Personalized Education: Blind Knowledge Distillation

Xiang Deng[1], Jian Zheng[2], and Zhongfei Zhang[1]

[1] Sate University of New York at Binghamton
xdeng7@binghamton.edu, zhongfei@cs.binghamton.edu
[2] Amazon
nzhengji@amazon.com

Abstract. Knowledge distillation compresses a large model (teacher) to a smaller one by letting the student imitate the outputs of the teacher. An interesting question is why the student still typically underperforms the teacher after the imitation. The existing literature usually attributes this to model capacity differences between them. However, capacity differences are unavoidable in model compression, and even large capacity differences are desired for achieving high compression rates. By designing exploratory experiments with theoretical analysis, we find that model capacity differences are not necessarily the root reason; instead the distillation data matter when the student capacity is greater than a threshold. In light of this, we propose personalized education (PE) to first help each student adaptively find its own blind knowledge region (BKR) where the student has not captured the knowledge from the teacher, and then teach the student on this region. Extensive experiments on several benchmark datasets demonstrate that PE substantially reduces the performance gap between students and teachers, even enables small students to outperform large teachers, and also beats the state-of-the-art approaches. Code link: https://github.com/Xiang-Deng-DL/PEBKD

Keywords: knowledge distillation, model compression, classification

1 Introduction

The successes of deep neural networks (DNNs) [23, 10] are accompanied with the requirements of large amounts of computation and memory, which seriously restricts their deployment on resource-limited devices. One widely used solution is knowledge distillation (KD) [16] that compresses a large model (teacher) to a small one (student) by enforcing the student to mimic the outputs of the teacher. However, there is typically still a performance gap between them even if the student has imitated the outputs of the teacher. Figuring out the reason for this gap is essential for further improving the student performance.

Mirzadeh et al. [27] argue that the model capacity difference causes the failure for transferring the knowledge from a large teacher to a small student, thus leading to a performance gap. Similarly, Cho and Hariharan [5] point out that as

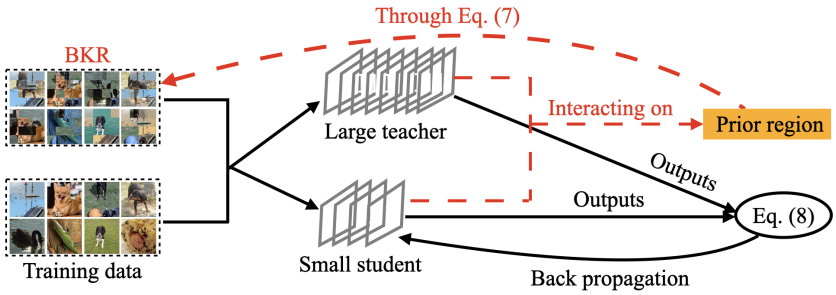


Fig. 1. Overview of PE

the teacher grows in capacity and accuracy, it is difficult for the student to emulate the teacher. However, the capacity reason is trivial for improving the student performance, since capacity differences are unavoidable in model compression. More essentially, large capacity differences are desired in model compression for achieving high compression rates. In light of this, we conduct simulation experiments (in Section 4) and find that in most experimental settings of the existing literature [34], the reason for the performance gap is not necessarily the capacity difference as the student is powerful enough to memorize the teacher's outputs. Instead, the reason lies in *distillation data* on which the knowledge is transferred.

In reality, it is not rare for human students to do better than their teachers. These excellent human students not only well capture the knowledge from their teachers but also learn more related knowledge on their own. This gives an insight for students in KD to match or outperform their teachers. We find that the students in KD have not well captured the knowledge from their teachers as they only mimic the behavior of the teachers on sparse training data points. We thus propose to go beyond the sparse, in-distribution distillation. However, simply going beyond distribution may not be optimal as different students master different levels of knowledge from the teacher. Similar to human students needing personalized education based on their own situations, we propose personalized education (PE) for KD to assist each student to spot and learn its own blind knowledge region (BKR) where the student fails to learn well from the teacher.

As image samples lie in a large, high-dimensional space, directly learning the BKRs in the full space is impossible and also not all samples in the large space are beneficial to the student. We thus learn BKR from a prior region where the samples share similar patterns with the training data (in-distribution) samples. We propose MixPatch as the prior region that linearly combines the patches in two images with different coefficients. MixPatch is inspired by but different from Mixup [43] that linearly combines two full images with a coefficient. Mixup is thus a special case of MixPatch when the patch size in MixPatch is set to the full image size. MixPatch theoretically can generate any image when the patch size is set to 1×1. MixPatch is also different from CutMix [40] that cuts a patch from one image and pastes it to another image. Unlike Mixup and CutMix, MixPatch is specially designed for KD instead of standard supervised learning, since it is almost impossible to directly generate the corresponding labels for MixPatch

images in standard supervised learning while Mixup and CutMix can linearly combine the ground-truth labels to generate labels for new samples. Thanks to the pretrained teacher which can provide supervision signals, MixPatch can be specially used in KD. With MixPatch as the prior region, as shown in Figure 1, PE first lets a student interact with the teacher to find its own BKR from the prior region and then lets it learn from the teacher on the BKR.

