Appendix for 'Personalized Education: Blind Knowledge Distillation'

A Theoretical Proof of Proposition 1

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.*

Theoretical Proof: The empirical proof has been given in the paper. We now provide the theoretical proof below. The goal of KD is to find a student function $S_\theta \in \mathcal{F}$ to approximate a powerful teacher T on a target data distribution $P(x)$. The approximation risk is written as:

$$R(S_\theta, T, P(x)) = \mathop{\mathbb{E}}_{x \sim P(x)} \ell(T(x), S_\theta(x)) \tag{1}$$

where ℓ is a loss function such as KL-divergence or mean square error.

Directly minimizing $R(S_\theta, T, P(x))$ is typically impossible as $P(x)$ is unknown. KD instead minimizes the empirical risk over a training dataset $X_t = \{x_i\}_{i=1}^n$ drawn from $P(x)$:

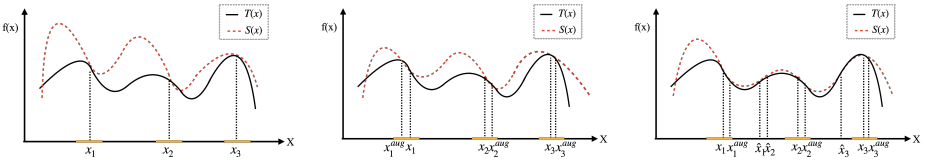
$$\hat{R}(S_\theta, T, X_t) = \frac{1}{n} \sum_{i=1}^{n} \ell(T(x_i), S_\theta(x_i)) \tag{2}$$

Theorem 1. *When ℓ is bounded, we have the following bound with probability at least 1-δ:*

$$R(S_\theta, T, P(x)) \leq \hat{R}(S_\theta, T, X_t) + \sqrt{\frac{C + \log\frac{2}{\delta}}{2n}} \tag{3}$$

where C denotes the complexity of $\mathcal{F}$. This bound can be obtained from ERM [11, 5].

We next connect CSTs and ISTs with the theoretical error bound (3). CSTs have enough capacities to memorize the outputs of the teachers so that they can achieve a small empirical risk $\hat{R}(S_\theta, T, X_t)$. In this case, as seen from (3), if we increase the number (i.e., n) of distillation data points, the bound can be further decreased and even achieves nearly 0. This is consistent with the empirical results in the simulation experiments. Therefore, for CSTs, the capacity differences are not necessarily the root reason and instead the distillation data matter. In contrast, the capacities of ISTs are too small to memorize the outputs of the teachers, which leads to a large empirical risk $\hat{R}(S_\theta, T, X_t)$. In this case, the $\hat{R}(S_\theta, T, X_t)$ is dominant in the bound (3) and thus further increasing the



(a) KD with training data (b) KD with regular data augmentation (c) KD with OOD assistance

Fig. 1. An intuitive illustration of OOD assistance on 2-D space: Yellow regions denote the real data distribution $P(x)$; x_i denotes training samples; x_i^{aug} denotes augmented samples; $\hat{x}_i$ denotes OOD samples.

number of distillation data may not lead to a smaller bound and even may further cause a larger empirical error $\hat{R}(S_\theta, T, X_t)$. Therefor, for ISTs, capacity differences cause the performance gap. By connecting CSTs, ISTs, and the bound (3), we theoretically prove Proposition 1.

B Theoretical Proof of Proposition 2

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

Theoretical Proof: When OOD samples X_o are used to assist distillation, the distillation dataset X_d is changed from X_t to $X_t \cup X_o$. $X_t \cup X_o$ follows an unknown distribution $Q(x)$. $Q(x)$ is different from the real data distribution $P(x)$. Thus, we have:

Theorem 2. *When ℓ is bounded, we have the following bound with probability at least 1-δ:*

$$R(S_\theta, T, P(x)) \leq \hat{R}(S_\theta, T, X_t \cup X_o) + \sqrt{\frac{C + \log\frac{2}{\delta}}{2m}} + D(P, Q) \tag{4}$$

where m is the total number of the samples in $X_t \cup X_o$; $D(P,Q) = \int \ell\left(T(x), S_\theta(x)\right)\left(P(x) - Q(x)\right) dx$.

