International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01792-z

Multi-target Knowledge Distillation via Student Self-reflection

Jianping Gou[1] · Xiangshuo Xiong[2] · Baosheng Yu[3] · Lan Du[4] · Yibing Zhan[5] · Dacheng Tao[3]

Received: 11 July 2022 / Accepted: 27 March 2023

Abstract
Knowledge distillation is a simple yet effective technique for deep model compression, which aims to transfer the knowledge learned by a large teacher model to a small student model. To mimic how the teacher teaches the student, existing knowledge distillation methods mainly adapt an unidirectional knowledge transfer, where the knowledge extracted from different intermedicate layers of the teacher model is used to guide the student model. However, it turns out that the students can learn more effectively through multi-stage learning with a self-reflection in the real-world education scenario, which is nevertheless ignored by current knowledge distillation methods. Inspired by this, we devise a new knowledge distillation framework entitled multi-target knowledge distillation via student self-reflection or MTKD-SSR, which can not only enhance the teacher's ability in unfolding the knowledge to be distilled, but also improve the student's capacity of digesting the knowledge. Specifically, the proposed framework consists of three target knowledge distillation mechanisms: a stage-wise channel distillation (SCD), a stage-wise response distillation (SRD), and a cross-stage review distillation (CRD), where SCD and SRD transfer feature-based knowledge (i.e., channel features) and response-based knowledge (i.e., logits) at different stages, respectively; and CRD encourages the student model to conduct self-reflective learning after each stage by a self-distillation of the response-based knowledge. Experimental results on five popular visual recognition datasets, CIFAR-100, Market-1501, CUB200-2011, ImageNet, and Pascal VOC, demonstrate that the proposed framework significantly outperforms recent state-of-the-art knowledge distillation methods.

Keywords Knowledge distillation · Self-reflection learning · Model compression · Deep learning

1 Introduction

In the past few years, deep neural networks have achieved state-of-the-art performances on natural language processing (NLP) and computer vision (CV) tasks, including language modelling (Li et al., 2021; Yuan et al., 2021; You et al., 2021; Hagström & Johansson, 2021), object detection/recognition (Huang et al., 2022; Shu et al., 2021; Yang et al., 2022; Gou et al., 2021; He et al., 2022), semantic segmentation (Shu et al., 2021; Liu et al., 2020; Wang et al., 2020), pose estimation (Wang et al., 2020; Yu & Tao, 2021), video retrieval (Kordopatis-Zilos et al., 2022), as well as visual-linguistic tasks (Fang et al., 2021; Peng et al., 2021; Ma et al., 2022). Nevertheless, a trade-off between a computationally demanding model with millions of parameters, if not billions, and the SOTA performance impedes their deploy-

Communicated by Christoph H. Lampert.

✉ Baosheng Yu
baosheng.yu@sydney.edu.au

Jianping Gou
cherish.gjp@gmail.com

Xiangshuo Xiong
2221908048@stmail.ujs.edu.cn

Lan Du
lan.du@monash.edu

Yibing Zhan
zhanyibing@jd.com

Dacheng Tao
dacheng.tao@sydney.edu.au

[1] College of Computer and Information Science, College of Software, Southwest University, Chongqing 400715, China

[2] School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

[3] School of Computer Science, The University of Sydney, Sydney, NSW 2008, Australia

[4] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

[5] JD Explore Academy, JD Technology, Beijing 100176, China

Published online: 25 April 2023

ment in computing resource-limited environments, such as mobile devices. As one of the most simple and effective model compression techniques, knowledge distillation overcomes this impediment by transferring the knowledge of a large-scale teacher model to improve the generalizability of a light-weight student model with fewer parameters (Hinton et al., 2015).

Recently, there has been many knowledge distillation mechanisms (Gou et al., 2021) that vary in either the distillation mechanisms (e.g., offline distillation, online distillation, and self-distillation) or the types of knowledge distilled (e.g., logits, feature knowledge, and relational knowledge). Specifically, offline distillation utilises a two-stage approach that requires the teacher model to be pre-trained as a prior and fixed for knowledge transfer (Liu et al., 2021; Zhao et al., 2022; Mirzadeh et al., 2020; Wu et al., 2020; Chen et al., 2022), while online and self-distillation instead argues that a pre-trained large teacher model is not always available for the tasks of interest, thus propose to update both the teacher model and the student model simultaneously in an end-to-end manner (Yuan et al., 2020; Mobahi et al., 2020; Zhang et al., 2022; Zhao et al., 2021; Xu et al., 2022; Shen et al., 2022; Gou et al., 2022; Li et al., 2018; Wu & Gong, 2021; Zhang et al., 2018).

