Switchable Online Knowledge Distillation

Biao Qian[1], Yang Wang[1(✉)], Hongzhi Yin[2], Richang Hong[1], and Meng Wang[1]

[1] Key Laboratory of Knowledge Engineering with Big Data,
Ministry of Education, School of Computer Science and Information Engineering,
Hefei University of Technology, Hefei, China
yangwang@hfut.edu.cn

[2] The University of Queensland, Brisbane, Australia
h.yin1@uq.edu.au

Abstract. Online Knowledge Distillation (OKD) improves the involved models by reciprocally exploiting the difference between teacher and student. Several crucial bottlenecks over the gap between them — e.g., Why and when does a large gap harm the performance, especially for student? How to quantify the gap between teacher and student? — have received limited formal study. In this paper, we propose **Swit**chable **O**nline **K**nowledge **D**istillation (SwitOKD), to answer these questions. Instead of focusing on the accuracy gap at test phase by the existing arts, the core idea of SwitOKD is to adaptively calibrate the gap at training phase, namely distillation gap, via a switching strategy between two modes — expert mode (pause the teacher while keep the student learning) and learning mode (restart the teacher). To possess an appropriate distillation gap, we further devise an adaptive switching threshold, which provides a formal criterion as to when to switch to learning mode or expert mode, and thus improves the student's performance. Meanwhile, the teacher benefits from our adaptive switching threshold and keeps basically on a par with other online arts. We further extend SwitOKD to multiple networks with two basis topologies. Finally, extensive experiments and analysis validate the merits of SwitOKD for classification over the state-of-the-arts. Our code is available at https://github.com/hfutqian/SwitOKD.

1 Introduction

The essential purpose of Knowledge Distillation (KD) [7,13,14,16,24–26,28,33] is to improve the performance of a *low-capacity student network* (small size, compact) for model compression by distilling the knowledge from a high-capacity teacher network (large size, over parameterized)[1]. The conventional knowledge distillation [2,3,7,9,10,15,27,32] requires a pre-trained teacher to serve as the *expert* network in advance, to be able to provide better supervision for the student in place of one-hot labels. However, it is usually a two-stage offline process, which is inflexible and requires extra computational cost.

Unlike offline fashion, the goal of recently popular online knowledge distillation is to reciprocally train teacher and student from scratch, where they

[1] Throughout the rest of the paper, we regard high-capacity network as teacher and low-capacity network as student for simplicity.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13671, pp. 449–466, 2022.
https://doi.org/10.1007/978-3-031-20083-0_27

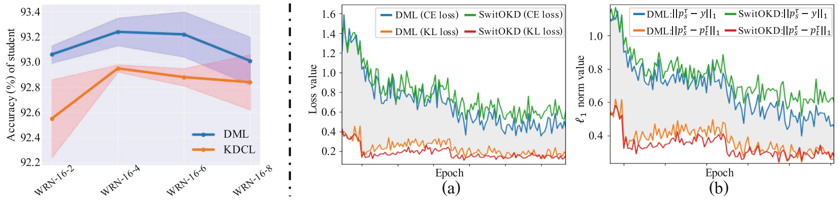


Fig. 1. **Left:** Illustration of how the large accuracy gap between teacher (WRN-16-2 to WRN-16-8) and student (ResNet-20) affects online distillation process on CIFAR-100 [11]. **Right:** DML [31] bears the emergency of escaping online KD under (a) large *accuracy gap* and (b) large *distillation gap*, whereas SwitOKD extends online KD's lifetime to avoid the degeneration.

learn extra knowledge from each other, and thus improve themselves simultaneously [1,4,22,31]. Typically, Deep Mutual Learning (DML) [31] encourages each network to mutually learn from each other by mimicking their predictions via Kullback Leibler (KL) divergence. Chen *et al.* [1] presents to improve the effectiveness of online distillation by assigning weights to each network with the same architecture. Further, Chung *et al.* [4] proposes to exchange the knowledge of feature map distribution among the networks via an adversarial means. Most of these approaches tend to *equally* train the same or different networks with *small accuracy gap*, where they usually lack richer knowledge from a powerful teacher. In other words, online fashion still fails to resolve the problem of student's performance impairment caused by a large accuracy gap [3,10,15] (see Fig. 1 Left), thus somehow violating the principle of KD. As inspired, we revisit such long-standing issue, and find the fundamental *bottlenecks* below: 1) when and how the gap has negative effect on online distillation process. For example, as the teacher turns from WRN-16-4 to WRN-16-8 (larger gap), the student accuracy rapidly declines (see Fig. 1 Left), while KL loss for the student degenerates into Cross-Entropy (CE) loss (see Fig. 1 Right(a)) as per loss functions in Table 1. To mitigate that, we raise 2) how to quantify the gap and automatically adapt to various accuracy gap, particularly large accuracy gap.

One attempt derives from Guo *et al.* [5], who studied the effect of large accuracy gap on distillation process and found that a large accuracy gap constitutes a certain harm to the performance of teacher. To this end, they propose KDCL, to admit the accuracy improvement of teacher by generating a high-quality soft target, so as to benefit the student. Unfortunately, KDCL pays more attention to teacher, which deviates from the essential purpose of KD; see Table 1.

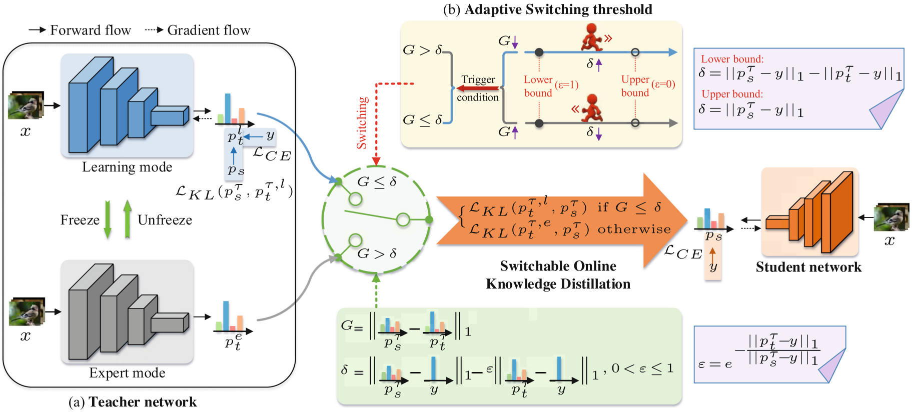
To sum up, the above online fashions overlook the principle of KD. Meanwhile, they focus on the accuracy gap that is merely obtained at *test* phase, which is not competent for quantifying the gap since it offers no guidance for the distribution alignment of distillation process performed at *training* phase. For instance, the accuracy just depends on the class with maximum probability given a 10-class output, while distillation usually takes into account all of the 10 classes. As opposed to them, we study the gap (the difference in class distribution

Table 1. The varied loss functions for typical distillation methods and the common form of their gradients. τ is set to 1 for theoretical and experimental analysis. *ensemble* is used to generate a soft target by combining the outputs of teacher and student. The gradient for KL divergence loss exactly reflects the difference between the predictions of student and teacher.