Our main contributions are summarized as follows:

- Different from the common belief that model capacity differences result in the performance gap between small students and large teachers, we find through designing exploratory experiments that capacity differences are not necessarily the root reason but the distillation data matter when student capacities are greater than a threshold.
- Different from the existing work focusing on designing different criteria to align representations or logits between teachers and students, we study knowledge distillation from a novel (data) perspective and accordingly propose personalized education (PE). PE goes beyond in-distribution distillation and adaptively learns the BKR for each student from a prior region.
- We propose a novel, simple yet effective data augmentation strategy (i.e., MixPatch) specially designed for KD, which addresses the limited image diversity issue of Mixup. It can be separately used or serve as a part of PE to enhance the student performance.
- Extensive experiments on several benchmark datasets demonstrate that PE reduces the student-teacher performance gap substantially, even enables small students to match or outperform large teachers, and also betas the existing SOTA approaches significantly. Furthermore, PE is also compatible with the existing SOTA approaches to further largely improve their performances.

2 Related Work

Knowledge Distillation. Hinton et al. [16] propose KD that trains a student network by using the softened logits of a teacher network as the targets. Different from one-hot labels, the soft targets contain instance-to-class similarity information (i.e., dark knowledge) learned by the teacher. However, KD only transfers the logits but fails to transfer the representations. Many approaches thus have been proposed to align the representations learned by a student and a teacher. Fitnets [31] lets a student imitate the intermediate features of a teacher through regressions. CRD [34] transfers representations by using contrastive learning. SSKD [38] uses extra self-supervised learning tasks to enhance the knowledge transfer. Other distillation approaches [42, 17, 39, 33, 37, 29, 28, 14, 15, 5, 2, 21, 1, 12, 24, 45, 44, 19, 8, 4, 18, 9, 3, 7] utilize different criteria to align the feature representations or logits between a teacher and a student. Different from these efforts focusing on designing different fitting criteria, we address knowledge distillation from a data perspective by exploring the blind knowledge regions for students.

Teacher-Student Performance Gaps. Mirzadeh et al. [27] observe that the model capacity gap results in the failure for transferring knowledge from a large

teacher to a small student, thus causing a performance gap. To reduce this gap, they propose a multi-step KD framework by using several intermediate-size networks (teacher assistants). However, the students still underperform the teachers substantially. Cho and Hariharan [5] argue that as the teacher grows in capacity and accuracy, it is difficult for the student to emulate the teacher. To reduce the influence of the large capacity gap, they regularize both the teacher and the distillation process by early stopping. We find that capacity differences are not necessarily the root reason when student capacities are greater than a threshold.

Mixup. MixPatch is proposed as a prior region in PE, which is inspired by Mixup [43] and aims to address the image diversity issue of Miuxp. Mixup [43] linearly interpolates a pair of training samples and their one-hot labels to generate new data. Linear combination of samples can preserve some patterns in the original samples, which can be beneficial to model learning. However, the simple linear interpolation can only generate limited samples, which is not enough for PE to learn the BKR. We thus propose MixPatch to address this issue by linearly combining the patches in two images with different coefficients. Mixup is thus a special case of MixPatch when the patch size in MixPatch is set to the image size. Furthermore, setting the patch size to 1 pixel can theoretically generate any image in the sample space. MixPatch is also substantially different from CutMix [40] where the patch is cut from one image and pasted into another image. Another related technique is ActiveMixup [36] which uses Mixup to generate a big image pool and then actively selects the images that the classifier has a low confidence on. However, the strategy ignores critical images that the classifier has a high confidence on but with a wrong prediction. The proposed PE addresses this issue by letting the student interact with the teacher to find its own BKR.

3 Reformulating KD

KD [16] aligns the outputs of a student and a teacher over training data $D_t = (X_t, Y_t) = \{(x_i, y_i)\}_{i=0}^{n}$ where X_t and Y_t are the training samples and the ground truth, respectively. The complete objective is written as:

$$\mathcal{L}_{KD} = \sum_{(x_t, y_t) \in D_t} [\alpha \mathcal{L}_{CE}(S_\theta(x_t), y_t) + \beta \mathcal{L}_{KL}(S_\theta, T, x_t)] \tag{1}$$

where α and β are the weights for balancing the contributions of the two terms; S_θ and T denote a student network with parameters θ and a pretrained teacher network, respectively; $\mathcal{L}_{CE}$ is the regular cross-entropy loss; $\mathcal{L}_{KL}$ is the distillation loss for transferring knowledge from the teacher to the student:

$$\mathcal{L}_{KL}\left(S_\theta, T, x_t\right) = \tau^2 \mathcal{K}\left(\sigma\left(\frac{T\left(x_t\right)}{\tau}\right), \sigma\left(\frac{S_\theta\left(x_t\right)}{\tau}\right)\right) \tag{2}$$

where σ is the softmax function; τ is a temperature to generate soft labels; $\mathcal{K}$ denotes KL-divergence. KD can be considered as using a function S_θ to fit another function T.

Note that in (1), $\mathcal{L}_{CE}$ requires both data samples X_t and the corresponding ground truth Y_t while $\mathcal{L}_{KL}$ only needs data samples X_t for distilling the teacher knowledge. In light of the difference, **we consider KD from a semi-supervised perspective** and reformulate (1) in a more general form:

$$\mathcal{L} = \sum_{(x_t, y_t) \in (X_t, Y_t)} \alpha \mathcal{L}_{CE}(S_\theta(x_t), y_t) + \sum_{x_d \in X_d} \beta \mathcal{L}_{KL}(S_\theta, T, x_d) \tag{3}$$

where we introduce a new concept: **distillation dataset** X_d is a set of samples on which the knowledge is transferred from a teacher to a student. The first term in the right hand side of (3) is supervised while the second term is unsupervised. It is obvious that the widely used objective (1) is a special case of (3) when X_d is set to X_t.