Learning actually requires the two-way communication: it depends on not only the teacher's ability of unfolding the knowledge but also the capacity of the student in digesting the knowledge. To the best of our knowledge, existing KD methods mainly focus on how effective a teacher can be in knowledge distillation, ignoring the knowledge transfer from the student perspective, i.e., a student can learn different knowledge from both the teacher and self-reflection. Therefore, it is of great importance to pay attention to how a student can learn effectively from both the teacher and itself, where self-reflection allows a student to analyse, evaluate, and improve its own learning process by reviewing the learned knowledge after each stage of learning. As a simple yet effective way, self-distillation has shown to be very effective for self-reflection, for example, to accelerate the inference of large language model (Liu et al., 2020). Meanwhile, under the teacher's guidance, the overall learning process usually can be divided into multiple stages with different levels of intellectual challenges, such as adaptive distillation with multiple paths and gradient similarity (Chennupati et al., 2021; Zhu & Wang, 2021) and block-wise architectures (Li et al., 2020).

To address the above-mentioned issues in knowledge distillation, we propose a novel framework, named multi-target knowledge distillation via student self-reflection or MTKD-SSR, which inherits the advantages of multi-stage learning and self-reflection. Specifically, the proposed method consists of three major distillation modules: (1) a stage-wise channel distillation (SCD) module assures the feature maps generated by the teacher and the student at each layer (e.g., a residual block and a transformer) to be as close as possible to each other; (2) a stage-wise response distillation (SRD) module trained with multi-task loss augments each layer at different depths with an early existing classifier for both the teacher model and the student model (Phuong & Lampert, 2019); and (3) a cross-stage review distillation (CRD) module mimics how a real-world student carries out self-reflection on its own learning, which transfers the response-based knowledge (i.e., logits) from shallower layers to deeper layers (i.e., a reverse direction of knowledge distillation used in self-distillation). Actually, each layer has its own dark knowledge to be distilled for the next layers. Compared to the deep layer of the student network, the dark knowledge from its shallow layer can be regarded as the informative knowledge previously learned from the perspective of human learning, and thus can be transferred to facilitate the learning of the following layers via self-distillation. Furthermore, a student can also be greatly improved by even a weak "teacher" with noise and/or poor performance (Yuan et al., 2020), which corresponds to the early exiting classifiers associated with the preceding layers in our case.

With the above-mentioned three important modules trained jointly, it assures that the student model iteratively improves itself and learns from the teacher model at different levels of intellectual challenges. To the best of our knowledge, this is the first multi-target knowledge distillation mechanism that explores the ideas of both multi-stage learning and self-reflection in a unified framework. To evaluate the proposed framework, we perform extensive experiments with comprehensive ablation studies on five popular visual recognition datasets, where the results show that the proposed MTKD-SSR framework significantly outperforms recent KD methods, including self-distillation (SD) (Zhang et al., 2019) and distillation via knowledge review (DKR) (Chen et al., 2021). In summary, our main contributions are as follows:

- We devise a new knowledge distillation method called multi-target knowledge distillation via student self-reflection (MTKD-SSR): It inherits the advantage of both offline and self distillation with different types of knowledge in the unified distillation framework.
- We introduce a novel cross-stage review for self-distillation, which distills the knowledge from the preceding layers to the subsequent layers in the student network.
- We evaluate the proposed method for knowledge distillation by performing extensive experiments on five visual recognition tasks, where the proposed MTKD-SSR framework outperforms recent state-of-the art KD methods with a clear margin.

The remainder of this paper is structured as follows. Section 2 outlines the related work. Section 3 introduces

our proposed MTKD-SSR method. Section 4 describes the experimental results to show the effectiveness of the proposed MTKD-SSR. Section 5 provides comprehensive ablation studies about different distillation and knowledge. Lastly, we conclude the proposed method in Sect. 6.

2 Related Work

In this section, we briefly review knowledge distillation methods that are closely related to ours in terms of different distillation schemes and different types of knowledge. For a comprehensive review of knowledge distillation, we refer the interested readers to (Gou et al., 2021; Wang & Yoon, 2022).