Method	Loss function of the networks	The common form of the gradient	Focus on student or not
KD [7] (NeurIPS 2015)	$\mathcal{L} = \alpha\mathcal{L}_{CE}(y, p_s^1) + (1-\alpha)\tau^2\mathcal{L}_{KL}(p_t^\tau, p_s^\tau)$	$(p_s^1 - y) + (p_s^\tau - p_t^\tau)$	✓
KDCL [5] (CVPR 2020)	$\mathcal{L} = \sum_i \mathcal{L}_{CE}^i(y, p_i^1) + \tau^2\mathcal{L}_{KL}^i(p_m, p_i^\tau)$, $p_m = ensemble(p_s^\tau, p_t^\tau)$	$(p_i^1 - y) + (p_i^\tau - p_m)$, $i = s, t$	✗
DML [31] (CVPR 2018)	$\mathcal{L}_s = \mathcal{L}_{CE}(y, p_s^1) + \mathcal{L}_{KL}(p_t^\tau, p_s^\tau)$, $\mathcal{L}_t = \mathcal{L}_{CE}(y, p_t^1) + \mathcal{L}_{KL}(p_s^\tau, p_t^\tau)$	$(p_s^1 - y) + (p_s^\tau - p_t^\tau)$, $(p_t^1 - y) + (p_t^\tau - p_s^\tau)$	✗
SwitOKD (Ours)	$\mathcal{L}_s = \mathcal{L}_{CE}(y, p_s^1) + \alpha\tau^2\mathcal{L}_{KL}(p_t^\tau, p_s^\tau)$, $p_t^\tau = p_t^{\tau,l} \Leftrightarrow p_t^\tau = p_t^{\tau,e}$	$(p_s^1 - y) + (p_s^\tau - p_t^{\tau,l})$ $\Updownarrow$ $(p_s^1 - y) + (p_s^\tau - p_t^{\tau,e})$	✓

between teacher and student) at *training* phase, namely *distillation gap*, which is quantified by ℓ_1 norm of the gradient (see Sect. 2.1), and how it affects online distillation process from student's perspective. Taking DML [31] as an example, we observe that the gradient for KL loss $||p_s^\tau - p_t^\tau||_1$ increasingly degenerates into that for CE loss $||p_s^\tau - y||_1$ given a large gap; see Fig. 1 Right(b). In such case, the student suffers from *the emergency of escaping online KD process.*

In this paper, we study online knowledge distillation and come up with a novel framework, namely **Swit**chable **O**nline **K**nowledge **D**istillation(SwitOKD), as illustrated in Fig. 2, which stands out new ways to mitigate the adversarial impact of large distillation gap on student. The basic idea of SwitOKD is to calibrate the distillation gap by adaptively pausing the teacher to wait for the learning of student during the *training* phase. Technically, we specify it via an adaptive switching strategy between two types of training modes: namely *learning mode* that is equivalent to reciprocal training from scratch and *expert mode* that freezes teacher's weights while keeps the student learning. Notably, we devise an adaptive switching threshold to endow SwitOKD with the capacity to yield an appropriate distillation gap that is conducive for knowledge transfer from teacher to student. Concurrently, it is nontrivial to devise an "ideal" switching threshold (see Sect. 2.5) due to: 1) not too large — a large threshold aggressively pushes *learning mode* and enlarges the distillation gap, resulting the student into the emergency of escaping online KD process; such fact, as expanded in Sect. 2.5, will further trap teacher to be paused constantly; as opposed to 2) not too small — the teacher constantly performs *expert mode* and receives poor accuracy improvement, suffering from no effective knowledge distilled from teacher to student. The above two conditions lead to 3) adaptiveness — the threshold is adaptively calibrated to balance learning mode and expert mode for extending online KD's lifetime. *Following* SwitOKD, we further establish two fundamental basis topologies to admit the extension of multi-network setting. The extensive experiments on typical datasets demonstrate the superiority of SwitOKD.



Forward flow
Gradient flow
x
Learning mode
p_t^l
y
$\mathcal{L}_{CE}$
p_s
$\mathcal{L}_{KL}(p_s^\tau, p_t^{\tau,l})$
Freeze
Unfreeze
x
Expert mode
p_t^e
(a) Teacher network
(b) Adaptive Switching threshold
$G > \delta$
$G \leq \delta$
Trigger condition
$G\downarrow$
$G\uparrow$
$\delta\uparrow$
$\delta\downarrow$
Lower bound (ε=1)
Upper bound (ε=0)
Lower bound:
$\delta = ||p_s^\tau - y||_1 - ||p_t^\tau - y||_1$
Upper bound:
$\delta = ||p_s^\tau - y||_1$
Switching
$G \leq \delta$
$G > \delta$
$\begin{cases} \mathcal{L}_{KL}(p_t^{\tau,l}, p_s^\tau) \text{ if } G \leq \delta \\ \mathcal{L}_{KL}(p_t^{\tau,e}, p_s^\tau) \text{ otherwise} \end{cases}$
Switchable Online Knowledge Distillation
p_s
$\mathcal{L}_{CE}$
y
Student network
x
$G = \| p_s^\tau - p_t^\tau \|_1$
$\delta = \| p_s^\tau - y \|_1 - \varepsilon \| p_t^\tau - y \|_1, 0 < \varepsilon \leq 1$
$\varepsilon = e^{-\frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1}}$

Fig. 2. Illustration of the proposed SwitOKD framework. Our basic idea is to adaptively pause the training of teacher while make the student continuously learn from teacher, to mitigate the adversarial impact of large distillation gap on student. Our framework is achieved by an adaptive switching strategy between two training modes: *learning mode* that is equivalent to training two networks reciprocally and *expert mode* that freezes teacher's parameters while keeps the student learning. Notably, we devise an adaptive switching threshold (b) to admit automatic switching between learning mode and expert mode for an appropriate distillation gap (quantified by G). See Sect. 2.6 for the detailed switching process.

