What Role Does Data Augmentation Play in Knowledge Distillation?

Wei Li[1] (✉), Shitong Shao[1], Weiyan Liu[1], Ziming Qiu[1], Zhihao Zhu[1], and Wei Huan[1]

School of Instrument Science and Engineering, Southeast University, Nanjing, Jiangsu 210096, China
{li-wei, shaoshitong, liuweiyan, qiuziming, zhuzhihao, huan-wei}@seu.edu.cn

Abstract. Knowledge distillation is an effective way to transfer knowledge from a large model to a small model, which can significantly improve the performance of the small model. In recent years, some contrastive learning-based knowledge distillation methods (i.e., SSKD and HSAKD) have achieved excellent performance by utilizing data augmentation. However, the worth of data augmentation has always been overlooked by researchers in knowledge distillation, and no work analyzes its role in particular detail. To fix this gap, we analyze the effect of data augmentation on knowledge distillation from a multi-sided perspective. In particular, we demonstrate the following properties of data augmentation: **(a)** data augmentation can effectively help knowledge distillation work even if the teacher model does not have the information about augmented samples, and our proposed diverse and rich **J**oint **D**ata **A**ugmentation (JDA) is more valid than single *rotating* in knowledge distillation; **(b)** using diverse and rich augmented samples to assist the teacher model in training can improve its performance, but not the performance of the student model; **(c)** the student model can achieve excellent performance when the proportion of augmented samples is within a suitable range; **(d)** data augmentation enables knowledge distillation to work better in a few-shot scenario; **(e)** data augmentation is seamlessly compatible with some knowledge distillation methods and can potentially further improve their performance. Enlightened by the above analysis, we propose a method named **C**osine **C**onfidence **D**istillation (CCD) to transfer the augmented samples' knowledge more reasonably. And CCD achieves better performance than the latest SOTA HSAKD with fewer storage requirements on CIFAR-100 and ImageNet-1k. *Our code is released at* `https://github.com/shaoshitong/CCD`*.*

1 Introduction

With the vigorous development of deep learning, numerous excellent models (e.g., ResNet [14], ShuffleNet [40], ViT [7]) have been proposed. In this trend, the evaluation metrics of some image upstream and downstream tasks have been greatly improved [26,34,20,21,11]. For instance, ResNet50 has achieved 77.15% on the ImageNet-1k [27] classification task in 2015, and VIT-H/14 has achieved 88.55% accuracy on the same task in 2020. However, ResNet50 has only 25.5 million parameters, compared to 632 million parameters of VIT-H/14. The massive storage requirements of the large models render them to deploy in real-time applications challengingly. For the purpose of developing efficient models, knowledge distillation [15], as an effective technique, has been widely used in model compression [1]. To be specific, knowledge distillation aims to transfer knowledge from a pre-trained teacher network with big-scale parameters to a lightweight student network. This significant training technique generally enables the student model to outperform traditional training techniques by a large margin. Commonly, the cases where the teacher model is fixed and not fixed are referred to as offline knowledge distillation [37,32,42] and online knowledge distillation [10,2,41], respectively. And online knowledge distillation can improve the performance of student models more effectively than offline knowledge distillation but requires more computational and storage costs [9]. To make the design choices that other researchers can better apply, we choose offline knowledge distillation as the standard for the study in this work.

Contrastive learning aims to encode the correlations between a sample pair $(\mathcal{X}_i, \mathcal{X}_j)$. Specifically, if $\mathcal{X}_i$ and $\mathcal{X}_j$ are similar, contrastive learning makes the distance between them close; otherwise, makes the distance between them as far as possible. In recent years, contrastive learning has been considered as an effective solution in the self-supervised domain. The popular contrastive learning methods, such as MOCO [13] and SimCLR [3], have been widely recognized and applied by related researchers. It is worth noting that, contrastive learning has also been applied in knowledge distillation as a novel way to transfer knowledge. The knowledge distillation methods, SSKD [35] and HSAKD [4], based on contrastive learning and self-supervised representational learning, utilize the same data augmentation (i.e., rotations $\{0°, 90°, 180°, 270°\}$). And the above methods are state-of-the-art (SOTA) in 2020 and 2021, respectively. However, the phenomenon that data augmentation changes the information of the training samples is not discussed in their works. In addition, because SSKD and HSAKD use *rotating* in training to duplicate the samples, their steps n_{step} (defined in Equation 1, where n_{iter}, n_{bs} and n_{epoch} refer to the number of batch sizes in an epoch, batch size, and the number of epochs, respectively) were four times higher than vanilla Knowledge Distillation (vanilla KD) in their comparative experiments. Based on the above analysis, we know that using data augmentation in SSKD and HSAKD changes the information of the training samples and n_{step}. So it is unclear that whether the performance improvement of SSKD and HSAKD is brought by data augmentation. Due to these reasons, we urgently need to analyze the role that data augmentation plays in knowledge distillation.

$$n_{step} = n_{iter} \times n_{bs} \times n_{epoch} \tag{1}$$

Some works such as [8], [33] and [6] have noted the role of data augmentation in knowledge distillation. They find that the knowledge distillation approaches, including vanilla KD and CRD [31], can also improve the performance of the student model to a certain extent by means of only data augmentation. However, the above work only discusses the role of data augmentation in knowledge distillation from a one-sided perspective. So they do not consider the impact of multiple factors, including different n_{step}, the diversity of data augmentation, the proportion of augmented samples in the all training samples, and the few-shot scenario, on the performance of student models. In order to fill this gap, we further evaluate the effect of data augmentation on knowledge distillation under different factors, and the main conclusions we find in our extensive experiments can be summarized as follows:

- Data Augmentation is effective in knowledge distillation, even though the teacher model has no information about the augmented sample. The effectiveness of self-supervised methods (i.e., SSKD and HSAKD) can be attributed to *rotating*, to a certain extent.
- By increasing n_{step}, knowledge distillation both with and w/o (i.e., without) data augmentation will improve the performance of the student model. In addition, the knowledge distillation with diverse and rich data augmentation is more valid than single rotational knowledge distillation.
- Transferring knowledge only from the augmented samples doesn't necessarily work, but transferring knowledge from both the original and augmented samples can effectively make the augmented samples work.
- Using diverse and rich augmented samples to assist the teacher model in training can improve its performance, but not the performance of the student model.
- The student model can achieve excellent performance when the proportion of augmented samples is within a suitable range. But too many augmented samples will lead to a drop in performance.
- Data augmentation enables knowledge distillation to work better in few-shot scenarios.
- Data augmentation is seamlessly compatible with some knowledge distillation methods and can potentially further improve their performance.

Inspired by these conclusions, we propose a method named **C**osine **C**onfidence **D**istillation (CCD) to transfer the probabilistic knowledge of augmented samples more reasonably. And CCD achieves better performance compared to the latest SOTA HSAKD with fewer storage requirements on CIFAR-100 and ImageNet-1k.

