacher models distill knowledge to one ResNet18 student model, denoted as $T_{1s}T_{2w}S_s$.

Table A.5
Few-shot KD: Comparison of our two-teacher ensemble KD against the black-box few-shot unsupervised KD method FS-BBT, and the data-free KD method MAD on CIFAR-100 and ImageNet datasets. without any additional models like VAE or EMA image generators, our proposed method achieves comparable student performance under the black-box few-shot unsupervised scenario.

Dataset	Method	Val-Acc
Cifar100	FS-BBT	0.5628
	MAD	0.6405
	Ours(2T) few-shot Bl-Un[a]	**0.6534**
	Ours(2T) few-shot Wh-Su[b]	0.7065
	Ours(2T) full	0.8195
ImageNet	FS-BBT	0.4329
	MAD	**0.4548**
	Ours(2T) few-shot Bl-Un[a]	0.4068
	Ours(2T) few-shot Wh-Su[b]	0.4258
	Ours(2T) full	0.6917

[a] Bl-Un: black-box teachers, unsupervised student.
[b] Wh-Su: white-box teachers, supervised student.

Table A.5 illustrates that, without the need for additional models such as the VAE used in Nguyen et al. (2022) or EMA image generators from Do et al. (2022), our proposed approach achieves comparable student performance in the black-box few-shot unsupervised scenario. Specifically, in this setting, Ours(2T) outperforms FS-BBT by 9.06% and outperforms MAD by 1.29% on the CIFAR-100 dataset. Meanwhile, on the ImageNet dataset, the performance of Ours(2T) is comparable to both the FS-BBT and MAD methods.

The effectiveness of our proposed method stems from the teacher ensemble structure combined with diverse data augmentation strengths, which enhance the model's attention mechanisms.

A.9. Comparison with model quantization

This section focuses on model quantization and presents experiments using the Post Training Static Quantization technique (Nagel et al., 2021) on our pre-trained teacher model with ImageNet dataset. This technique reduces the model's 32-bit floating-point numbers to 8-bit integers. We implemented this using torch.ao.quantization, and since PyTorch does not provide quantized operator implementations on CUDA, the experiments were conducted on a CPU, which is the only supported device for testing quantized models. The experimental results are summarized in Table A.6. For comparison, we include the baseline results for the pre-trained teacher ResNet50 model and the Ours(2T) Student ResNet18 model, as presented in Table 2.

Table A.6 shows that, compared to our proposed KD method, static quantization on the pre-trained teacher model achieves a higher validation accuracy (by 4.49%) and results in a smaller model size (by 20.8MB). However, due to PyTorch quantization's lack of CUDA support, the average inference time is significantly higher, nearly 1 s per image, while our KD method only takes 0.0036 s. To address this issue, effective model quantization requires dedicated efforts in its implementation, which can be model-specific and often necessitates

Table A.6
Model quantization experimental results using Post Training Static Quantization technique, on our pre-trained teacher model with ImageNet dataset. This technique reduces the 32-bit floating-point numbers in the model to 8-bit integers. For comparison, the baseline (pre-trained teacher model ResNet50) and Ours(2T) student model ResNet18 are also listed.

Method	Val-Acc	Avg Infer time	Model Size
Baseline[a]	**0.7620**	0.0059 s	102.5 MB
Static Quant[b]	0.7366	0.9143 s	**26.1 MB**
Ours(2T)[c]	0.6917	**0.0036 s**	46.9 MB

[a] Baseline: Teacher ResNet50.
[b] Static Quant: Static Quantization ResNet50.
[c] Ours(2T): Student ResNet18.

hardware-level adaptations. Such processes, along with quantization-aware training, can be complex. In contrast, the KD technique offers greater flexibility: once trained, the student model is ready for deployment, with its architecture or size easily customizable.

Data availability

The data and codes are already shared online.

References

Allen-Zhu, Z., & Li, Y. (2023). Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International conference on learning representations*.

Asif, U., Tang, J., & Harrer, S. (2019). Ensemble knowledge distillation for learning improved and efficient networks. In *European conference on artificial intelligence*.