2 Switchable Online Knowledge Distillation

Central to our method are three aspects: (i) quantifying the distillation gap between teacher and student, and analyzing it for online distillation (Sect. 2.1 and 2.2); (ii) an adaptive switching threshold to mitigate the adversarial impact of large distillation gap from student's perspective (Sect. 2.5); and, (iii) extending SwitOKD to multiple networks (Sect. 2.7).

2.1 How to Quantify the Distillation Gap Between Teacher and Student?

Thanks to varied random starts and differences in network structure (*e.g.*, layer, channel, etc.), the prediction difference between teacher and student always exists, which is actually exploited to benefit online distillation. Since the accuracy obtained at *test* phase is not competent to quantify the gap, we propose to quantify the gap at *training* phase, namely *distillation gap*, by computing ℓ_1 norm of the gradient for KL divergence loss (see Table 1), denoted as G, which is more suitable for capturing the same elements (0 entries) and the element-wise difference between the predictions of student and teacher, owing to the *sparsity* property of ℓ_1 norm. Concretely, given a sample x, let p_t^τ and p_s^τ represent the softened outputs of a teacher network $\mathcal{T}(x, \theta_t)$ and a student network $\mathcal{S}(x, \theta_s)$, respectively, then G is formulated as

$$G = ||p_s^\tau - p_t^\tau||_1 = \frac{1}{K}\sum_{k=1}^{K} |p_s^\tau(k) - p_t^\tau(k)|, G \in [0, 2], \tag{1}$$

where $|.|$ denotes the absolute value and τ is the temperature parameter. The k-th element of the softened output p_f^τ is denoted as $p_f^\tau(k) = \frac{exp(z_f(k)/\tau)}{\sum_j^K exp(z_f(j)/\tau)}, f = s, t$; $z_f(k)$ is the k-th value of the logit vector z_f. K is the number of classes. Prior work observes that a great prediction difference between teacher and student has a negative effect on distillation process [3,5,10]. Next, we discuss how the distillation gap affects online distillation process from student's perspective.

2.2 Why is an Appropriate Distillation Gap Crucial?

It is well-accepted that knowledge distillation loss for student is the KL divergence of the soften outputs of teacher p_t^τ and student p_s^τ [7], defined as

$$\mathcal{L}_{KL}(p_t^\tau, p_s^\tau) = \frac{1}{K}\sum_{k=1}^{K} p_t^\tau(k) log\frac{p_t^\tau(k)}{p_s^\tau(k)} = \mathcal{L}_{CE}(p_t^\tau, p_s^\tau) - H(p_t^\tau), \tag{2}$$

where $p_t^\tau(k)$ and $p_s^\tau(k)$ are the k-th element of the output vector p_t^τ and p_s^τ, respectively. $\mathcal{L}_{CE}(.,.)$ represents the Cross-Entropy loss and $H(\cdot)$ means the entropy value. Notably, when p_t^τ stays away from p_s^τ (large distillation gap appears), p_t^τ goes to y, then $\mathcal{L}_{KL}(p_t^\tau, p_s^\tau)$ will degenerate into $\mathcal{L}_{CE}(y, p_s^\tau)$ below:

$$\lim_{p_t^\tau \to y} \mathcal{L}_{KL}(p_t^\tau, p_s^\tau) = \lim_{p_t^\tau \to y} (\mathcal{L}_{CE}(p_t^\tau, p_s^\tau) - H(p_t^\tau)) = \mathcal{L}_{CE}(y, p_s^\tau), \tag{3}$$

where $H(y)$ is a constant (*i.e.*, 0) since y is the one-hot label. The gradient of $\mathcal{L}_{KL}$ w.r.t. z_s also has

$$\lim_{p_t^\tau \to y} \frac{\partial \mathcal{L}_{KL}}{\partial z_s} = \lim_{p_t^\tau \to y} \frac{1}{\tau}(p_s^\tau - p_t^\tau) = \frac{1}{\tau}(p_s^\tau - y), \tag{4}$$

where the gradient for KL loss increasingly degenerates into that for CE loss, resulting student into the emergency of escaping online KD process. The results in Fig. 1 Right also confirm the above analysis. As opposed to that, when p_t^τ goes to p_s^τ (the distillation gap becomes small), $\lim_{p_t^\tau \to p_s^\tau} \mathcal{L}_{KL}(p_t^\tau, p_s^\tau) = 0$, therefore no effective knowledge will be distilled from teacher to student.

How to Yield an Appropriate Gap? Inspired by the above, we need to yield an appropriate distillation gap G to ensure that student can always learn effective knowledge from the teacher throughout the training. In other words, the learning pace of student should continuously keep consistent with that of teacher. Otherwise, the online KD process will terminate. To this end, we propose to maintain an appropriate G. When G is larger than a threshold δ, namely *switching threshold*, we terminate teacher and keep only student learn from teacher, such training status is called *expert mode*. When expert mode progresses, G will decrease until less than δ, it will switch to the other training status of mutually learning between teacher and student, namely *learning mode*. The above two modes alternatively switch under an appropriate δ to keep improving the student's performance. Next, we will offer the details for learning mode (see Sect. 2.3) and expert mode (see Sect. 2.4), which pave the way to our proposed adaptive switching threshold δ (see Sect. 2.5).

2.3 Learning Mode: Independent *vs* Reciprocal

Unlike [3,7,10] that pre-train a teacher network in advance, the goal of learning mode is to reduce the distillation gap by training teacher and student network from scratch. Naturally, one naive strategy is to train the teacher independently with the supervision of one-hot label. Then the loss function of teacher and student is given as

$$\mathcal{L}_s^l = \mathcal{L}_{CE}(y, p_s^1) + \alpha\tau^2\mathcal{L}_{KL}(p_t^{\tau,l}, p_s^\tau), \quad \mathcal{L}_t^l = \mathcal{L}_{CE}(y, p_t^{1,l}), \tag{5}$$

where $p_t^{\tau,l}$ and p_s^τ are the predictions of teacher and student, respectively. α is a balancing hyperparameter. Unfortunately, the independently trained teacher provides poor improvement for student (see Sect. 3.3). Inspired by the fact that the teacher can benefit from reciprocal training [5,31] and, in turn, admit better guidance for student, we propose to reciprocally train student and teacher, therefore $\mathcal{L}_t^l$ in Eq. (5) can be upgraded to

$$\mathcal{L}_t^l = \mathcal{L}_{CE}(y, p_t^{1,l}) + \beta\tau^2\mathcal{L}_{KL}(p_s^\tau, p_t^{\tau,l}), \tag{6}$$

where β is a balancing hyperparameter. Thus we can compute the gradient of $\mathcal{L}_s^l$ and $\mathcal{L}_t^l$ w.r.t. z_s and z_t, *i.e.*,

$$\partial\mathcal{L}_s^l/\partial z_s = (p_s^1 - y) + \alpha\tau(p_s^\tau - p_t^{\tau,l}), \quad \partial\mathcal{L}_t^l/\partial z_t = (p_t^{1,l} - y) + \beta\tau(p_t^{\tau,l} - p_s^\tau). \tag{7}$$

In learning mode, the teacher usually converges faster (yield higher accuracy), owing to its superior learning ability, therefore *the distillation gap will increasingly grow as the training progresses. Meanwhile, for the student, KL loss exhibits a trend to be functionally equivalent to CE loss, causing the effect of knowledge distillation to be weakened.* In this case, SwitOKD will switch to expert mode.