2 Related Work

Knowledge Distillation with Self-Supervision. Recently, knowledge distillation methods (i.e., SSKD and HSAKD) with self-supervision have achieved state-of-the-art on both CIFAR-100 and ImageNet-1k. Among them, SSKD and HSAKD adopt the idea of contrastive learning directly and indirectly, respectively. Specifically, SSKD tends to compute the sample-based metric matrices for the teacher and student models separately and align them to achieve self-supervision and knowledge distillation. And HSAKD generates bivariate distribution labels based on augmented and semantic categories, then minimizes the loss between logits of the teacher and student models. The core idea of the above two methods is to transfer the information learned by the teacher model through self-supervision to the student model, which needs to be implemented through data augmentation. Therefore, it is clear that data augmentation has become an essential part of self-supervised knowledge distillation [28,30,19].

Data Augmentation. In the field of computer vision, data augmentation is a simple and effective way to improve model performance [16,18,5,36,12]. For instance, data augmentation rules such as *rotating*, shear, and contrast are widely applied in visual tasks (e.g., object classification and object detection) and avoid the overfitting problem of the model to a certain extent [22,29]. If data augmentation really works, SSKD and HSAKD only use *rotating* to augment data, which obviously lacks variety. So for our study, We turn our attention to other more diverse data augmentation. AutoAugment [5], an effective and popular data augmentation method, adopts 16 commonly used data augmentation rules as its sub-policies. These sub-policies efficiently augment the dataset by searching for their optimal hyperparameters through reinforcement learning (RL). However, for knowledge distillation, we do not know what individualized samples are adapted for a particular pair of teacher and student models. And finding the optimal conversion probability again requires a lot of training costs. Therefore, we will apply 14 data augmentation (i.e., the sub-policies in AutoAugment) to convert the original samples in random order and with the same probability in our study.

In a word, inspired by SSKD and HSAKD applying data augmentation for their self-supervised methods, we employ a wide and abundant variety of data augmentation rules to conduct our research.

3 Contributions

For all experimental results shown in this Sec. 3.1 to this Sec. 3.5, we conduct evaluations on the standard CIFAR-10 [17] benchmark across the ResNet56-ResNet20 [14] pair and the standard CIFAR-100 [17] benchmark across the WRN-40-2-WRN-16-2 [38] pair. Note that the top-1 test accuracy of teachers ResNet-56 and WRN-40-2 are 93.56% and 76.44%, respectively. And all black horizontal lines in figures of this paper represent the test accuracy of the teacher model. All "×**number**" in this paper represents a multiple of the increase about n_{step} compared to the benchmark, and more detailed benchmark settings can be found in Appendix **??**. Besides, we utilize the standard training settings following [35,4] on CIFAR-100 and following [23] on CIFAR-10. *All teacher models do not utilize the augmented samples for representation learning, unless otherwise specified in this paper*. To get plausible results, we report the mean test accuracy with 3 runs. Note that to introduce our research more logically, we will present our contributions following the form of progressive exploration.

3.1 Inspired by SSKD and HSAKD

By regarding the similarity between self-supervised samples as the transferring knowledge, SSKD has achieved excellent performance. But when investigating previous work, we find that SSKD is sensitive in certain scenes. For example, the validation accuracy of ResNet-18 (student) trained on ImageNet-1k with ResNet-34 (teacher), obtained by the original work, is 71.62% [35]. Still, the validation accuracy got by work [23] is 70.09%. There is a 1.53 percentage point difference between the above two results, which is relatively large under the same hyperparameter configuration. Meanwhile, the validation accuracy for vanilla KD got by work [23] is 71.23% and got by us is 71.16%. Intuitively, SSKD is inferior to vanilla KD under certain circumstances, which drives us to rethink the effectiveness of SSKD.

Table 1: These two tables show the experimental results of decoupling of SSKD and HSAKD on CIFAR-10 and CIFAR-100, respectively. Where “Baseline” stands for training using only vanilla KD. In addition, “(number)” refer to the increased validation accuracy compared to the baseline.

CIFAR-10

n_{step}	Methods	Options	Acc.(%)
×2	Baseline	$\mathcal{L}_{ce} + \mathcal{L}_{kd}$	93.24
×2	SSKD	L_{SSKD}	$92.56_{(-0.68)}$
×2	SSKD	$L_{SSKD} - L_{ss}$	$92.72_{(-0.52)}$
×2	HSAKD	L_{HSAKD}	$93.22_{(-0.02)}$
×2	HSAKD	$L_{HSAKD} - L_{kl\_q}$	$92.68_{(-0.56)}$
×4	Baseline	$\mathcal{L}_{ce} + \mathcal{L}_{kd}$	93.47
×4	SSKD	L_{SSKD}	$92.73_{(-0.74)}$
×4	SSKD	$L_{SSKD} - L_{ss}$	$92.73_{(-0.74)}$
×4	HSAKD	L_{HSAKD}	$93.46_{(-0.01)}$
×4	HSAKD	$L_{HSAKD} - L_{kl\_q}$	$92.88_{(-0.59)}$

CIFAR-100

n_{step}	Methods	Options	Acc.(%)
×2	Baseline	$\mathcal{L}_{ce} + \mathcal{L}_{kd}$	74.72
×2	SSKD	L_{SSKD}	$75.51_{(+0.79)}$
×2	SSKD	$L_{SSKD} - L_{ss}$	$75.31_{(+0.59)}$
×2	HSAKD	L_{HSAKD}	$76.73_{(+2.01)}$
×2	HSAKD	$L_{HSAKD} - L_{kl\_q}$	$75.55_{(+0.83)}$
×4	Baseline	$\mathcal{L}_{ce} + \mathcal{L}_{kd}$	74.87
×4	SSKD	L_{SSKD}	$76.16_{(+1.29)}$
×4	SSKD	$L_{SSKD} - L_{ss}$	$76.31_{(+1.44)}$
×4	HSAKD	L_{HSAKD}	$77.20_{(+2.33)}$
×4	HSAKD	$L_{HSAKD} - L_{kl\_q}$	$76.05_{(+1.18)}$

A natural direction to tackle this problem is to decouple SSKD and HSAKD. To this end, we first give the standard cross-entropy (CE) loss of original samples in Equation 2:

$$\begin{aligned}\mathbf{p}^S(\mathbf{x};\tau) &= \textbf{softmax}\left(f^S(\mathbf{x})/\tau\right),\\ \mathcal{L}_{ce} &= \mathbb{E}_{\mathbf{x}\in\mathcal{X}}\,\textbf{CE}\left(\mathbf{p}^S(\mathbf{x};1),\mathbf{y}\right),\end{aligned} \tag{2}$$

where τ, $f^S(\cdot)$, $\textbf{CE}(\cdot,\cdot)$, $\mathcal{X}$ and $\mathbf{y}$ refer to the temperature hyperparameter, the student backbone network, the cross entropy loss function, the original sample set and the hard label about $\mathbf{x}$, respectively. Then we denote the vanilla KD loss of original samples as Equation 3.