4 Why Small Students Underperform Large Teachers?

In this section, we systematically examine the reason for the performance gap between small students and large teachers. We first introduce several definitions.

Definition 1. *Memorization Error (ME): For a given task with data sample distribution $P(x)$, ME measures the degree of a student S_θ fitting the outputs of a teacher T over $P(x)$:*

$$E(S_\theta, T, P) = \mathop{\mathbb{E}}_{x \sim P(x)} M(T(x), S_\theta(x)) \tag{4}$$

where M denotes a distance metric such as KL-divergence or mean square error.

When ME is (or extremely close to) 0, it means that the student can completely memorize the outputs of the teacher over the data distribution. In this paper, we take KL-divergence as M.

Definition 2. *Capable Students (CSTs) and Incapable Students (ISTs): network S_θ is a CST of teacher T on a task with data sample distribution P if there exists θ such that $E(S_\theta, T, P)$=0; otherwise, it is an IST.*

Note that whether a student is a CST or IST is determined by its own capacity, the complexity of the teacher function T, and the task data distribution P. Obviously, a CST is able to fully fit the teacher outputs over data distribution $P(x)$. In contrast, an IST does not have the capacity to fit the teacher. For ISTs, the common belief holds that the student-teacher capacity gap causes the performance gap. For example, we cannot expect a two-layer neural network with 1000 parameters to fit the outputs of ResNet-101 with 1.7M parameters on CIFAR-100. However, in the current SOTA approaches and applications [34], the commonly used students are modern neural network architectures, such as ResNet-32, ResNet8×4, VGG-8, and WRN-16-2. We empirically show that these models are CSTs on commonly used benchmark datasets, i.e., CIFAR-10, CIFAR-100, and Tiny ImageNet.

Table 1. ME of different networks on CIFAR-10, CIFAR-100, and Tiny ImageNet. All the values are accurate to 1 decimal place.

Teacher (#Params)	WRN-40-2 (2.26M)	VGG-13 (9.46M)	ResNet32×4 (7.43M)	ResNet-110 (1.74M)	ResNet32×4 (7.43M)	VGG-13 (9.46M)	VGG-13 (9.46M)
Student (#Params)	WRN-16-2 (0.70M)	VGG-8 (3.97M)	ResNet8×4 (1.23M)	ResNet-32 (0.47M)	ShuffleNetV2 (1.36M)	SN2 (0.0284M)	SN3 (0.0298M)
CIFAR-10	0.0	0.0	0.0	0.0	0.0	1.7	0.1
CIFAR-100	0.0	0.0	0.0	0.0	0.0	2.4	0.3
Tiny ImageNet	0.0	0.0	0.0	0.0	0.0	4.2	1.9

To check whether student S_θ is a CST of teacher T on a task, we minimize ME to check whether $E(S_\theta, T, P)$ can achieve 0. However, in practice, it is impossible to calculate $E(S_\theta, T, P)$ as the data distribution P is typically unknown. Fortunately, we have the access to a set of training data (X_t, Y_t). With the training data, we approximate ME $E(S_\theta, T, P)$ with the empirical error:

$$E_{em}(S_\theta, T, X_t) = \frac{1}{|X_t|} \sum_{x_t \in X_t} M(T(x_t), S_\theta(x_t)) \tag{5}$$

For comparison, we also evaluate two tiny neural networks which are expected to be ISTs, i.e., SN-2 and SN-3 with two and three layers, respectively. We report the ME in Table 1, where we adopt the students and the teachers that share the same architectures (e.g., WRN-40-2 and WRN-16-2) or use different architectures (e.g., ResNet32×4 and ShuffleNetV2). As expected, the widely used students achieve ME 0.0[3] on all the three benchmark datasets while the tiny networks (i.e., SN2 and SN3) have large ME (e.g., 2.4 and 4.2), which demonstrates that the widely used students are CSTs. However, as shown in Table 2, these students (i.e., KD) after distillation still underperform the teachers by a significant margin on the test data. This indicates that these students have well learned the knowledge on sparse training data points but have not well captured the local function shapes of the teachers within the data distribution so that they fail on the test data. This suggests the following proposition:

Proposition 1 *In KD, for CSTs, only fitting the teacher function on sparse training data points cannot enable them to well capture the in-distribution function shape of the teacher (i.e., in-distribution knowledge), thus leading to a performance gap. For ISTs, capacity differences cause the performance gap.*

We further conduct exploratory experiments to verify this proposition and the theoretical proof of the proposition is given in Appendix A. The exploratory experiments compare the student performances in the following two settings: (a) setting the distillation dataset to training data points; (b) setting the distillation dataset to real data distribution $P(x)$. As $P(x)$ is typically unknown in

[3] Note that all the ME values in Table 1 are accurate to 1 decimal place. 0.0 is not exact 0 but is extremely close to 0.

Table 2. Simulation results on CIFAR-100 in terms of test accuracy (%). Underline denotes that small students outperform large teachers.