As seen from (4), when the OOD samples are far from $P(x)$, $D(P,Q)$ is large, which theoretically explains why not all samples are useful. On the contrary, when the OOD samples are not far from $P(x)$, $D(P,Q)$ is small and $\hat{R}$ is also small for CSTs. In this case, when $m > \frac{C+\log\frac{2}{\delta}}{2(\mathcal{B}_p - D(P,Q))^2}$ where $\mathcal{B}_p$ denotes the error bound by only using X_t, the OOD samples are beneficial to distillation. We thus theoretically prove Proposition 2.

We also provide an intuitive example in a 2D space about the usefulness of OOD distillation in Figure 1. As shown in Figure 1(a), even if student S perfectly fits teacher T at each training data point, i.e., x_1, x_2, and x_3, their local

shapes near these samples can still be highly different. To mitigate this issue, the typically used strategy is the regular data augmentation (e.g. padding and cropping) formulated by the *Vicinal Risk Minimization* [2] principle. Nevertheless, it has a severe limitation that a newly generated data point is very close to the original training data point, since they contain almost the identical objective only with different backgrounds caused by padding or cropping. Due to this limitation, as shown in Figure 1(b), even if the student S fits the teacher T at all the training data points plus the augmented data points (i.e., x_1^{aug}, x_2^{aug}, and x_3^{aug}), their local shapes can still differ substantially. We thus propose to go beyond in-distribution distillation by enforcing students to mimic the behavior of teachers on OOD samples. As shown in Figure 1(c), OOD samples (i.e., $\hat{x}$) can assist the student to better capture the local, in-distribution shape of the teacher function.

C Comparison with Both non-Data-Driven and Data-Driven SOTA Approaches

PE is a data-driven approach, which is along a different line from the existing state-of-the-art (SOTA) distillation approaches. The existing SOTA approaches are not data-driven as they focus on using different criteria to align the representations or the logits between students and teachers. We have shown in the paper that PE is compatible with these approaches and enhances their performances substantially. In this part, we compare PE with these SOTA approaches including FitNet [8], AT [13], SP [10], VID [1], RKD [6], PKT [7], AB [3], FT [4], CRD [9], and SSKD [12]. All these SOTA approaches use the standard data augmentation strategy, except for SSKD using strong data augmentation in the extra self-supervised task.

The comparison results with SOTA approaches on CIFAR-100 are reported in Table 1. It is observed that PE betas these SOTA approaches significantly. Note that although SSKD uses extra tasks (i.e., self-supervised learning tasks) and extra data (i.e., strong data augmentation in self-supervised learning tasks), PE is still able to outperform it. These observations demonstrate the superiority of PE and the promise of addressing knowledge distillation from data perspectives.

The comparison results on Tiny ImageNet are reported in Table 2. It is observed that PE also performs much better than these SOTA approaches, which demonstrates the effectiveness of PE.

We further compare PE with SOTA approaches on large dataset ImageNet, where we follow the existing literature to adopt ResNet34 and ResNet18 as the teacher and the student, respectively. As shown in Table 3, PE also beats these SOTA approaches on ImageNet significantly, which demonstrates the applicability and usefulness of PE on large datasets.

Table 1. 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**
FitNet	75.75	71.06	73.54	74.31	73.84	75.11	75.55
AT	75.28	72.31	73.62	74.26	73.45	75.30	75.61
SP	75.34	72.69	73.44	74.74	73.86	75.15	75.56
VID	74.79	72.61	73.96	74.82	73.75	75.78	75.36
RKD	75.40	71.82	73.72	74.46	73.73	75.74	75.45
PKT	76.01	72.61	73.37	74.17	73.53	75.18	75.51
AB	68.89	70.98	74.27	74.45	74.20	75.66	76.58
FT	75.15	72.37	73.42	75.02	73.58	74.95	75.18
CRD	76.04	73.75	74.06	75.90	74.42	75.72	75.96
SSKD	76.04	73.60	75.33	76.20	75.76	78.61	77.40