$$\begin{aligned}\mathbf{p}^T(\mathbf{x};\tau) &= \textbf{softmax}\left(f^T(\mathbf{x})/\tau\right),\\ \mathcal{L}_{kd} &= \tau^2\mathbb{E}_{\mathbf{x}\in\mathcal{X}}\,\textbf{KL}\left(\mathbf{p}^T(\mathbf{x};\tau)\,\|\mathbf{p}^S(\mathbf{x};\tau)\right),\end{aligned} \tag{3}$$

where $f^T(\cdot)$ denotes the teacher backbone network and $\textbf{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. Thereby, the vanilla KD loss of augmented samples can be defined as follows:

$$\mathcal{L}_T = \tau^2\mathbb{E}_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}}\,\textbf{KL}\left(\mathbf{p}^T(\widetilde{\mathbf{x}};\tau)\,\|\mathbf{p}^S(\widetilde{\mathbf{x}};\tau)\right), \tag{4}$$

where $\widetilde{\mathcal{X}}$ stands for the augmented sample set. Hence, we give the corresponding Equations 5 for SSKD and HSAKD:

$$\begin{aligned}\mathcal{L}_{SSKD} &= \lambda_1 * \mathcal{L}_{ce} + \lambda_2 * \mathcal{L}_{kd} + \lambda_3 * \mathcal{L}_{ss} + \lambda_4 * \mathcal{L}_T,\\ \mathcal{L}_{HSAKD} &= \mathcal{L}_{ce} + \mathcal{L}_{kl\_q} + \mathcal{L}_{kl\_p}.\end{aligned} \tag{5}$$

In $\mathcal{L}_{SSKD}$, λ_1, λ_2, λ_3 and λ_4 are the balancing weights. And the sub-task loss function $\mathcal{L}_{ss}$ concurs with L_{ss} in [35]. In $\mathcal{L}_{HSAKD}$, $\mathcal{L}_{kl\_q}$ (feature-based) and $\mathcal{L}_{kl\_p}$ (response-based) are the loss functions as mentioned in the work of [4][1]. Although $\mathcal{L}_{SSKD}$ and

[1] For the sake of simplicity, $\mathcal{L}_{kl\_q}$ and $\mathcal{L}_{kl\_p}$ here have an additional process of calculating mathematical expectations compared to the original paper.

$\mathcal{L}_{HSAKD}$ are not similar in form, by Equation 6 we can rewrite the form of $\mathcal{L}_{kl\_p}$.

$$\begin{aligned}\mathcal{L}_{kl\_p} &= \tau^2 \mathbb{E}_{\mathbf{x}\in\mathcal{X}} \frac{1}{M} \sum_{j=1}^{M} \mathbf{KL}\left(\mathbf{p}^T\left(t_j\left(\mathbf{x}\right);\tau\right) \|\mathbf{p}^S\left(t_j\left(\mathbf{x}\right);\tau\right)\right), \\ &= \tau^2 \mathbb{E}_{\mathbf{x}\in\mathcal{X}} \frac{1}{M} \mathbf{KL}\left(\mathbf{p}^T\left(\mathbf{x};\tau\right) \|\mathbf{p}^S\left(\mathbf{x}\right);\tau\right) + \tau^2 \mathbb{E}_{\widetilde{\mathbf{x}}\in\widetilde{\mathcal{X}}} \frac{M-1}{M} \mathbf{KL}\left(\mathbf{p}^T\left(\widetilde{\mathbf{x}};\tau\right) \|\mathbf{p}^S\left(\widetilde{\mathbf{x}};\tau\right)\right), \\ &= \frac{1}{M}\mathcal{L}_{kd} + \frac{M-1}{M}\mathcal{L}_T,\end{aligned} \tag{6}$$

where M and $\{t_j\left(\cdot\right)\}_{j=1}^{M}$ refer to the number of rotation operators (i.e., rotations $\{0°, 90°$ $, 180°, 270°\}$) and a set of data augmentation operators, respectively. According to the code and original paper provided by the author, we can default M to 4. Furthermore, we find that λ_2/λ_4 is $1/3$ in SSKD-related codes. Then an obvious conclusion is that $\mathcal{L}_{kl\_p} \propto \lambda_2 * \mathcal{L}_{kd} + \lambda_4 * \mathcal{L}_T$.

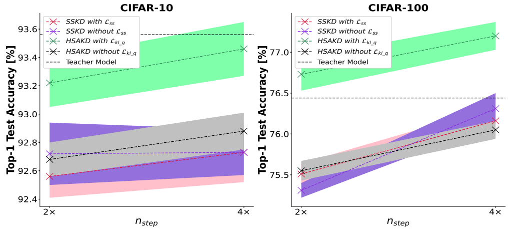

Fig. 1: This line graph is used to clearly demonstrate the roles $\mathcal{L}_{kl\_q}$ and $\mathcal{L}_{ss}$ play in knowledge distillation, and the shaded area in this figure indicates the standard deviation. Of note is that, $\mathcal{L}_{ss}$ does not play a positive role. In contrast, $\mathcal{L}_{kl\_q}$ can significantly improve the performance of the student model.

The above analysis shows that the essential difference between SSKD and HSAKD is that SSKD transfers the correlation information between samples after the global average pooling (GAP) layer ($\mathcal{L}_{ss}$). In contrast, HSAKD transfers the self-supervised augmented distribution information of the outputs of middle layers ($\mathcal{L}_{kl\_q}$).

For the purpose of verifying the role of $\mathcal{L}_{ss}$, $\mathcal{L}_{kl\_q}$ and *rotating*, we conduct the decoupling experiment on the two benchmarks mentioned above. For a fair comparison, we quadruple the baseline's n_{epoch} to ensure its n_{step} being the same as the other methods. Then, we get the results shown in Tab. 1 and Fig. 1. The results displayed in Fig. 1 show that ***it is effective to transfer the self-supervised augmented distribution information of middle layers' outputs***. On the contrary, transferring the correlation information of GAP's output is ineffective. Meanwhile, the analysis in Tab. 1 shows that SSKD without $\mathcal{L}_{ss}$ and HSAKD without $\mathcal{L}_{kl\_q}$ (i.e., $\lambda_1 * \mathcal{L}_{ce} + \lambda_2 * \mathcal{L}_{kd} + \lambda_4 * \mathcal{L}_T$ and $\mathcal{L}_{ce} + \mathcal{L}_{kl\_p}$) have a slight decrease in performance compared to the baseline on CIFAR-10. Still, there is a manifest improvement in the performance on CIFAR-100 compared with the baseline. Intuitively, the common part of SSKD and HSAKD is the traditional CE loss on the original training samples and the vanilla KD loss on both original and rotated samples. Just relying on this common part, the student model has a significant performance gain on CIFAR-100 compared with baseline. Thus we can illustrate that ***rotating is effective in knowledge distillation***.