Teacher	ResNet32×4	WRN-40-2	VGG-13	ResNet32×4	VGG-13	VGG-13
Student	ResNet8×4	WRN-16-2	VGG-8	ShuffleNetV2	SN2	SN3
Teacher	79.63	76.46	75.38	79.63	75.38	75.38
Vanilla Student	72.51	73.64	70.68	73.12	26.29	55.31
Student Type	CST	CST	CST	CST	IST	IST
KD	73.33	74.92	72.98	74.45	26.04	55.32
Simulation KD	**79.91**	**78.46**	**77.99**	**81.64**	25.58	57.50

practice, we conduct a simulation experiment on CIFAR-100. We suppose that the union of the training dataset and the test dataset in CIFAR-100 can accurately represent the real data distribution for this task. Then we randomly draw data samples from the vicinity around the training data and the test data as the distillation dataset, i.e., X_d in Eq. (3). Consequently, the distillation dataset can sufficiently represent the real data sample distribution. Note that in this experiment, we **never spy the ground truth of the test samples**, since the distillation dataset X_d does not use ground truth as shown in Eq. (3). This means that the students are trained without any additional supervision compared with the teachers as training datset (X_t, Y_t) in Eq. (3) does not change. Since CSTs are able to fully memorize the outputs of the teachers, we expect them to achieve the same accuracies as or higher accuracies than those of the teachers. In contrast, we expect ISTs to achieve lower accuracies than those of the teachers. Table 2 shows the simulation results. As expected, all the CSTs outperform the teachers in the simulation experiments (i.e., Simulation KD). This is due to the following facts: first, by using the simulated distillation dataset, the distillation objective in Eq. (3) makes the CSTs fully capture the knowledge of the teachers within the data distribution; second, the cross-entropy objective in Eq. (3) enables the CSTs to learn their own knowledge. Consequently, CSTs contain both the teacher knowledge and the knowledge learned on their own, which results in better performances than those of the teachers. SN2 and SN3 still underperform the teachers in the simulation experiments due to their limited capacities. These results empirically validate the proposition.

The simulation experiments also suggest a way for CSTs to outperform large teachers. That is to sufficiently distill the knowledge from the teachers with a well representative distillation dataset. Unfortunately, it is impossible to have such a distillation dataset as the real data sample distribution $P(x)$ is typically unknown in reality. As directly modelling the high-dimensional data distribution from sparse data points is even more difficult than training the classifier itself, we propose to go beyond in-distribution distillation by using out-of-distribution (OOD) samples to assist distillation based on the following proposition:

Proposition 2 *Out-of-distribution (OOD) samples can be beneficial to knowledge distillation, but not all OOD samples are beneficial.*

The theoretical proof of this proposition is given in Appendix B.

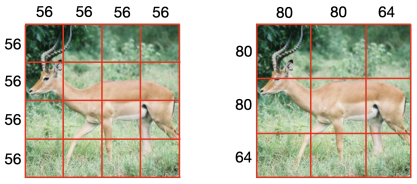
5 Personalized Education

Although OOD samples can be useful for KD, simply going beyond the distribution may not be optimal, given the fact that different students master different levels of knowledge from the teacher. We thus further propose personalized education (PE) which automatically spots the blind knowledge region (BKR) for each student where the student has not well learned from the teacher. Due to the large, high-dimensional image space, finding the beneficial BKR from the full space is difficult or even impossible. We thus learn the BKR from a prior region where the OOD samples share similar local patterns to the original training data (i.e., in-distribution data).

5.1 MixPatch

Mixup [43] linearly interpolates two training images to generate a new image which can preserver or share similar patterns to the original images. However, the diversities of the generated images are limited by the simple linear interpolation. CutMix [40] suffers a similar issue as it only cuts one patch from one image and pastes it to another image. To increase the image sample diversity and preserve similar local patterns to the original image, we propose MixPatch that linearly interpolates the patches in two images with different coefficients.

MixPatch has two hyper-parameters, i.e., patch size $s_h \times s_w$ where s_h and s_w are the height and width of the patch, and beta-distribution parameter a that is used to generate the coefficients. Suppose that the image size is $h \times w$ where h and w are the height and the width of the image, respectively (the channel size is omitted here); then the image is split into $m = \lceil \frac{h}{s_h} \rceil \times \lceil \frac{w}{s_w} \rceil$ patches, where $\lceil x \rceil$ is the ceiling function that returns the smallest integer that is greater than or equal to x. When h and w are divisible by s_h and s_w respectively, the sizes of the m patches are all equal to $s_h \times s_w$. One intuitive example is given in Figure 2(a) where all the 16 patches are 56×56. On the contrary, when the image size is not divisible by the patch size, some of the patches have a size smaller than $s_h \times s_w$, with one example shown in Figure 2(b) where 80×64 smaller than 80×80.



(a) $s_h \times s_w$ is set to 56×56. (b) $s_h \times s_w$ is set to 80×80.

Fig. 2. An image of size 224×224 are split into patches with different $s_h \times s_w$.