2.1 Distillation Scheme

Since the first introduction of offline knowledge distillation (Hinton et al., 2015), which relies on a unidirectional two-stage training procedure, there have existed a variety of knowledge distillation methods that explore different ways of transferring knowledge. For example, Zhang et al. (2020) proposed task-oriented feature distillation (TOFD) where several auxiliary classifiers are attached to different depths of a shared backbone network. Huang et al. (2022) proposed a promising feature map distillation framework of thin nets in an offline manner. Chen et al. (2021) revealed the significance of a cross-level connection path between teacher and student and proposed a new review mechanism, DRK, which uses shallower classifier/features of teacher to guide the student training. In order to obtain a good teacher, Park et al. (2021) further proposed to pre-train a student-friendly teacher by exploring some branches of the student network. However, a pre-trained teacher is not always necessary for online knowledge distillation. For example, Zhu and Gong (2018) proposed an on-the-fly native ensemble learning strategy for one-stage online distillation by constructing a multi-branch structure with each single branch as a student. Apart from additional teacher models, self-knowledge distillation aims to train a student by exploring its own knowledge. For example, Zhang et al. (2019) proposed a self-distillation (SD) training framework to distill knowledge within a network itself, which is similar to the deeply-supervised learning of the deep model itself (Sun et al., 2019). Additionally, Ji et al. (2021) proposed to utilize an auxiliary self-teacher network to transfer a refined knowledge for the classifier network.

Most aforementioned knowledge distillation methods follow a multi-branch approach, including both online and offline distillation, while all of them fail to consider the student self-reflection via reviewing its own knowledge. Meanwhile, self-distillation often distills the knowledge from the last layer of the student network, and a few distillation ones also distill knowledge from deeper layers to shallow layers within the student network itself (Zhang et al., 2019; Sun et al., 2019) or between the teacher and student networks (Chen et al., 2021). Besides, most existing knowledge distillation methods, which distill the knowledge at a certain layer of the teacher network using only one kind of distillation strategy, could not detect more informative knowledge from multiple layers of the teacher network. To address the issues, we take the advantages of both offline distillation and self-distillation into our proposed MTKD-SSR via both stage-wise and cross-stage distillation at different layers to improve the performance of student network.

2.2 Teacher's Knowledge

Different types of knowledge has been used for distillation, including logits (Zhao et al., 2022; Zhang et al., 2022; Phan et al., 2022), feature maps (Yim et al., 2017; Zagoruyko & Komodakis, 2017; Heo et al., 2019; Huang et al., 2022), and sample relationships (Tung & Mori, 2019; Yang et al., 2022). Recently, different types of channel features have also been explored as the knowledge for distillation (Shu et al., 2021; Liu et al., 2021; Li et al., 2021; Muhammad et al., 2021; Zhou et al., 2006; Qu et al., 2020; Fan et al., 2022; Ge et al., 2019). Specifically, Shu et al. (2021) proposed channel-wise knowledge distillation, which distills the knowledge from the channel-wise probability maps of the teacher network. Zhou et al. (2006) proposed channel distillation, which transfers the channel information from a teacher network to a student network. In (Li et al., 2021), the authors proposed a new channel correlation structure (CCS), which aims to guide the training of a student by applying fine-grained supervision, i.e., both inter- and intra-instance relationships. In (Liu et al., 2021), diversity-preserved knowledge was designed by discovering the teacher knowledge from inter-channel correlation. Qu et al. (2020) designed a hybrid attention transfer (H-AT), which calculates the channel-based attention knowledge through the output of the teacher's middle layers and then transfers it to the student network. Fan et al. (2022) proposed an online distillation method via channel self-supervision, which considers the sample, target, and network diversity knowledge on the dual-network multi-branch structure. Muhammad et al. (2021) proposed a new robust knowledge transfer via distilling activated channel maps to overcome the susceptibility of the natural samples. Besides, except knowledge distillation, the channel features have been fully used for improving the performance of the other deep learning methods (Lou & Loew, 2021; Chen et al., 2021; Yan et al., 2021). In this paper, we also use the channel-based knowledge to improve the distillation performance, but consider multiple types of knowledge from the channel features and the logit outputs at different layers, and accordingly both stage-wise channel distillation and stage-wise response distillation are devised.