From the results presented in Tab. 1, we find that adding the rotated samples into CIFAR-10 will cause performance damage during the training process. We argue that this phenomenon is due to the rotation operator itself. In follow-up experiments, we demonstrate that utilizing more diverse data augmentation operators is equally effective on CIFAR-10.

3.2 The Role of Data Augmentation

Although in Sec. 3.1 we have bespoken that *rotating* is beneficial to knowledge distillation. However, only utilizing *rotating* as data augmentation in training lacks diversity, so it cannot demonstrate that other data augmentation is also effective. Similarly, in the work [33], only CutMix [36] and Mixup [39] are discussed. To ensure the diversity of data augmentation, we propose the **J**oint **D**ata **A**ugmentation (JDA), which is composed of cascaded sub-policies. Assuming that we define N various sub-policies $\{\mathbf{sp}_i(\cdot)\}_{i=1}^{N}$. And we also define a Bernoulli operator $g(f;q)$ as shown in Equation 7.

$$g(f;q) = \begin{cases} f(\cdot) & ,w.p.\ q \quad , \\ \textbf{identity}(\cdot) & ,w.p.\ 1-q \quad , \end{cases} \tag{7}$$

identity$(\cdot)$ refers to the identity transformation, i.e. **identity**$(x) = x$. Then for an original sample $\mathbf{x} \in \mathcal{X}$, we can denote its augmented sample $\widetilde{\mathbf{x}}$ in Equation 8.

$$\widetilde{\mathbf{x}} = g(\mathbf{sp}_N;q) \circ g(\mathbf{sp}_{N-1};q) \cdots g(\mathbf{sp}_2;q) \circ g(\mathbf{sp}_1;q)(\mathbf{x}), \tag{8}$$

where $\circ$ denotes composition, and note that all sub-policies in Equation 8 have the same probability of occurrence. This approach can ensure that all sub-policies have similar effects on the original sample and provide the convenience for related experiments in this paper. Moreover, JDA not only is easy to be set up but also guarantees a huge difference compared to the optimal hyperparameters searched by AutoAugment. JDA also eliminates the possibility that the data augmentation work is a result of AutoAugment's hyperparameter search. In particulat, as demonstrated in Appendix **??**, JDA is more effective than AutoAugment because it can perform richer and more varied transformation on one single image. In more detail, q is set to 0.5 by default unless otherwise specified in our experiments. We consider choosing 14 sub-policies in AutoAugment for JDA, and the detailed hyperparameter settings for 14 sub-policies can be apparent from Appendix **??**. Furthermore, we introduce the mini-batch component of model training to explain how our data augmentation works. For the original sample set $\mathcal{X}$ (i.e., the original mini-batch), our proposed JDA transforms all elements in it in turn and composes a new augmented sample set $\widetilde{\mathcal{X}}$. Then the new mini-batch composed of $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ serves as the real input of the model.

Table 2: **Left:** *CIFAR-10*. **Right:** *CIFAR-100*. **KD+JDA:** *vanilla KD+joint data augmentation*. The numbers and numerical subscripts in the table represent the test accuracy and standard deviation, respectively.

SSKD	HSAKD	KD+JDA	SSKD	HSAKD	KD+JDA
$92.73_{(\pm 0.16)}$	$93.46_{(\pm 0.19)}$	$93.51_{(\pm 0.14)}$	$76.17_{(\pm 0.17)}$	$77.20_{(\pm 0.17)}$	$76.86_{(\pm 0.16)}$

In SSKD and HSAKD, hard labels for classification, which are provided by augmented samples, are not applied to supervise the student model in training. SSKD

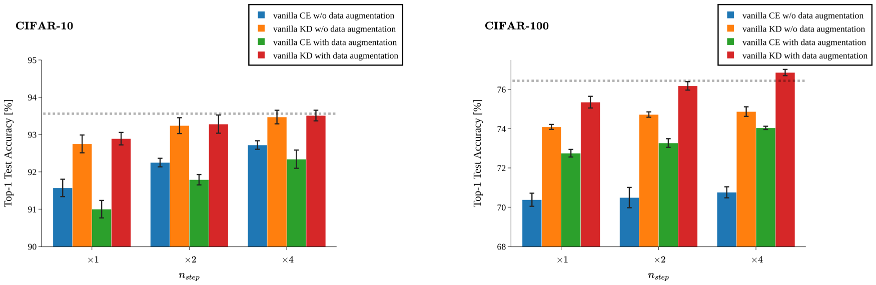

Fig. 2: This two figures show the comparative experimental results on CIFAR-10 and CIFAR-100. Among them, blue ●, orange ●, green ● and red ● represent four different loss combinations of $\{\mathcal{L}_{ce}\}$, $\{\mathcal{L}_{ce}, \mathcal{L}_{kd}\}$, $\{\mathcal{L}_{ce}, \mathcal{L}_{A}\}$ and $\{\mathcal{L}_{ce}, \mathcal{L}_{A}, \mathcal{L}_{kd}, \mathcal{L}_{T}\}$, respectively. And error bars in figures indicate standard deviation.

considers it is unnecessary for the student model to correctly identify these labels in knowledge distillation, and HSAKD inherits this behavior from SSKD. Although the knowledge distillation methods mentioned above do not attach importance to the hard labels provided by augmented samples, we will utilize them and manifest them can work (Sec. 3.3). We give the CE loss with respect to the augmented samples in the following Equation 9.

$$\mathcal{L}_A = \mathbb{E}_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} \, \mathbf{CE}\left(\mathbf{p}^S\left(\widetilde{\mathbf{x}}; 1\right), \widetilde{\mathbf{y}}\right), \tag{9}$$

where $\widetilde{\mathbf{y}}$ stands for the hard label about $\widetilde{\mathbf{x}}$. So in our proposed data-augmented knowledge distillation, the overall loss can be expressed as $\mathcal{L}_{oa}$ (in Equation 10, where $|\mathcal{X}|$ and $|\widetilde{\mathcal{X}}|$ refer to the number of elements in the original sample set and the augmented sample set, respectively). At the same time, this method can be denoted as KD+JDA.

$$\mathcal{L}_{oa} = \frac{|\mathcal{X}|}{|\mathcal{X}| + |\widetilde{\mathcal{X}}|}\mathcal{L}_{ce} + \frac{|\widetilde{\mathcal{X}}|}{|\mathcal{X}| + |\widetilde{\mathcal{X}}|}\mathcal{L}_{A} + \frac{|\mathcal{X}|}{|\mathcal{X}| + |\widetilde{\mathcal{X}}|}\mathcal{L}_{kd} + \frac{|\widetilde{\mathcal{X}}|}{|\mathcal{X}| + |\widetilde{\mathcal{X}}|}\mathcal{L}_{T}. \tag{10}$$

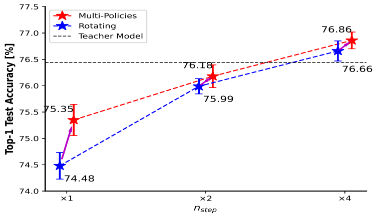

Fig. 3: In this figure, the purple arrows represent the changes in test accuracy from single *rotating* to JDA. So this scattergraph shows that our proposed JDA outperforms single *rotating*.