Each images x_i can thus be represented by m patches, i.e., $[h_{i0}, h_{i1}, ..., h_{im}]$, where h_{ij} is the jth patch in image x_i. MixPatch generates a new image $\hat{x}_z$ by linearly combining the patches of two random training images x_i and x_k with different coefficients:

$$\hat{h}_{zj} = \lambda_j h_{ij} + (1 - \lambda_j) h_{kj},\ j = 0, 1, 2, ..., m \tag{6}$$

where $\hat{h}_{zj}$ is the jth patch of new sample $\hat{x}_z$, i.e., $\hat{x}_z = [\hat{h}_{z0}, \hat{h}_{z1}, ..., \hat{h}_{zm}]$, and $\lambda_j \sim Beta(a, a)$ where a is the beta-distribution parameter.

We can see that when patch size $s_h \times s_w$ is set to image size $h \times w$, MixPath is reduced to Mixup, which indicates that Mixup is a special case of MixPatch. Furthermore, when patch size is set to 1×1 (i.e., 1 pixel), MixPatch is able to generate any images by tuning λ_j but loses the information of the local patterns in the original images. Therefore, the patch size controls the contributions of the new image diversity and the local pattern information. Another difference between MixPatch and Mixup is that MixPatch is specially designed for KD as it is difficult to directly generate labels for these new images in standard supervised learning. Thanks to the pretrained teacher in KD, the student can make use of these MixPatch samples by imitating the outputs of the teacher on them.

MixPatch samples can be used in two ways to assist distillation. One way is to directly use them to enhance the student knowledge by manually tuning and then freezing hyper-parameter patch size $s_h \times s_w$ and parameter a, but it may not be the optimal. The other way is to adaptively update the patch size and the distribution parameter by letting the student interact with the teacher during the training process.

5.2 Blind Knowledge Region Discovery

PE aims to help each small student spot its own BKR from a prior region by the interaction with the teacher. We propose a gradient-free search strategy to find the BKR for each student from the MixPatch prior region. Since the input images to modern DNNs are typically fix-sized with the height equal to the width (e.g., 224×224 for ImageNet image), for simplicity, we use one parameter s to represent the patch size by assuming $s = s_h = s_w$, and one parameter h to denote the image size by assuming $h = w$.

The MixPatch region is then determined by two hyper-parameter a and s that control coefficient (i.e., λ) distribution and the patch size, respectively. PE spots the BKR for a student from a set of candidate regions. The candidate values for a are typically set to $\mathbf{a} = \{0.1, 0.5, 1.0\}$ and those for s are set to $\mathbf{s} = \{h, \frac{h}{2}, ..., \frac{h}{n}\}$ where n is an integer. As a BKR is where the student has not well learned the knowledge from the teacher, we can search the BKR by maximizing the output differences of the student and the teacher on the prior region:

$$\text{argmax}_{a,s} \mathop{\mathbb{E}}_{\hat{x} \sim \text{MixPatch}(a,s)} \mathcal{L}_{KL}(S_\theta, T, \hat{x}), \text{ for } a \in \mathbf{a}\ s \in \mathbf{S} \tag{7}$$

PE adaptively searches the BKR with (7) every k epochs to update a and s. Note that this search process is very fast as it is gradient-free.

The BKR samples are then added to the distillation dataset X_d in Eq. (3). PE thus can be considered as a data-driven approach that enhances the student by enforcing it to mimic the behavior of the teacher on its own BKR:

$$\mathcal{L}_{\text{PE}} = \mathcal{L}_{\text{KD}} + \sum_{\hat{x} \in \text{BKR}} \mathcal{L}_{KL}(S_\theta, T, \hat{x}) \tag{8}$$

where BKR = MixPatch$(\overline{a}, \overline{s})$; $\overline{a}$ and $\overline{s}$ are the optimally identified values by (7) and updated every k epochs. By adaptively learning the BKR, PE is expected to

Table 3. Effects of different components in PE.

Teacher	Student	Teacher	Vanilla student	KD	KD+Mixup	KD+MixPatch	PE
WRN-40-2	WRN-16-2	76.46	73.64	74.92	75.56	75.94	**76.57**
VGG-13	VGG-8	75.38	70.68	72.98	74.58	74.93	**75.41**

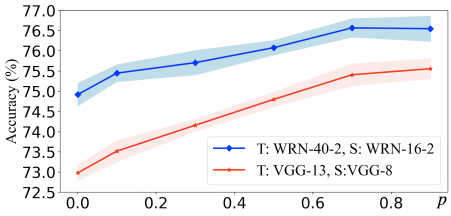


Fig. 3. Effects of p.

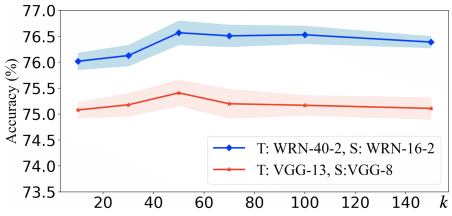


Fig. 4. Effects of update interval k.

enhance CSTs to match the performance of the teacher or largely reduce their performance gap.

6 Experiments

In this section, we aim to answer the following questions:
Q1: Ablation studies regarding MixPatch and the personalized education in PE.
Q2: Can PE enable small students to outperform large teachers or substantially reduce their performance gap?
Q3: Can PE outperform other similar data-driven approaches?
Q4: Is PE compatible with other SOTA distillation approaches?
Q5: Does PE indeed reduce teacher-student function shape differences?
Q6: What do the BKR images from the MixPatch prior region look like?