Table 2. 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	59.09±0.19	82.24±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**
FitNet	55.41±0.31	79.75±0.40	55.26±0.20	78.70±0.44	55.85±0.14	78.39±0.16
AT	55.84±0.41	80.03±0.18	56.82±0.46	80.35±0.37	56.62±0.19	79.27±0.12
SP	54.09±0.26	79.26±0.36	56.99±0.42	79.87±0.16	58.25±0.41	81.90±0.35
CC	55.10±0.43	79.05±0.30	54.14±0.19	77.45±0.23	54.61±0.30	78.27±0.19
VID	56.07±0.23	80.47±0.22	54.57±0.26	77.59±0.23	55.79±0.18	78.90±0.29
RKD	55.37±0.29	79.94±0.19	56.60±0.13	78.33±0.39	56.35±0.24	78.83±0.27
PKT	56.31±0.22	81.20±0.27	56.36±0.17	78.98±0.21	56.75±0.15	79.31±0.13
CRD	56.75±0.33	81.64±0.19	59.95±0.23	81.33±0.31	58.80±0.39	82.05±0.23
SSKD	56.90±0.37	81.75±0.21	59.87±0.27	81.15±0.10	58.89±0.33	81.95±0.17

D Implementation Details and Datasets

We first introduce the adopted benchmark datasets and the standard data augmentation on each dataset as follows:

Table 3. Comparison results on ImageNet with SOTA approaches.

	Teacher	Student	KD	AT	SP	CC	SSKD	CRD	PE(ours)
Top-1 (%)	73.3	69.8	70.7	70.7	70.2	70.0	71.6	71.4	**71.9**
Top-5 (%)	91.4	89.1	89.9	90.0	89.8	89.2	90.7	90.5	**90.7**

CIFAR-100: CIFAR-100 is an image classification dataset with 100 classes, containing 50,000 training images and 10,000 test images with image size 32 $\times$ 32 in the RGB space. The standard data augmentation on CIFAR datasets is as follows: during training time, 4 pixels are padded on each side of an image and then are randomly flipped horizontally; finally the image is randomly cropped to 32 $\times$ 32 size.

Tiny ImageNet: Tiny ImageNet i.e., a subset of ImageNet, is an image classification dataset with 200 classes, containing 100,000 training images and 10,000 test images with size 64 $\times$ 64 in the RGB space. At training time, 8 pixels are padded on each side of an image and then are randomly flipped horizontally; finally the image is randomly cropped to 64 $\times$ 64 size.

ImageNet: ImageNet is a large-scale image classification dataset with 1000 classes, containing 1.28 million training images and 50,000 validation images with different sizes in the RGB space. On ImageNet, we use the standard scale and aspect ratio augmentation strategy. Test images are resized so that the shorter side is set to 256, and then are cropped to size 224 $\times$ 224.

On the exploratory experiments, the architectures of SN2 and SN3 are: Conv(128)-BN-AvgPooling(32)-FC and Conv(128)-BN-ReLU-Covn(256)-BN -ReLU-AvgPooling(16)-FC, respectively.

We set α, β, and τ to 0.1, 0.9, and 4, respectively, on all the datasets except on ImageNet where we follow the existing literature to set $\alpha = 1$ and $\tau = 2$. Probability p of using BKR samples is set to 0.5 on ImageNet and 0.7 or 0.9 on the other two datasets. The other hyper-parameters can be found in Github link: https://github.com/Xiang-Deng-DL/PEBKD

References

1. 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)
2. Chapelle, O., Weston, J., Bottou, L., Vapnik, V.: Vicinal risk minimization. In: Advances in neural information processing systems. pp. 416–422 (2001)
3. 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)
4. 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)
5. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018)
6. 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)
7. 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)
8. 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)
9. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2020)
10. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1365–1374 (2019)
11. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
12. 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)
13. 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)