We should verify the validity of $\mathcal{L}_{oa}$ through rigorous experiments that n_{step} is the same for all comparative methods. Therefore, we set up three various n_{step} (i.e., ×1, ×2, ×4) in our experiments. Then we compare four different methods, and finally show the results in Fig. 2.

From Fig. 2, we can find that $\mathcal{L}_{oa}$ achieves the best performance on both CIFAR-10 and CIFAR-100. Especially on CIFAR-100, when n_{step} required for all methods is ×4, we can see that its test accuracy has exceeded SSKD and is only slightly lower than HSAKD by observing Tab. 2. So we can infer that **data augmentation is useful in knowledge distillation, and the**

Table 3: The data are obtained under the premise that the teacher model uses only the original samples in training. And ✓and × represent whether they use the corresponding loss or not, respectively.

CIFAR-10 • Acc. of teacher model:93.56%						CIFAR-100 • Acc. of teacher model:76.44%					
$\mathcal{L}_{ce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_A$	$\mathcal{L}_T$	Acc.(%)	Std.(%)	$\mathcal{L}_{ce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_A$	$\mathcal{L}_T$	Acc.(%)	Std.(%)
✓	×	✓	×	91.79	0.13	✓	×	✓	×	73.27	0.22
✓	✓	✓	✓	93.28	0.24	✓	✓	✓	✓	76.18	0.22
✓	✓	×	×	92.70	0.21	✓	✓	×	×	74.66	0.12
✓	×	×	✓	91.88	0.17	✓	×	×	✓	74.57	0.16
✓	×	×	×	92.14	0.08	✓	×	×	×	71.17	0.12
✓	✓	×	✓	93.15	0.11	✓	✓	×	✓	75.32	0.16

magnitude of its improvement depends on the nature of the dataset itself. Intuitively, the green rectangle is lower than blue rectangle on CIFAR-10, while the conclusion is converse on CIFAR-100. This illustrates that **adding data augmentation to the single vanilla CE loss during the training phase may harm the performance of the student model**.

Our proposed JDA composed of cascade sub-policies can indeed better transfer the "dark knowledge" from the teacher model to the student model. Unfortunately, both $\mathcal{L}_{SSKD} - \mathcal{L}_{ss}$ and $L_{HSAKD} - L_{kl\_q}$ lack $\mathcal{L}_A$ compared to $\mathcal{L}_{oa}$, so we cannot conclude that diverse and abundant data sub-policies are more effective than single *rotating* by directly comparing the relevant experimental results. In order to thoroughly verify the conclusion mentioned above under the premise of a fair comparison, we first add $\mathcal{L}_A$ for utilizing single *rotating* in training, and then conduct additional experiments on the CIFAR-100 benchmark and draw the results in Fig. 3. We can find that diverse and rich sub-policies have improved the test accuracy to some extent compared with single *rotating* at different n_{step}. In this way, we can conclude that **diverse and rich data augmentation is more valid than single *rotating***.

3.3 Decoupling the Overall Loss

Sec. 3.2 has demonstrated that $\mathcal{L}_{oa}$ in knowledge distillation is excellent. And this also effectively shows that data augmentation is quite helpful for the performance improvement of the student model. Therefore, another question that needs to be answered urgently is thrown: Which part of $\mathcal{L}_{oa}$ plays a positive role? Is it $\mathcal{L}_{ce}$, $\mathcal{L}_{kd}$, $\mathcal{L}_A$, or $\mathcal{L}_T$?

Following the above analysis, we argue that decoupling the overall loss is the necessary work. For the experiments in this subsection, the setting of n_{step} we uniformly adopted is $\times 2$, and the real input of the student model and the teacher model is the same as that of Sec. 3.2. Finally the decoupling experimental results can be found in Tab. 3. By analyzing the results in Tab. 3, we can find that some conclusions drawn on CIFAR-10 and CIFAR-100 are not consistent. In instance, simply adding $\mathcal{L}_T$ to $\mathcal{L}_{ce}$ hurts performance on the CIFAR-10 benchmark, but improves it on the CIFAR-100 benchmark. Of course, this similar conclusion also appears in the previous Section. So we only discuss the common conclusions on these two benchmarks. Intuitively, we conclude that $\mathcal{L}_T$ plays a weaker role than $\mathcal{L}_{kd}$ in knowledge distillation by comparing the results of two different loss combinations (i.e., $\{\mathcal{L}_{ce}, \mathcal{L}_{kd}\}$ and $\{\mathcal{L}_{ce}, \mathcal{L}_T\}$). On the other hand, when we add the combination of $\mathcal{L}_A$ and $\mathcal{L}_T$ to the combination of $\mathcal{L}_{ce}$ and $\mathcal{L}_{kd}$. The results are always some improvement on both CIFAR-10 and CIFAR-100. To sum up, **transferring knowledge only from the augmented samples does not necessarily work, but when the real input contains original samples in training, the augmented samples can effectively play an auxiliary advantage in improving the performance of the student model**. So we infer that **the augmented samples play an auxiliary advantage and the original samples play a key advantage in knowledge transfer**. In addition, by comparing combinations $\{\mathcal{L}_{ce}, \mathcal{L}_{kd}, \mathcal{L}_A, \mathcal{L}_T\}$ and $\{\mathcal{L}_{ce}, \mathcal{L}_{kd}, \mathcal{L}_T\}$ in Tab. 3, and even Fig. 3 and Tab. 1, we can find that the traditional CE loss of the augmented samples (i.e., $\mathcal{L}_A$) has an apparent positive effect. Therefore, our opinion is different from that in the SSKD paper [35]. We argue that the traditional CE loss of the augmented samples is also an essential part of knowledge distillation based on data augmentation. However, the teacher model does not use both original and augmented samples for the above experiments in training. As a result, it's natural to trust that transferring knowledge using only augmented samples is less effective than using only original samples, which might be because the teacher model lacks relevant knowledge. Considering this problem, we let the teacher model learn the information of the augmented samples and conduct the decoupling experiment again, and finally present the results in Tab. 4.

Table 4: The data are obtained under the premise that the teacher model uses both the original samples and the augmented samples in training. And ✓and × represent whether they use the corresponding loss or not, respectively. "(number)" in this table refers to the increased validation accuracy compared to that in Table 3.