Bai, Y., Wang, Z., Xiao, J., Wei, C., Wang, H., Yuille, A., Zhou, Y., & Xie, C. (2023). Masked autoencoders enable efficient knowledge distillers. In *Computer vision and pattern recognition*.

Cubuk, E. D., Dyer, E. S., Lopes, R. G., & Smullin, S. (2021). Tradeoffs in data augmentation: An empirical study. In *ICLR*.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2019). AutoAugment: Learning augmentation policies from data. In *Computer vision and pattern recognition*.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*.

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

Do, K., Le, H., Nguyen, D., Nguyen, D., Harikumar, H., Tran, T., Rana, S., & Venkatesh, S. (2022). Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation. In *Advances in neural information processing systems*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Gou, J., Chen, Y., Yu, B., Liu, J., Du, L., Wan, S., & Yi, Z. (2024). Reciprocal teacher-student learning via forward and feedback knowledge distillation. In *IEEE transactions on multimedia*.

Gou, J., Sun, L., Yu, B., Wan, S., & Tao, D. (2023). Hierarchical multi-attention transfer for knowledge distillation. In *ACM trans. multimedia comput. commun. appl.*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning* (pp. 1321–1330). JMLR.org.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*: *Technical report*, University of Toronto.

Lao, S., Song, G., Liu, B., Liu, Y., & Yang, Y. (2023). Unikd: Universal knowledge distillation for mimicking homogeneous or heterogeneous object detectors. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6362–6372).

Lewy, D., & Mańdziuk, J. (2023). AttentionMix: Data augmentation method that relies on BERT attention mechanism. arXiv preprint arXiv:2309.11104.

Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., & Liang, D. (2022). Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *AAAI*.

Li, W., Shao, S., Liu, W., Qiu, Z., Zhu, Z., & Huan, W. (2022). What role does data augmentation play in knowledge distillation? In *Proceedings of the Asian conference on computer vision* (pp. 2204–2220).

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *IEEE conference on computer vision and pattern recognition*.

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., van Baalen, M., & Blankevoort, T. (2021). A white paper on neural network quantization. arXiv preprint arXiv:2106.08295.

Nguyen, D., Gupta, S., Do, K., & Venkatesh, S. (2022). Black-box few-shot knowledge distillation. In *European conference on computer vision*.

Özdemir, Ö., & Sönmez, E. B. (2022). Attention mechanism and mixup data augmentation for classification of COVID-19 computed tomography images. *Journal of King Saud University - Computer and Information Sciences*, *34*(8, Part B), 6199–6207. http://dx.doi.org/10.1016/j.jksuci.2021.07.005.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Shen, R., Bubeck, S., & Gunasekar, S. (2022). Data augmentation as feature manipulation. In *Proceedings of the 39th international conference on machine learning* (pp. 19773–19808). PMLR.

Shrivastava, A., Qi, Y., & Ordonez, V. (2023). Estimating and maximizing mutual information for knowledge distillation. In *CVPR workshop*.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., & Wilson, A. G. (2021). Does knowledge distillation really work? In *Advances in neural information processing systems* (pp. 6906–6919). Curran Associates, Inc..

Sun, S., Ren, W., Li, J., Wang, R., & Cao, X. (2024). Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Tian, S., & Chen, D. (2022). Attention based data augmentation for knowledge distillation with few data. *Journal of Physics: Conference Series*, *2171*(1), Article 012058.

Tsantekidis, A., Passalis, N., & Tefas, A. (2021). Diversity-driven knowledge distillation for financial trading using deep reinforcement learning. *Neural Networks*, *140*, 193–202. http://dx.doi.org/10.1016/j.neunet.2021.02.026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*.

Wang, G.-H., Ge, Y., & Wu, J. (2022). Distilling knowledge by mimicking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 8183–8195. http://dx.doi.org/10.1109/TPAMI.2021.3103973.

Xiang, L., & Ding, G. (2020). Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. arXiv:2001.01536.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Computer vision and pattern recognition*.