2.4 Expert Mode: Turn to Wait for Student

To mitigate the adversarial impact of large distillation gap on student, SwitOKD attempts to pause the training of teacher while make student continuously learn from teacher, to keep the learning pace consistent, that sets it apart from previous online distillation methods [5,31]. Indeed, a teacher that is suitable for student rather than one who perfectly imitates one-hot label, can often improve student's performance, in line with our view of an appropriate distillation gap. Accordingly, the loss function for student is similar in spirit to that of Eq. (5):

$$\mathcal{L}_s^e = \mathcal{L}_{CE}(y, p_s^1) + \alpha\tau^2\mathcal{L}_{KL}(p_t^{\tau,e}, p_s^\tau), \tag{8}$$

where $p_t^{\tau,e}$ is the prediction of teacher network under expert mode. Thus the gradient of $\mathcal{L}_s^e$ w.r.t. z_s is computed as

$$\partial\mathcal{L}_s^e/\partial z_s = (p_s^1 - y) + \alpha\tau(p_s^\tau - p_t^{\tau,e}). \tag{9}$$

In such mode, the student will catch up or even surpass teacher as the training progresses, resulting into no effective knowledge distilled from teacher to student. Then SwitOKD will switch back to learning mode based on our adaptive switching threshold. We discuss that in the next section.

2.5 Adaptive Switching Threshold: Extending Online Knowledge Distillation's Lifetime

Intuitively, a naive strategy is to *manually* select a *fixed* value of δ, which, however, is inflexible and difficult to yield an appropriate distillation gap for improving the student (see Sect. 3.3). We propose an adaptive switching threshold for δ, which offer insights into how to *automatically* switch between learning mode and expert mode. First, observing that the distillation gap $G = ||p_s^\tau - p_t^\tau||_1 < ||p_s^\tau - y||_1$ on average because the teacher is superior to student, and

$$||p_s^\tau - p_t^\tau||_1 = ||(p_s^\tau - y) - (p_t^\tau - y)||_1 \geq ||p_s^\tau - y||_1 - ||p_t^\tau - y||_1, \tag{10}$$

which further yields $||p_s^\tau - y||_1 - ||p_t^\tau - y||_1 \leq G < ||p_s^\tau - y||_1$, leading to

$$\underbrace{||p_s^\tau - y||_1 - ||p_t^\tau - y||_1}_{\textbf{lower bound}} \leq \delta < \underbrace{||p_s^\tau - y||_1}_{\textbf{upper bound}}, \tag{11}$$

which, as aforementioned in Sect. 1, ought to be neither too large nor too small. To this end, we propose to adaptively adjust δ. Based on Eq. (11), we can further reformulate δ to be:

$$\delta = ||p_s^\tau - y||_1 - \varepsilon||p_t^\tau - y||_1, 0 < \varepsilon \leq 1. \tag{12}$$

It is apparent that ε approaching either 1 or 0 is equivalent to lower or upper bound of Eq. (11). Unpacking Eq. (12), the effect of ε is expected to be: when G becomes large, δ will be decreased towards $||p_s^\tau - y||_1 - ||p_t^\tau - y||_1$ provided ε approaching 1, then $G > \delta$ holds, which naturally enters into expert mode, and switches back into learning mode vice versa; see Fig. 2(b).

Discussion on $\boldsymbol{\varepsilon}$. As per Eq. (12), the value of δ closely relies on ε, which actually plays the role of tracking the changing trend of G. Intuitively, once the teacher learns faster than student, G will be larger, while $||p_t - y||_1 < ||p_s - y||_1$ holds from Eq. (12). Under such case, small value of δ is expected, leading to a larger value of ε, and vice versa. Hence, ε is inversely proportional to $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1}$. However, if G is very large, the student cannot catch up with the teacher; worse still, the teacher is constantly paused (trapped in expert mode) and cannot improve itself to distill knowledge to student, *making the online KD process terminated.* Hence, we decrease δ, so that, observing that p_t and p_s are very close during the early training time, the teacher can pause more times initially to make student to be in line with teacher, to avoid being largely fall behind at the later training stage (see Sect. 3.2 for detailed validations). Following this, we further decrease the value of r, such that $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1 + ||p_t^\tau - y||_1}$, to balance learning mode and expert mode. For normalization issue, we reformulate $\varepsilon = e^{-r}$, leading to the final adaptive switching threshold δ to be:

$$\boxed{\delta = ||p_s^\tau - y||_1 - e^{-\frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1 + ||p_t^\tau - y||_1}} ||p_t^\tau - y||_1.} \tag{13}$$

Algorithm 1. SwitOKD: Switchable Online Knowledge Distillation

```
Input: learning rate $\eta_1$, $\eta_2$, student network $\mathcal{S}$ parameterized by $\theta_s$, teacher network $\mathcal{T}$ parameterized by $\theta_t$
Output: Trained $\mathcal{S}$, $\mathcal{T}$
Randomly initialize $\mathcal{S}$ and $\mathcal{T}$.
for number of training iterations do
    Compute $G = ||p_s^\tau - p_t^\tau||_1$.
    Compute $\delta$ by Eqn. (13).
    if $G \leq \delta$ then
        # Learning Mode
        Estimate $\mathcal{L}_s^l$, $\mathcal{L}_t^l$ with Eqn. (6).
        Update $\theta_s$, $\theta_t$:
            $\theta_s \leftarrow \theta_s - \eta_1 \frac{\partial \mathcal{L}_s^l}{\partial z_s} \frac{\partial z_s}{\partial \theta_s}$
            $\theta_t \leftarrow \theta_t - \eta_2 \frac{\partial \mathcal{L}_t^l}{\partial z_t} \frac{\partial z_t}{\partial \theta_t}$
    else
        # Expert Mode
        Estimate $\mathcal{L}_s^e$ with Eqn. (8).
        Freeze $\theta_t$, update $\theta_s$:
            $\theta_s \leftarrow \theta_s - \eta_1 \frac{\partial \mathcal{L}_s^e}{\partial z_s} \frac{\partial z_s}{\partial \theta_s}$
    end if
end for
```

Unlike the existing arts [5,31], where they fail to focus on student to follow the principle of KD, SwitOKD can *extend online knowledge distillation's lifetime*, and thus largely improve student's performance, while keep our teacher be basically on par with theirs, thanks to Eq. (13); see Sect. 3.4 for validations.