CIFAR-10 • Acc. of teacher model:$94.16_{(+0.60)}$%						CIFAR-100 • Acc. of teacher model:$77.81_{(+1.37)}$%					
$\mathcal{L}_{ce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_A$	$\mathcal{L}_T$	Acc.(%)	Std.(%)	$\mathcal{L}_{ce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_A$	$\mathcal{L}_T$	Acc.(%)	Std.(%)
✓	×	✓	×	$91.79_{(+0.00)}$	$0.13_{(+0.00)}$	✓	×	✓	×	$73.27_{(+0.00)}$	$0.22_{(+0.00)}$
✓	✓	✓	✓	$92.94_{(-0.33)}$	$0.15_{(-0.10)}$	✓	✓	✓	✓	$75.81_{(-0.37)}$	$0.27_{(+0.06)}$
✓	✓	×	×	$93.25_{(+0.54)}$	$0.23_{(+0.01)}$	✓	✓	×	×	$75.69_{(+1.03)}$	$0.09_{(-0.04)}$
✓	×	×	✓	$91.77_{(-0.11)}$	$0.15_{(-0.02)}$	✓	×	×	✓	$74.23_{(-0.34)}$	$0.27_{(+0.10)}$
✓	×	×	×	$92.14_{(+0.00)}$	$0.08_{(+0.00)}$	✓	×	×	×	$71.17_{(+0.00)}$	$0.12_{(+0.00)}$
✓	✓	×	✓	$92.99_{(-0.16)}$	$0.15_{(+0.03)}$	✓	✓	×	✓	$75.59_{(+0.27)}$	$0.12_{(-0.04)}$

Comparing Tab. 3 and Tab. 4, it is surprising that when we utilize the teacher model to transfer knowledge with the augmented sample information, the performance of the student model has not been improved except that a loss combination $\{\mathcal{L}_{ce}, \mathcal{L}_{kd}\}$ is applied. It can be clearly inferred that **in the training stage of the teacher model, combing the original with augmented samples can effectively improve the performance of the teacher model compared with only using the original samples. But this is not available for the student model**. Meanwhile, when we observe Figure **??** in Appendix **??**, we can easily find that the augmented sample information is rather incorrect when the teacher model only uses the original samples in training. In contrast, the augmented sample information produced by the teacher model trained with both the original and augmented samples is relatively correct. So this means that **for augmented samples, "dark knowledge", which leads to misclassification, also plays a significant role in knowledge distillation**.

In particular, the teacher model transfers relatively correct information hurting the performance of the student model, and transfers relatively incorrect information improving the performance of the student model. This fact is contrary to our experience. By carefully observing Figure **??** (CIFAR-10) in Appendix **??**, we can find that the visualization of the original samples is changed after the teacher model has been trained with the additional augmented samples. So we argue that the reason causes the above conclusion is that **the teacher model trains both the original and augmented samples, which will damage the original samples' reasonable information**.

3.4 The Probability of Data Augmentation

This subsection will analyze the data augmentation based on the value of p. We know that p refers to the probability of each sub-policy being executed. Specifically, the larger p is, the greater the number of executed sub-policies in the augmented samples is. Generally, executing data augmentation with a high probability is not the best choice. We believe that this conclusion typically applies to JDA. To fully explore the effects of different p values, we set p to be 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0 for experiments, respectively. Finally, we show the results in Fig. 4 **(a) and (b)**.

As displayed in Fig. 4 **(a) and (b)**, we can clearly determine that the curve has a peak in each subfigure and illustrate the regularity of both curves. On a deeper level, this phenomenon implies that JDA's benefits are only available when p is inside a specified range. So we can draw a conclusion that p **is sensitive in training, and** p **with a reasonable setting can effectively achieve knowledge transfer**. Furthermore, although these experimental results guarantee that each mini-batch has at least half of the original samples, the student model performs poorly on both benchmarks when p is greater than 0.5. Therefore, we can infer that **the ratio of the augmented samples should not be set too large when using the augmented samples to assist knowledge distillation. Otherwise, the student model cannot obtain excellent performance**.

3.5 Few-Shot Analysis

In the real world, many datasets usually do not have a large amount of labeled data. Therefore, the study of few-shot learning becomes extremely important for solving this problem. For SSKD and HSAKD, they verified that their methods are robust by simulating a few-shot scenario that has only a small amount of labeled training samples. But since we have proved that data augmentation can strongly improve the student model's performance, it is reasonable for us to deduce that data augmentation may have contributed to the performance improvement of both SSKD and HSAKD in a few-shot scenario. In order that we can evaluate the role of data augmentation in a few-shot scenario, we follow the training setting in SSKD and HSAKD and randomly retain 25%, 50%, 75%, and 100% training samples in CIFAR-100. Particularly, we compare the performance of vanilla KD with and w/o JDA by training $\times 2$, and the experimental results can be found in Fig. 4 **(c) and (d)**.

Fig. 4 **(c) and (d)** illustrate that vanilla KD with data augmentation further strengthens the generalization ability of models when the labeled data are insufficient. Specifically, we can discover that when fewer training samples are retained, the student model

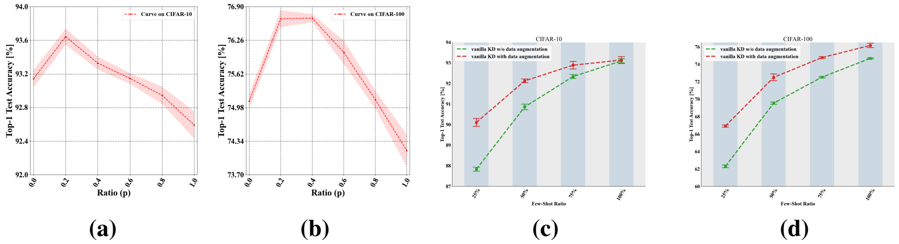

Fig. 4: **(a) and (b):** These two figures show the performance of the student model for different p values on CIFAR-10 and CIFAR-100, and the shaded area indicates the standard deviation. In addition, all experiments are performed with n_{step} set to $\times 2$. **(c) and (d):** These two figures demonstrate that our proposed joint data augmentation can work greatly in few-shot scenarios. Moreover, the error bars in figures refer to standard deviation.

with JDA can outperform the student model w/o JDA better. It means that **the potential of knowledge distillation with data augmentation in few-shot learning is enormous**.

3.6 Wide Comparison

In theory, JDA, as a data-focused method, can be perfectly combined with other methods that focus on "what to distill". To make sure data augmentation is robust in various teacher-student pairs and can be seamlessly compatible with some knowledge distillation methods, we conduct more extensive experiments on CIFAR-100. The hyperparameter settings of the experiments are the same as the CIFAR-100 benchmark mentioned in Sec. 3. And the results are shown in Tab. 5. By comparing whether to add JDA to vanilla KD, SPKD [32], and CRD [31][2], we can intuitively find that JDA effectively improves the performance of all methods. In addition, as can be seen in Appendix **??**, comparing the latest SOTA HSAKD and JDA, HSAKD requires additional computational modules, and more complex feature-based distillation, and spending two times n_{step} to get the same results. The above analysis fully demonstrates that **data augmentation is broadly effective and can be easily combined with other knowledge distillation methods**.