The experiments are conducted on the three widely used knowledge distillation benchmark datasets, i.e., CIFAR-100 [22], Tiny ImageNet [4], and ImageNet [6]. For a fair comparison, we adopt the architectures used in the existing literature (**we have shown that most of them are CSTs in Section 4.**) including ResNet [13], WRN [41], VGG [32], and ShuffleNet [26].

6.1 Answers to Q1

The ablation studies are conducted on CIFAR-100 with two teacher-student pairs of WRN-40-2 and WRN-16-2, and VGG-13 and VGG-8.

Effects of the Components in PE. PE uses the proposed MixPatch as the prior region, which is inspired by Mixup. We thus first show the superiority of MixPatch over Mixup and then show the effectiveness of (7) for spotting BKRs. As shown in Table 3, KD+MixPatch performs better than KD+Mixup on both pairs of teachers and students, which validates the superiority of MixPatch

[4] https://tiny-imagenet.herokuapp.com

Table 4. Comparison results on CIFAR-100 in terms of test accuracy (%). Underline denotes students matching or outperforming teachers.

Teacher (#Params)	WRN-40-2 (2.26M)	resnet-110 (1.74M)	VGG-13 (9.46M)	ResNet32×4 (7.43M)	ResNet50 (23.71M)	ResNet32×4 (7.43M)	WRN-40-2 (2.26M)
Student (#Params)	WRN-16-2 (0.70M)	ResNet-32 (0.47M)	VGG-8 (3.97M)	ResNet8×4 (1.23M)	VGG-8 (3.97M)	ShuffleNetV2 (1.36M)	ShuffleNetV1 (0.95M)
Teacher	76.46	74.31	75.38	79.63	79.10	79.63	76.46
Vanilla Student	73.64	71.14	70.68	72.51	70.68	73.12	70.77
KD	74.92±0.28	73.08±0.18	72.98±0.19	73.33±0.25	73.81±0.13	74.45±0.27	74.83±0.17
GANKD	75.05±0.30	72.87±0.35	73.09±0.28	73.46±0.33	73.68±0.27	74.61±0.29	74.88±0.22
Mixup	75.56±0.15	73.67±0.22	74.58±0.27	75.86±0.30	75.15±0.23	77.63±0.20	77.05±0.29
ActiveMixKD	75.60±0.35	73.21±0.28	74.50±0.29	75.98±0.19	75.13±0.25	77.22±0.17	77.01±0.20
PE(Ours)	**76.57±0.23**	**74.35±0.19**	**75.41±0.25**	**76.27±0.20**	**75.81±0.23**	**79.83±0.14**	**77.78±0.22**

over Mixup for assisting KD. Furthermore, personalized education (PE) further enhances the performances of KD+MixPatch significantly on both pairs, even beating the teachers, which indicates the effectiveness of Eq. (7) for spotting BKRs.

Effects of the Number of BKR Samples. PE can theoretically find infinite BKR samples to assist distillation. However, the diversity of them is also limited by the original data. We randomly use the BKR samples with probability p and the training samples with $1-p$ as the distillation data in each optimization step so that we can control the number of BKR samples. Figure 3 presents the effects of p on CIFAR-100. As expected, with the number of BKR samples increasing, the performance first increases and then becomes stable.

Effects of the Update Frequency. PE updates the BKR every k epochs in the training process. The effects of k are presented in Figure 4. It is observed that overall the performance of PE is not sensitive to k. However, when the BKR is updated too frequently (i.e., too small k), the student cannot have enough epochs to learn the current BKR, which leads to an inferior performance. Similarly, when the update frequency is too small, the performance of the student is also inferior, since not enough BKRs are found in the training process.

6.2 Answers to Q2 and Q3

We examine whether PE can indeed enable small students to outperform large teachers or substantially reduce their performance gap and compare it with the approaches along the same line. PE is a data-driven approach while the existing approaches focus on using different criteria to align the representations or logits of the teacher and the student. We thus design three data-driven baselines for reference (the comparison with non-data-driven SOTA approaches such as CRD [34] and SSKD [38] is reported in Appendix C):

(1) GANKD: As analysed in Section 4, only fitting the teacher outputs at sparse data points cannot enable students to well capture the in-distribution shape of the teacher function. One natural idea is to use generative adversarial networks (GANs) [11, 25] to learn the data distribution and then use the generator to generate fake data to assist distillation.

Table 5. Comparison results on Tiny ImageNet.

Teacher Student	WRN-40-2 WRN-40-1		VGG-13 VGG-8		WRN-40-2 VGG-8	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
Teacher	61.84	84.11	61.62	81.71	61.84	84.11
Vanilla student	55.39	79.87	55.46	78.15	55.46	78.15
KD	56.25±0.15	81.11±0.19	60.19±0.21	81.61±0.30	58.25±0.25	82.51±0.18
GANKD	55.71±0.26	80.13±0.16	60.32±0.11	81.90±0.23	58.49±0.35	82.73±0.20
MixupKD	56.80±0.28	81.29±0.15	61.64±0.15	82.81±0.13	59.16±0.19	82.10±0.10
ActiveMixKD	56.66±0.16	81.52±0.30	61.23±0.29	83.07±0.12	59.09±0.14	82.24±0.19
PE (Ours)	**58.74±0.17**	**82.59±0.13**	**62.40±0.29**	**83.57±0.16**	**59.91±0.18**	**82.98±0.07**

(2) MixupKD[5]: MixupKD uses Mixup [6] samples to assist distillation.
(3) ActiveMixKD: ActiveMixup [36] is originally proposed to train a classifier in data-few cases. It first uses Mixup to generate an image pool and then selects the images that the classifier has a low confidence on. ActiveMixKD adds these ActiveMixup samples to the distillation dataset.