2.6 Optimization

The above specifies the adaptive switching strategy between two training modes. Specifically, we kick off SwitOKD with learning mode to minimize $\mathcal{L}_s^l$ and $\mathcal{L}_t^l$, then the training mode is switched into expert mode to minimize $\mathcal{L}_s^e$ when $G > \delta$. Following that, SwitOKD switches back to learning mode when $G \leq \delta$. The whole training process is summarized in Algorithm 1.

2.7 Multi-network Learning Framework

To endow SwitOKD with the extendibility to multi-network setting with large distillation gap, we divide these networks into multiple teachers and students, involving switchable online distillation between teachers and students, which is built by two types of fundamental basis topologies below: multiple teachers *vs* one student and one teacher *vs* multiple students. For ease of understanding, we take 3 networks as an example and denote the basis topologies as **2T1S** and **1T2S**, respectively; see Fig. 3. Notably, the training between each teacher-student pair directly follows SwitOKD, while two teachers for **2T1S** (or two students for **1T2S**) mutually transfer knowledge in a conventional two-way manner. Note that, for **1T2S**, only when the switching conditions between teacher and both students are triggered, will the teacher be completely suspended. The detailed validation results are reported in Sect. 3.8.

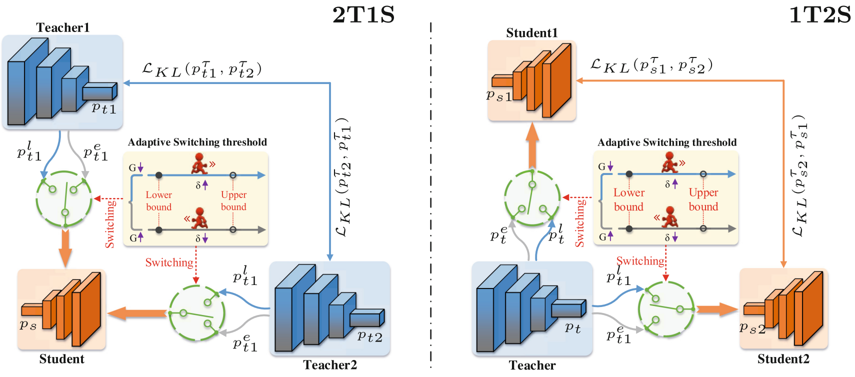


Fig. 3. The multi-network framework for training 3 networks simultaneously, including two fundamental basis topologies: **2T1S** (**Left**) and **1T2S** (**Right**).

3 Experiment

To validate the effectiveness of SwitOKD, we experimentally evaluate various state-of-the-art backbone networks via student-teacher pair below: MobileNet [8], MobileNetV2 [20] (sMobileNetV2 means the width multiplier is s), ResNet [6] and Wide ResNet (WRN) [30] over the typical image classification datasets: **CIFAR-10** and **CIFAR-100** [11] are natural image datasets, including 32 × 32 RGB images containing 10 and 100 classes. Both of them are split into a training set with 50 k images and a test set with 10k images. **Tiny-ImageNet** [12] consists of 64 × 64 color images from 200 classes. Each class has 500 training images, 50 validation images, and 50 test images. **ImageNet** [19] contains 1 k object classes with about 1.2 million images for training and 50 k images for validation.

3.1 Experimental Setup

We implement all networks and training procedures with pytorch [17] on an NVIDIA GeForce GTX 1080 Ti GPU and an Intel(R) Core(TM) i7-6950X CPU @ 3.00 GHz. For **CIFAR-10/100**, we use Adam optimizer with the momentum of 0.9, weight decay of 1e−4 and set batch size to 128. The initial learning rate is 0.01 and then divided by 10 at 140, 200 and 250 of the total 300 epochs. For **Tiny-ImageNet** and **ImageNet**, we adopt SGD as the optimizer, and set momentum to 0.9 and weight decay to 5e−4. Specifically, for Tiny-ImageNet, we set batch size to 128, the initial learning rate to 0.01, and the learning rate is dropped by 0.1 at 100, 140 and 180 of the total 200 epochs. For ImageNet, batch size is 512, while the initial learning rate is 0.1 (dropped by 0.1 every 30 epochs and trained for 120 epochs). As for the hyperparameter, we set α, β and τ to 1, and $\tau = \{2, 3\}$ for the classic distillation [5,7].

Previous sections (Sects. 2.3, 2.4, 2.5 and 2.7) explicate how adaptive switching strategy benefits a student. We offer practical insights into why SwitOKD

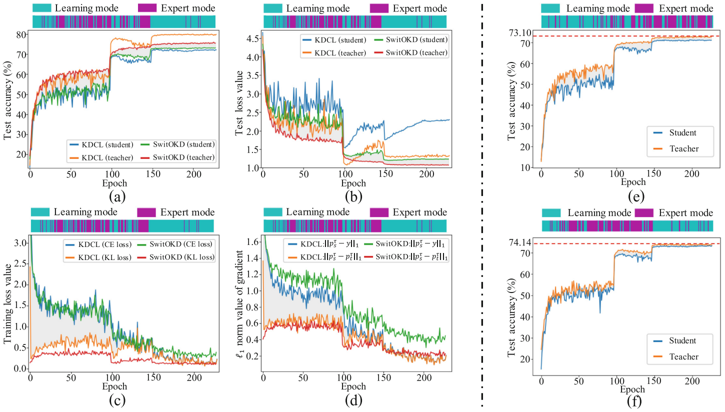


Fig. 4. Left: Illustration of test accuracy (a) and loss (b) for SwitOKD and KDCL [5]. From the perspective of student, (c) shows the comparison of CE loss and KL loss, while (d) is ℓ_1 norm value of the gradient for CE loss and KL loss. **Right:** Illustration of why the parameter r should be decreased from $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1}$ (e) to $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1 + ||p_t^\tau - y||_1}$ (f). The color bar shows the switching process of SwitOKD, where the cyan and the magenta denote learning mode and expert mode, respectively. (Color figure online)

works well, including ablation study and comparison with the state-of-the-arts, as well as extendibility to multiple networks.