3.7 Cosine Confidence Knowledge Distillation

Inspired by Sec. 3.4, the strength of the augmented sample must be in a suitable range to exert a positive effect. If the model can adaptively assign appropriate weight to each augmented sample, the knowledge imparted by the teacher model can be more reasonable. As a result, here we propose a method called **C**osine **C**onfidence **D**istillation (CCD) to help transfer the knowledge of the augmented samples. First, as denoted in

[2] The reason JDA is not added to SSKD and HSAKD is that these methods themselves use *rotating* as their data augmentation. If we are to force the inclusion of JDA, it will destroy the original character of these approaches.

Table 5: Top-1 test accuracy (%) comparison of different methods across various teacher-student pairs on CIFAR-100. The results of SSKD and HSAKD are copied from [4], and the remaining results are obtained by our run. "($\pm number$)" in this table refers to the standard deviation, and the red number in the upper left corner of the accuracy symbolizes the ranking of closely-related methods. In particular, the rightmost column represents each method's n_{step} in our experiments.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet32×4 ResNet8×4	VGG13 MobileNetV2	n_{step}
Teacher	76.44	76.44	73.44	79.63	74.64	
Student	^{11}73.57$_{(\pm0.23)}$	^{11}71.95$_{(\pm0.59)}$	^{11}69.62$_{(\pm0.26)}$	^{11}72.95$_{(\pm0.24)}$	^{11}73.51$_{(\pm0.26)}$	×2
SSKD	^{7}76.16$_{(\pm0.17)}$	^{7}75.84$_{(\pm0.04)}$	^{10}70.80$_{(\pm0.02)}$	^{7}75.83$_{(\pm0.29)}$	^{7}76.21$_{(\pm0.16)}$	×4
HSAKD	^{2}77.20$_{(\pm0.17)}$	^{1}77.00$_{(\pm0.21)}$	^{4}72.58$_{(\pm0.33)}$	^{2}77.26$_{(\pm0.14)}$	^{5}77.45$_{(\pm0.21)}$	×4
KD	^{10}74.36$_{(\pm0.11)}$	^{10}73.21$_{(\pm0.10)}$	^{9}71.68$_{(\pm0.30)}$	^{10}72.34$_{(\pm0.12)}$	^{10}75.94$_{(\pm0.21)}$	×2
KD+JDA	^{5}76.80$_{(\pm0.13)}$	^{4}76.18$_{(\pm0.18)}$	^{6}72.37$_{(\pm0.28)}$	^{6}76.50$_{(\pm0.22)}$	^{3}77.64$_{(\pm0.23)}$	×2
SPKD	^{9}74.84$_{(\pm0.38)}$	^{9}73.51$_{(\pm0.17)}$	^{7}72.11$_{(\pm0.10)}$	^{9}72.77$_{(\pm0.25)}$	^{9}76.13$_{(\pm0.25)}$	×2
SPKD+JDA	^{6}76.58$_{(\pm0.31)}$	^{4}76.18$_{(\pm0.26)}$	^{3}72.73$_{(\pm0.11)}$	^{5}76.64$_{(\pm0.36)}$	^{6}77.33$_{(\pm0.14)}$	×2
CRD	^{8}74.88$_{(\pm0.16)}$	^{8}74.43$_{(\pm0.16)}$	^{8}71.94$_{(\pm0.20)}$	^{8}73.58$_{(\pm0.20)}$	^{8}76.14$_{(\pm0.17)}$	×2
CRD+JDA	^{4}76.84$_{(\pm0.23)}$	^{3}76.27$_{(\pm0.16)}$	^{5}72.38$_{(\pm0.08)}$	^{4}77.12$_{(\pm0.11)}$	^{4}77.61$_{(\pm0.06)}$	×2
CCD(ours)+JDA	^{3}77.16$_{(\pm0.14)}$	^{6}76.07$_{(\pm0.10)}$	^{2}72.82$_{(\pm0.16)}$	^{3}77.16$_{(\pm0.18)}$	^{2}77.71$_{(\pm0.12)}$	×2
CCD(ours)+JDA	^{1}77.34$_{(\pm0.12)}$	^{2}76.78$_{(\pm0.11)}$	^{1}73.24$_{(\pm0.08)}$	^{1}77.59$_{(\pm0.18)}$	^{1}78.11$_{(\pm0.10)}$	×4

Table 6: Top-1 accuracy (%) and Top-5 accuracy (%) comparison on ImageNet-1k. We follow the experimental setting in [31,35,4] and mark the highest Top-1 validation accuracy by **bold black**.

Teacher	Student	Acc.	Teacher	Student	KD	AT [37]	CC [25]	SPKD	RKD [24]	CRD	SSKD	HSAKD	DKD [42]	KD+JDA	CCD(ours)+JDA
ResNet-34	ResNet-18	Top-1	73.31	69.75	70.66	70.70	69.96	70.62	71.34	71.38	71.62	72.16	71.70	72.16	**72.22**
		Top-5	91.42	89.07	89.88	90.00	89.17	89.80	90.37	90.49	90.67	90.85	90.41	90.99	90.86
		n_{step}	-	×1	×1	×1	×1	×1	×1	×1	×4	×4	×1	×4	×4

Equation 11, we need to calculate the confidence of the teacher model with the augmented samples, which is measured by the cosine distance.

$$d = \textbf{cosine}(\widetilde{\mathbf{x}}, \mathbf{x}) = \frac{\left\langle f^T\left(\widetilde{\mathbf{x}}\right), f^T\left(\mathbf{x}\right)\right\rangle}{\|f^T\left(\widetilde{\mathbf{x}}\right)\|_2 \cdot \|f^T\left(\mathbf{x}\right)\|_2}. \tag{11}$$

Of particular note is that d provides a way of quantitatively presenting the strength of the augmented samples, the basis of which is clearly shown in Fig. 5. For the given augmented samples, the strength of their data augmentation is negatively correlated with their cosine confidence weight. Thus, the cosine distance is reasonable to measure whether the augmented samples have a high confidence level to facilitate distillation. Due to $d \in [-1, 1]$, utilizing it directly as weight makes the expectation of KL loss close to zero and model optimization difficult. We multiply $d+1$ as a weight $\in [0, 2]$ by $\textbf{KL}\left(\mathbf{p}^T\left(\widetilde{\mathbf{x}}; \tau\right) \| \mathbf{p}^S\left(\widetilde{\mathbf{x}}; \tau\right)\right)$. This means that the stronger an augmented sample is, the greater the distance between the original sample and the augmented sample is, and the smaller d is. Thus, less knowledge is transferred from the teacher model to the student model. So we denote new $\mathcal{L}_{\hat{T}}$ in Equation 12 instead of $\mathcal{L}_T$. And the new overall loss is shown in Equation 13.