CIFAR-100 The performances on CIFAR-100 are reported in Table 4. It is observed that there is an obvious performance gap between the students and the teachers in KD for different pairs. By enhancing each student on its own BKR, PE enables small students to **match or outperform** large teachers on five of the seven teacher-student pairs, and also substantially reduces the performance gap on the other two pairs. For example, the performance gap between teacher ResNet32×4 and student ResNet8×4 is reduced from 6.30 (KD) to 3.36 (PE). Note that there is no guarantee for PE to make small students match or outperform teachers as the BKR in PE cannot fully compensate for the unknown data sample distribution. Furthermore, PE also outperforms the data-driven competitors significantly. The reason for the superior performances of PE over GANKD is that modelling high-dimensional data distribution from sparse data points is even more challenging than training the classifier itself, which results in a nontrivial data discrepancy between the generated images and the real training images. This also causes that GANKD sometimes even underperforms KD. PE also beats ActiveMixKD significantly, since ActiveMixup ignores the critical samples that the student has a high confidence but with a wrong prediction and it also fails to interact with the teacher. In contrast, PE adaptively spots the BKRs for each student by interacting with the teacher. PE also outperforms Mixup due to the diverseness of MixPatch images in PE.

Tiny ImageNet We further evaluate PE on Tiny ImageNet. As shown in Table 5, PE beats all the competitors in terms of both Top-1 and Top-5 accuracies on

[5] CutMix performs almost the same with Mixup for assisting KD but is slower so that we simply adopt Mixup as the baseline here.
[6] The core code for this baseline is borrowed from this publicly accessible implementation: https://github.com/facebookresearch/mixup-cifar10.

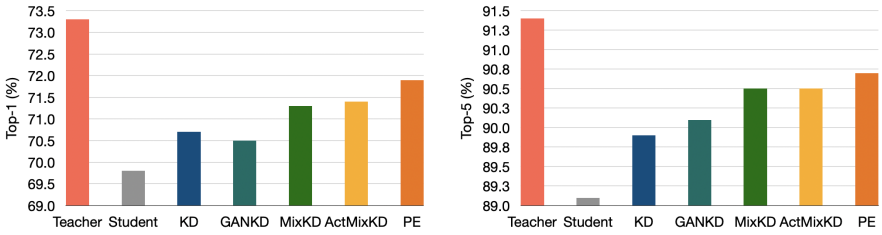


Fig. 5. Comparison on ImageNet

all the three pairs, and even enables the small student VGG-8 to significantly outperform the teacher VGG-13, which demonstrates the superiority of PE and the promise of addressing knowledge distillation from data perspectives.

ImageNet To investigate whether PE is applicable to large scale datasets, we conduct experiments on ImageNet. We follow CRD [34] to use ResNet-34 and ResNet-18 as the teacher and the student, respectively. As shown in Figure 5, by exploring the knowledge in the BKR, PE reduces the teacher-student performance gap significantly and also outperforms the competitors in terms of Top-1 and Top-5 accuracies, which demonstrates the applicability and usefulness of PE on large scale datasets. Nevertheless, the small student still underperforms the large teacher. Further examination reveals that ResNet-18 is an IST of ResNet-34 on the large and complex dataset ImageNet with ME 0.8, which indicates that the capacity difference can be the reason for the performance gap on ImageNet.

6.3 Answers to Q4

We further explore whether PE can be generalized to other SOTA distillation approaches inlcuding FitNet [31], AT [42], SP [35], CC [30], VID [2], RKD [28], PKT [29], AB [15], FT [20], NST [17], CRD [34], and SSKD [38]. The BKR searched by PE is used to enhance these SOTA approaches. We use WRN-40-2 and WRN-16-2 as the teacher and the student, respectively. Table 6 reports the results on CIFAR-100. It is observed that the enhanced counterparts (PE+Methods) consistently and substantially outperform the original methods even for the strong baselines like CRD and SSKD, which demonstrates the compatibility and usefulness of PE on different distillation approaches.

6.4 Answers to Q5

We further investigate whether students trained by PE indeed better capture the local, in-distribution shapes of the teachers than those trained by KD. The local shape of a function can be represented by a set of points (x, y) on the function graph where x is the input and y is the function output. To measure the shape difference, we report the average mean square **student-teacher output logit differences (S-T DIFs)** by using test data as inputs. As shown in Table 7, S-T DIFs of PE are consistently smaller than those of KD on all the seven pairs,

Table 6. Generalization (or Compatibility) of PE on SOTA distillation methods.