3.2 Why Does SwitOKD Work Well?

One of our aims is to confirm that the core idea of our SwitOKD — using an adaptive switching threshold to achieve adaptive switching strategy — can possess an appropriate distillation gap for the improvement of student. The other is to verify why the parameter r (Sect. 2.5) should be decreased. We perform online distillation with a compact student network (ResNet-32) and a powerful teacher network (WRN-16-8 and WRN-16-2) on CIFAR-100.

Figure 4(a) (b) illustrate that the performance of student is continuously improved with smaller accuracy gap (gray area) compared to KDCL, confirming that our switching strategy can effectively calibrate the distillation gap to extend online KD's lifetime, in keeping with the core idea of SwitOKD. As an extension of Fig. 1, Fig. 4(c) reveals that KL loss for SwitOKD keeps far away from CE loss throughout the training unlike KDCL. Akin to that, the gradient for KL loss $||p_s^\tau - p_t^\tau||_1$ (refer to Eq. (1)) keeps divergent from that for CE loss $||p_s^\tau - y||_1$; see Fig. 4(d). Especially, *the color bar* illustrates the process of switching two modes on top of each other: when to pause the teacher — expert mode (magenta), or restart training — learning mode (cyan), reflecting that an appropriate gap holds with adaptive switching operation.

Figure 4(e)(f) validate the findings below: when $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1}$ (e), the teacher is rarely paused at the early stage of training, then the student largely falls behind at the later stage, leading to poor teacher (73.10% *vs* 74.14%) and student (71.89% *vs* 73.47%), confirming our analysis in Sect. 2.5 — $r = \frac{||p_t^\tau - y||_1}{||p_s^\tau - y||_1 + ||p_t^\tau - y||_1}$ (f) is desirable to balance learning mode and expert mode.

3.3 Ablation Studies

Table 2. Ablation study about the effectiveness of each component, of which constitutes SwitOKD. The best results are reported with **boldface**.

Case	Threshold δ	Switching or not	Teacher's loss $\mathcal{L}_t^l$	CIFAR-100
A	-	✗	Eqn.(6)	72.91
B	$\delta = 0.2$ $\delta = 0.6$ $\delta = 0.8$	✓	Eqn.(6)	72.92 72.83 73.00
C	Eqn.(13)	✓	Eqn.(5)	72.72
D	Eqn.(13)	✓	Eqn.(6)	**73.47**

Is Each Component of SwitOKD Essential? To verify the effectiveness of several components constituting SwitOKD — *switching strategy*, *adaptive switching threshold* and *teacher's training strategy*, we construct ablation experiments with ResNet-32 (student) and WRN-16-2 (teacher) from the following cases: **A**: SwitOKD without switching; **B**: SwitOKD with fixed δ (*i.e.*, $\delta \in \{0.2, 0.6, 0.8\}$); **C**: teacher's loss $\mathcal{L}_t^l$ (Eq. (6) *vs* Eq. (5)); **D**: the proposed SwitOKD. Table 2 summarizes our findings, which suggests that SwitOKD shows great superiority (73.47%) to other cases. Especially for case **B**, the manual δ fails to yield an appropriate distillation gap for improving the performance of student, confirming the importance of adaptive switching threshold, subject to our analysis (Sect. 2.5). Notably, the student for case **C** suffers from a large accuracy loss, verifying the benefits of reciprocal training on improving the performance of student (Sect. 2.3).

Table 3. Ablation study about the effectiveness of varied temperature τ on CIFAR-100. The best results are reported with **boldface**.

τ	0.5	1	2	5	8	10
SwitOKD	64.95	67.24	**67.80**	66.30	66.00	65.64

Why does the Temperature τ Benefit SwitOKD? The temperature τ [7] usually serves as a factor to smooth the predictions of student and teacher. Empirically, temperature parameter enables the output of student to be closer to that of teacher (and thus reduce the distillation gap), improving the performance, in line with our perspective in Sect. 2.4. To highlight the effectiveness of SwitOKD, we simply set $\tau = 1$ for our experiments. To further verify the effectiveness of varied $\tau \in \{0.5, 1, 2, 5, 8, 10\}$, we perform the ablation experiments with MobileNetV2 (student) and WRN-16-2 (teacher). Table 3 summarizes the findings. The optimal student (67.80%) is achieved with a slightly higher $\tau^* = 2$, implying that τ contributes to the calibration of the distillation gap. Note that when $\tau = 10$, the accuracy of student rapidly declines, implying that excessive smoothness can make the gap beyond an optimal range and, in turn, harm the performance of student, consistent with our view of an appropriate gap in Sect. 2.2.

Table 4. Accuracy (%) comparison on Tiny-ImageNet and CIFAR-10/100. (.M) denotes the number of parameters. All the values are measured by computing mean and standard deviation across 3 trials with random seeds. The best results are reported with **boldface**.

Backbone		Vanilla	DML [31]	KDCL [5]	SwitOKD
Tiny-ImageNet					
Student	1.4MobileNetV2(4.7 M)	50.98 ± 0.32	55.70 ± 0.61	57.79 ± 0.30	**58.71 ± 0.11**
Teacher	ResNet-34(21.3 M)	63.18 ± 0.37	64.49 ± 0.43	**65.47 ± 0.32**	63.31 ± 0.04
Student	ResNet-20(0.28M)	52.35 ± 0.15	53.98 ± 0.26	53.74 ± 0.39	**55.03 ± 0.19**
Teacher	WRN-16-2(0.72 M)	56.59 ± 0.22	57.45 ± 0.19	**57.71 ± 0.30**	57.41 ± 0.06
CIFAR-10					
Student	WRN-16-1(0.18 M)	91.45 ± 0.06	91.96 ± 0.08	91.86 ± 0.11	**92.50 ± 0.17**
Teacher	WRN-16-8(11.0 M)	95.21 ± 0.12	95.06 ± 0.05	**95.33 ± 0.17**	94.76 ± 0.12
CIFAR-100					
Student	0.5MobileNetV2(0.81 M)	60.07 ± 0.40	66.23 ± 0.36	66.83 ± 0.05	**67.24 ± 0.04**
Teacher	WRN-16-2(0.70 M)	72.90 ± 0.09	73.85 ± 0.21	73.75 ± 0.26	**73.90 ± 0.40**

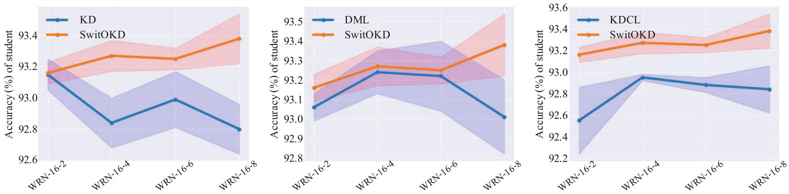


Fig. 5. Illustration of how an appropriate distillation gap yields better student. For KD, DML and KDCL, the accuracy of student (ResNet-20) rapidly declines as the teacher turns to higher capacity (WRN-16-2 to WRN-16-8). As opposed to that, SwitOKD grows steadily, owing to an appropriate distillation gap.