$$\mathcal{L}_{\hat{T}} = \tau^2 \mathbb{E}_{(\mathbf{x},\widetilde{\mathbf{x}})\sim(\mathcal{X},\widetilde{\mathcal{X}})} \left(\textbf{cosine}\left(\widetilde{\mathbf{x}},\mathbf{x}\right)+1\right) * \textbf{KL}\left(\mathbf{p}^T\left(\widetilde{\mathbf{x}};\tau\right)\|\mathbf{p}^S\left(\widetilde{\mathbf{x}};\tau\right)\right). \tag{12}$$

$$\mathcal{L}_{\hat{o}a} = \frac{|\mathcal{X}|}{|\mathcal{X}|+|\widetilde{\mathcal{X}}|}\mathcal{L}_{ce} + \frac{|\widetilde{\mathcal{X}}|}{|\mathcal{X}|+|\widetilde{\mathcal{X}}|}\mathcal{L}_{A} + \frac{|\mathcal{X}|}{|\mathcal{X}|+|\widetilde{\mathcal{X}}|}\mathcal{L}_{kd} + \frac{|\widetilde{\mathcal{X}}|}{|\mathcal{X}|+|\widetilde{\mathcal{X}}|}\mathcal{L}_{\hat{T}}. \tag{13}$$

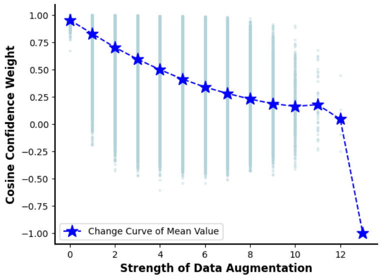

Fig. 5: The horizontal axis displays the number of transformations the augmented sample has undergone, and the vertical axis shows the cosine distance between the original sample and the associated augmented sample on CIFAR-100.

The apparent here to note in Equation 12 is that $\widetilde{\mathbf{x}}$ in sample pair $(\mathbf{x},\widetilde{\mathbf{x}})$ is transformed by $\mathbf{x}$. The comparative experimental results of CCD are also shown in Tab. 5. In addition, comparative results on the ImageNet-1k [27] benchmark can be found in Tab. 6. In particular, CCD is modified from the vanilla KD and exceeds the performance of KD+JDA. For n_{step} being $\times 2$, we can observe that CCD outperforms KD+JDA, SPKD+JDA, and CRD+JDA on all teacher-student pairs except WRN-40-2-WRN-40-1 in CIFAR-100, fully indicating that CCD is an excellent method. For n_{step} is $\times 4$, CCD surpasses HSAKD on all teacher-student pairs except WRN-40-2-WRN-40-1 in CIFAR-100. In the ImageNet-1k benchmark, we also achieve a SOTA distillation result. It should be emphasized that CCD achieves almost the same performance when the training time of HSAKD is twice that of CCD. Hence, CCD is more outstanding than HSAKD under all-around consideration.

4 Conclusion

In this paper, we conduct a multi-angle analysis of the role that data augmentation plays in knowledge distillation. Then we conclude that data augmentation can effectively improve the performance of knowledge distillation, and so forth (more detailed conclusions are shown at the end of Sec. 1). Furthermore, inspired by Sec. 3.4, we propose an excellent method named CCD to transfer knowledge of the augmented samples and the performance of CCD is better than that of the latest SOTA HSAKD. In future, our work will focus on "what kind of augmented samples should be used for distillation" or "how to better utilize the information of augmented samples", other than "what to distill".

Acknowledgement. This work was supported in part by the Aeronautical Science Foundation of China under Grant 20200058069001 and in part by the Fundamental Research Funds for the Central Universities under Grant 2242021R41094.

References

1. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10925–10934 (2022)
2. Chen, D., Mei, J.P., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3430–3437 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chuanguang Yang, Zhulin An, L.C.Y.X.: Hierarchical self-supervised augmented knowledge distillation. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). pp. 1217–1223 (2021)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
6. Das, D., Massa, H., Kulkarni, A., Rekatsinas, T.: An empirical analysis of the impact of data augmentation on distillation
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
8. Fu, J., Geng, X., Duan, Z., Zhuang, B., Yuan, X., Trischler, A., Lin, J., Pal, C., Dong, H.: Role-wise data augmentation for knowledge distillation. arXiv preprint arXiv:2004.08861 (2020)
9. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
10. Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., Luo, P.: Online knowledge distillation via collaborative learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11020–11029 (2020)
11. Guo, S.: Dpn: Detail-preserving network with high resolution representation for efficient segmentation of retinal vessels. Journal of Ambient Intelligence and Humanized Computing pp. 1–14 (2021)
12. Han, J., Fang, P., Li, W., Hong, J., Armin, M.A., Reid, I., Petersson, L., Li, H.: You only cut once: Boosting data augmentation with a single cut (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
14. 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)
15. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015). https://doi.org/10.48550/ARXIV.1503.02531, `https://arxiv.org/abs/1503.02531`
16. Ho, D., Liang, E., Chen, X., Stoica, I., Abbeel, P.: Population based augmentation: Efficient learning of augmentation policy schedules. In: International Conference on Machine Learning. pp. 2731–2741. PMLR (2019)
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

18. Lim, S., Kim, I., Kim, T., Kim, C., Kim, S.: Fast autoaugment. Advances in Neural Information Processing Systems **32** (2019)
19. Liu, S., Tian, Y., Chen, T., Shen, L.: Don't be so dense: Sparse-to-sparse gan training without sacrificing performance. International Journal of Computer Vision **20**(X) (2022)
20. Liu, S., Lin, T., He, D., Li, F., Deng, R., Li, X., Ding, E., Wang, H.: Paint transformer: Feed forward neural painting with stroke prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6598–6607 (October 2021)
21. Liu, Z., Farrell, J., Wandell, B.A.: Isetauto: Detecting vehicles with depth and radiance information. IEEE Access **9**, 41799–41808 (2021)
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
23. Matsubara, Y.: torchdistill: A modular, configuration-driven framework for knowledge distillation. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 24–44. Springer (2021)
24. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
25. 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/CVF International Conference on Computer Vision. pp. 5007–5016 (2019)
26. Razavi, M., Alikhani, H., Janfaza, V., Sadeghi, B., Alikhani, E.: An automatic system to monitor the physical distance and face mask wearing of construction workers in covid-19 pandemic. SN computer science **3**(1), 1–8 (2022)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
28. Sharma, S.: Game theory for adversarial attacks and defenses. arXiv preprint arXiv:2110.06166 (2021)
29. Singh, B., Najibi, M., Davis, L.S.: Sniper: Efficient multi-scale training. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), `https://proceedings.neurips.cc/paper/2018/file/166cee72e93a992007a89b39eb29628b-Paper.pdf`
30. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
31. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: International Conference on Learning Representations (2019)
32. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019)
33. Wang, H., Lohit, S., Jones, M., Fu, Y.: Knowledge distillation thrives on data augmentation. arXiv preprint arXiv:2012.02909 (2020)
34. Wieczorek, M., Rychalska, B., Dąbrowski, J.: On the unreasonable effectiveness of centroids in image retrieval. In: International Conference on Neural Information Processing. pp. 212–223. Springer (2021)
35. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 588–604. Springer International Publishing, Cham (2020)
36. 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)

37. 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(ICLR) (2016)
38. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
39. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
40. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
41. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
42. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11953–11962 (June 2022)