	FitNet	AT	SP	CC	VID	RKD	PKT	AB	FT	NST	CRD	SSKD
Methods	75.75	75.28	75.34	75.09	74.79	75.40	75.33	68.89	75.15	74.67	76.04	76.04
PE+Methods	**76.63**	**76.37**	**75.50**	**75.88**	**75.78**	**76.49**	**76.23**	**72.20**	**75.48**	**75.86**	**76.81**	**76.45**

Table 7. S-T DIFs (shape differences) on CIFAR-100

Teacher	WRN-40-2	resnet-110	VGG-13	ResNet32×4	ResNet-50	ResNet32×4	WRN-40-2
Student	WRN-16-2	resnet-32	VGG-8	ResNet8×4	VGG-8	ShuffleNetV2	ShuffleNetV1
KD	2.39	3.74	1.55	2.06	1.90	1.23	2.13
PE	1.74	2.92	1.15	1.53	1.37	0.77	1.52

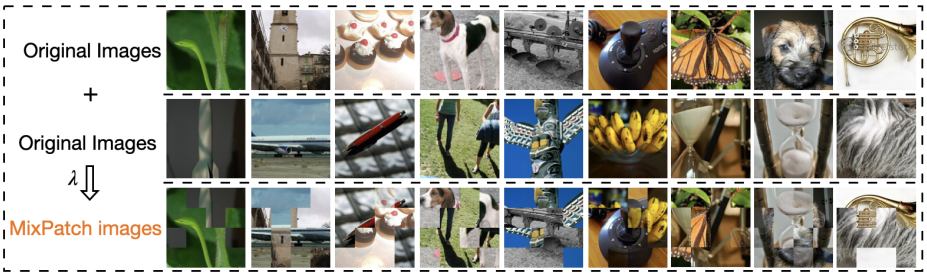


Fig. 6. MixPatch images from the BKR on ImageNet.

which validates that the student shapes of PE are closer to the teacher shapes and indicates that the BKRs in PE are indeed beneficial to the students for capturing the local shapes of the teacher functions.

6.5 Answers to Q6

PE adaptively learns the BKR from the MixPatch prior region for a student by interacting with the teacher. We show some BKR samples in Figure 6. It is observed that these samples share similar local patterns with the original images.

7 Conclusion

In this paper, we study why small students typically underperform large teachers in KD and how they can outperform large teachers. Through designing exploratory experiments with theoretical analysis, we find that model capacity differences are not necessarily the root reason and the distillation data matter when the student capacity is greater than a threshold. Inspired by this, we propose to our best knowledge the first personalized distillation approach PE that goes beyond in-distribution distillation and adaptively learns the blind knowledge region for each student through interacting with the teacher. Extensive experiments demonstrate that PE substantially reduces the student-teacher performance gap and even enables small students to outperform large teachers.

References

1. Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Fan, X., Guo, E.: Knowledge distillation from internal representations. arXiv preprint arXiv:1910.03723 (2019)
2. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9163–9171 (2019)
3. Chen, L., Wang, D., Gan, Z., Liu, J., Henao, R., Carin, L.: Wasserstein contrastive representation distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16296–16305 (2021)
4. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
5. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4794–4802 (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
7. Deng, X., Zhang, Z.: Comprehensive knowledge distillation with causal intervention. Advances in Neural Information Processing Systems **34** (2021)
8. Deng, X., Zhang, Z.: Graph-free knowledge distillation for graph neural networks. Proceedings of the 30th International Joint Conference on Artificial Intelligence (2021)
9. Deng, X., Zhang, Z.: Learning with retrospection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2021)
10. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT Press (2016)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
12. Han, X., Song, X., Yao, Y., Xu, X.S., Nie, L.: Neural compatibility modeling with probabilistic knowledge distillation. IEEE Transactions on Image Processing **29**, 871–882 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1921–1930 (2019)
15. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3779–3787 (2019)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
17. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017)
18. Huang, Z., Shen, X., Xing, J., Liu, T., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.S.: Revisiting knowledge distillation: An inheritance and exploration framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3579–3588 (2021)

19. Jing Yang, Brais Martinez, A.B., Tzimiropoulos, G.: Knowledge distillation vis softmax regression representation learning. In: International Conference on Learning Representations (2021)
20. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. In: Advances in Neural Information Processing Systems. pp. 2760–2769 (2018)
21. Koratana, A., Kang, D., Bailis, P., Zaharia, M.: Lit: Learned intermediate representation training for model compression. In: International Conference on Machine Learning. pp. 3509–3518 (2019)
22. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
24. Li, X., Li, S., Omar, B., Wu, F., Li, X.: Reskd: Residual-guided knowledge distillation. IEEE Transactions on Image Processing **30**, 4735–4746 (2021)
25. Liu, R., Fusi, N., Mackey, L.: Teacher-student compression with generative adversarial networks. arXiv preprint arXiv:1812.02271 (2018)
26. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 116–131 (2018)
27. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. AAAI Conference on Artificial Intelligence (2020)
28. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
29. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 268–284 (2018)
30. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5007–5016 (2019)
31. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (2015)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
33. Srinivas, S., Fleuret, F.: Knowledge transfer with Jacobian matching. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4723–4731. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018)
34. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020)
35. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1365–1374 (2019)
36. Wang, D., Li, Y., Wang, L., Gong, B.: Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1498–1507 (2020)

37. Wang, X., Zhang, R., Sun, Y., Qi, J.: Kdgan: Knowledge distillation with generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 775–786 (2018)
38. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European Conference on Computer Vision. pp. 588–604. Springer (2020)
39. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4133–4141 (2017)
40. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
41. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
42. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
43. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1Ddp1-Rb
44. Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., Zhang, Q.: Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. In: International Conference on Learning Representations (2021)
45. Zhu, J., Tang, S., Chen, D., Yu, S., Liu, Y., Rong, M., Yang, A., Wang, X.: Complementary relation contrastive distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9260–9269 (2021)