3.4 Comparison with Other Approaches

To verify the superiority of SwitOKD, we first compare with typical online KD methods, including: 1) DML [31] is equivalent to learning mode for SwitOKD; 2) KDCL [5] studies the effect of large accuracy gap at the test phase on online distillation process, but they pay more attention to teacher instead of student. For brevity, "vanilla" refers to the backbone network trained from scratch with classification loss alone. A compact student and a powerful teacher constitute the student-teacher network pair with *large distillation gap* at the training phase.

Table 4 and Fig. 5 summarize our findings below: *First*, switchable online distillation offers a significant and consistent performance improvement over the baseline (vanilla) and the state-of-the-arts for *student*, in line with the principle of KD process. Impressively, SwitOKD achieves 1.05% accuracy improvement to vanilla on CIFAR-10 (7.17% on CIFAR-100). Besides, SwitOKD also shows 0.54% and 0.54% (WRN-16-1/WRN-16-8) accuracy gain over DML and KDCL on CIFAR-100, respectively. Especially with 1.4MobileNetV2/ResNet-34, SwitOKD still obtains significant performance gains of 7.73%, 3.01% and 0.92% (the gains are substantial for Tiny-ImageNet) over vanilla, DML and KDCL.

Table 5. Accuracy (%) comparison of student network with offline KD methods (seen as expert mode of SwitOKD) on CIFAR-100. (.M) denotes the number of parameters. The best results are reported with **boldface**.

Backbone		Vanilla	KD [7]	FitNet [18]	AT [29]	CRD [23]	RCO [10]	SwitOKD
Student	WRN-16-2(0.70 M)	72.79	74.49	73.44	73.35	75.01	75.36	**75.95**
Teacher	WRN-40-2(2.26 M)	76.17	–	–	–	–	–	**76.54**

Table 6. Top-1 accuracy (%) on ImageNet dataset. (.M) denotes the number of parameters. The best results are reported with **boldface**.

Backbone		Vanilla	DML [31]	KDCL [5]	SwitOKD
Student	ResNet-18(11.7 M)	69.76	70.81	70.91	**71.75**
Teacher	ResNet-34(21.8 M)	73.27	73.47	**73.70**	73.65
Student	0.5MobileNetV2(1.97 M)	63.54	64.22	63.92	**65.11**
Teacher	ResNet-18(11.7 M)	69.76	68.30	**70.60**	68.08

Second, our teachers still benefit from SwitOKD and obtain accuracy improvement *basically on a par* with DML and KDCL, confirming our analysis about the adaptive switching threshold δ (see Eq. (13)) — balance of learning mode and expert mode. Note that, with 0.5MobileNetV2 and WRN-16-2 on CIFAR-100, our teacher (73.90%) upgrades beyond the vanilla (72.90%), even yields comparable accuracy gain (0.05% and 0.15%) over DML and KDCL. By contrast, KDCL has most of the best teachers, but with poor students, owing to its concentration on teacher only.

Finally, we also validate the effectiveness of SwitOKD for student even under a small distillation gap on CIFAR-10 (see Fig. 5), where the students (ResNet-20) still possess significant performance advantages, confirming the necessity of adaptively calibrating an appropriate gap with adaptive switching threshold δ in Sect. 2.5. Especially for Fig. 5 (b)(c), as the teacher turns to higher capacity (WRN-16-2 to WRN-16-8), students' accuracy from DML and KDCL rises at the beginning, then rapidly declines, and reaches the best results when the teacher is WRN-16-4. This, in turn, keeps consistent with our analysis (Sect. 2.2) — an appropriate distillation gap admits student's improvement.

To further validate the switching strategy between two modes, we also compare SwitOKD with offline knowledge distillation approaches (seen as expert mode of SwitOKD) including KD [7], FitNet [18], AT [29] and CRD [23] that require a fixed and pre-trained teacher. Especially, RCO [10] is similar to our approach, which maintains a reasonable performance gap by manually selecting a series of pre-trained intermediate teachers. Table 5 reveals that SwitOKD achieves superior performance over offline fashions, while exceeds the second best results from RCO by 0.59%, implying that SwitOKD strictly follows the essential principle of KD with the adaptive switching strategy.

Table 7. Accuracy (%) comparison with 3 networks on CIFAR-100. WRN-16-2 serves as either teacher (**T**) or student (**S**) for DML and KDCL, while is treated as S for **1T2S** and T for **2T1S**. The best results are reported with **boldface**.

Backbone	Vanilla	DML [31]	KDCL [5]	**SwitOKD (1T2S)**	**SwitOKD (2T1S)**
MobileNet	58.65(**S**)	63.75(**S**)	62.13(**S**)	**64.62**(**S**)	**65.03**(**S**)
WRN-16-2	73.37(**S**/**T**)	74.30(**S**/**T**)	73.94(**S**/**T**)	**75.02**(**S**)	71.73(**T**)
WRN-16-10	79.45(**T**)	77.82(**T**)	**80.71**(**T**)	77.33(**T**)	77.07(**T**)

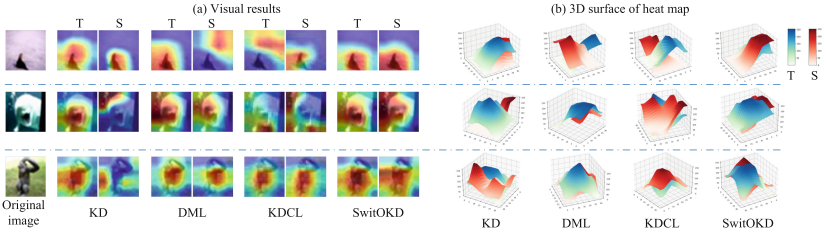


Fig. 6. Visual analysis of why SwitOKD works on CIFAR-100. **Left:**(a) The visual results by superimposing the heat map onto corresponding original image. **Right:**(b) 3D surface of heat maps for teacher and student (the more the peak overlaps, the better the student mimics teacher), where x and y axis denote the width and height of an image, while z axis represents the gray value of the heat map. T: ResNet-34 (teacher); S: ResNet-18 (student).

3.5 Extension to Large-Scale Dataset

Akin to [5,31], as a by-product, SwitOKD can effectively be extended to the large-scale dataset (*i.e.*, ImageNet), benefiting from its good generalization ability; see Table 6. It is observed that the students' accuracy is improved by 1.99% and 1.57% upon the vanilla, which are substantial for ImageNet, validating the scalability of SwitOKD. Particularly, for ResNet-34, our teacher (73.65%) outperforms the vanilla (73.27%) and DML (73.47%), highlighting the importance of our adaptive switching strategy upon δ to balance the teacher and student. Another evidence is shown for 0.5MobileNetV2 and ResNet-18 with larger distillation gap, our student outperforms DML and KDCL by 0.89% and 1.19%, while the teacher also yields comparable performance with DML, keeping consistent with our analysis in Sect. 2.5.

3.6 How About SwitOKD from Visualization Perspective?

To shed more light on why SwitOKD works in Sect. 3.2, we further perform a visual analysis with Grad-cam [21] visualization of image classification via a heat map (red/blue region corresponds to high/low value) that localizes the class-discriminative regions, to confirm that our adaptive switching strategy enables student to mimic teacher well, and thus improves the classification accuracy.

Figure 6(a) illustrates the visual results by superimposing the heat map onto corresponding original image, to indicate whether the object regions of the image is focused (the red area denotes more focus); Fig. 6(b) shows 3D surface of the heat map to reflect the overlap range of heat maps for teacher and student (the more the peak overlaps, the better the student mimics teacher). Combining Fig. 6(a) and (b), it suggests that SwitOKD focuses on the regions of student, which not only keep consistent with that of the teacher — mimic the teacher well (KL loss), but correctly cover the object regions — yield high precision (CE loss), in line with our analysis (Sect. 1): keep the gradient for KL loss divergent from that for CE loss. *The above further confirms the adversarial impact of large distillation gap — the emergency of escaping online KD process* (Sects. 2.2 and 2.3).

3.7 What Improves the Training Efficiency of SwitOKD?

Interestingly, SwitOKD has considerably raised the training efficiency of online distillation process beyond [5,31] since the training of teacher is *paused* (merely involve inference process) under expert mode (Sect. 2.4). We perform efficiency analysis on a single GPU (GTX 1080 Ti), where SwitOKD is compared with other online distillation methods, *e.g.*, DML [31] and KDCL [5]. Figure 7 shows that the time per iteration for SwitOKD (green line) varies greatly, owing to adaptive switching operation. Notably, the total training time is significantly reduced by 27.3% (9.77h *vs* 13.43h) compared to DML (blue line), while 34.8% (9.77h *vs* 14.99h) compared to KDCL (orange line).

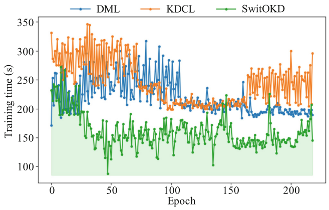


Fig. 7. Efficiency analysis with Mobile NetV2 (student) and ResNet-18 (teacher) on Tiny-ImageNet. (Color figure online)

3.8 Extension to Multiple Networks

To show our approach's extendibility for training multiple networks, we conduct the experiments based on three networks with large distillation gap, see Table 7. As can be seen, the students for **1T2S** and **2T1S** achieve significant accuracy gains (5.97%, 1.65%, and 6.38%) over vanilla and outperform other online distillation approaches (*i.e.*, DML [31] and KDCL [5]) with significant margins, while our teachers (WRN-16-10) are basically on a par with DML, consistent with the tendency of performance gain for SwitOKD in Table 4. By contrast, KDCL receives the best teacher (WRN-16-10), but a poor student (MobileNet), in that it pays more attention to teacher instead of student. Notably, **1T2S** achieves a better teacher (77.33% *vs* 77.07% for WRN-16-10) than **2T1S**; the reason is that the teacher for **1T2S** will be completely suspended when the switching conditions between teacher and both students are triggered (Sect. 2.7).

4 Conclusion

In this paper, we propose Switchable Online Knowledge Distillation (SwitOKD), to mitigate the adversarial impact of large distillation gap between teacher and student, where our basic idea is to calibrate the distillation gap by adaptively pausing the teacher to wait for the learning of student. We foster it throughout an adaptive switching strategy between learning mode and expert mode. Notably, an adaptive switching threshold is devised to endow SwitOKD with the capacity to automatically yield an appropriate distillation gap, so that the performance of student and teacher can be improved. Further, we verify SwitOKD's extendibility to multiple networks. The extensive experiments on typical classification datasets validate the effectiveness of SwitOKD.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under grant no U21A20470, 62172136, 61725203, U1936217. Key Research and Technology Development Projects of Anhui Province (no. 202004a5020043).

References

1. 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)
2. 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)
3. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
4. Chung, I., Park, S., Kim, J., Kwak, N.: Feature-map-level online adversarial knowledge distillation. In: International Conference on Machine Learning, pp. 2006–2015. PMLR (2020)
5. Guo, Q., et al.: Online knowledge distillation via collaborative learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. 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)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS (2015)
8. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
9. Huang, Z., et al.: 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)
10. Jin, X., et al.: Knowledge distillation via route constrained optimization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1345–1354 (2019)

11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
12. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N, **7N**(7), 3 (2015)
13. Li, T., Li, J., Liu, Z., Zhang, C.: Few sample knowledge distillation for efficient network compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14639–14647 (2020)
14. Menon, A.K., Rawat, A.S., Reddi, S., Kim, S., Kumar, S.: A statistical perspective on distillation. In: International Conference on Machine Learning, pp. 7632–7642. PMLR (2021)
15. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 5191–5198 (2020)
16. Passalis, N., Tzelepi, M., Tefas, A.: Heterogeneous knowledge distillation using information flow modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2339–2348 (2020)
17. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
18. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
19. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv 2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
22. Song, G., Chai, W.: Collaborative learning for deep neural networks. In: Advances in Neural Information Processing Systems, pp. 1832–1841 (2018)
23. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
24. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1365–1374 (2019)
25. 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)
26. Wang, Y.: Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) **17**(1s), 1–25 (2021)
27. 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.) ECCV 2020. LNCS, vol. 12354, pp. 588–604. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_34
28. 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)
29. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)

30. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
31. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
32. Zhu, J., et al.: Complementary relation contrastive distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9260–9269 (2021)
33. Zhu, Y., Wang, Y.: Student customized knowledge distillation: bridging the gap between student and teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5057–5066 (2021)