Most importantly, we consider not only the importance of teaching from teacher model, but also the knowledge transferring from previous stage to later stages within the student network, which we refer to it as the student self-reflection or self-knowledge review. By doing this, we take into account the advantages of both offline and self-distillation. That is, we further enable student to review and then consolidate its previous learned knowledge.

3 Method

In this section, we introduce the proposed multi-target knowledge distillation via student self-reflection. The overall MTKD-SSR framework is shown in Fig. 1, where: (1) the red arrow indicates the stage-wise channel distillation (SCD); (2) the green arrow indicates the stage-wise response distillation (SRD); and (3) the orange arrow indicates the cross-stage review distillation (CRD).

3.1 Overview

Given a student model s and a teacher model t, let a dataset denote as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$ from the classes $\{1, 2, \ldots, C\}$, where y_i indicates the one-hot label and C is the number of categories. As shown in Fig. 1, we attach an auxiliary classifier associated with the fully connected (FC) layer to each block during training. The logits at the i-th auxiliary classifier for the input $x \in \mathcal{D}$ can be denoted as $z_i^s \in \mathbb{R}^C$ and $z_i^t \in \mathbb{R}^C$ for the student and teacher, respectively. Let p_i^s and p_i^t denote the prediction probabilities at the i-th auxiliary classifier of the student and the teacher models, respectively. We then have the probabilities of the c-th category as

$$p_i^s(c) = \frac{\exp\left(z_i^s(c)/T\right)}{\sum_{j=1}^{C} \exp\left(z_i^s(j)/T\right)}, \tag{1}$$

$$p_i^t(c) = \frac{\exp\left(z_i^t(c)/T\right)}{\sum_{j=1}^{C} \exp\left(z_i^t(j)/T\right)}, \tag{2}$$

where $T > 0$ indicates the distillation temperature to smooth the probability.

Knowledge distillation methods mainly transfer the knowledge (either logits or features) from the teacher to the student model, but there is no attempt to distill the knowledge with several shallow blocks separately, thus failing to consider the knowledge transfer among different stages of the student network. The key difference between previous methods and the proposed method is that: previous methods usually consider this problem from a teacher view, while we take the advantages of both offline and self-distillation from a student view. That is, the knowledge should be transferred from not only the teacher model but also the student model itself, i.e., a student self-reflection via reviewing the knowledge from its previous stages. Therefore, we argue that the student knowledge review is of great importance for knowledge distillation and the overall framework is shown in Fig. 1, where we divide the student and teacher networks into several blocks/stages (e.g., according to the residual blocks in ResNets). In this paper, if not otherwise stated, we use four blocks for knowledge distillation and the auxiliary classifiers are utilized help training early blocks. All auxiliary classifiers are only applied to improve training student network in knowledge distillation period, which will be removed during the inference stage. Notably, during training, the auxiliary FC layer corresponding to the attached classifier at each block is first randomly initialised and then jointly optimized; During testing, all auxiliary classifiers can be removed and thus will not bring any extra parameters or computational cost.

In previous knowledge distillation methods, the knowledge from channel features and logit outputs is just transferred from the teacher network to the student network directly. Not only does our proposed method transfer the knowledge in the form of channel features and logit outputs from the teacher network to the student network, but also strengthens the student's early blocks via transferring knowledge from early blocks to deep blocks. Furthermore, we propose to bring the cross-block connections into the student model, where knowledge is transferred from shallow blocks to deep blocks, namely student knowledge review. Specifically, we use i-th block to review all previous shallow blocks' knowledge. We introduce the details of three knowledge distillation paths used in the proposed method as follows.

3.2 Stage-Wise Response Distillation

Previous knowledge distillation methods often transfer knowledge via minimizing the Kullback–Leibler (KL) divergence between the logits from the last layer of the teacher network and those from the last layer of the student network. Our method splits both the student and the teacher networks into several same level blocks, each of which is associated with an auxiliary classifier to get the softmax outputs. That is to say, each block's auxiliary classifier of the teacher network corresponds to the same block's auxiliary classifier of the student network. The logit outputs as the knowledge will then be distilled from the teacher network's auxiliary classifier to the student network's auxiliary classifier at the same stage, which can enable student network to learn more knowledge from teacher network at different levels. The distillation loss between student network and teacher network is formulated as:

$$\mathcal{L}_{SRD} = \sum_{i=1}^{K} \mathcal{L}_{KL}\left(p_i^s, p_i^t\right), \tag{3}$$

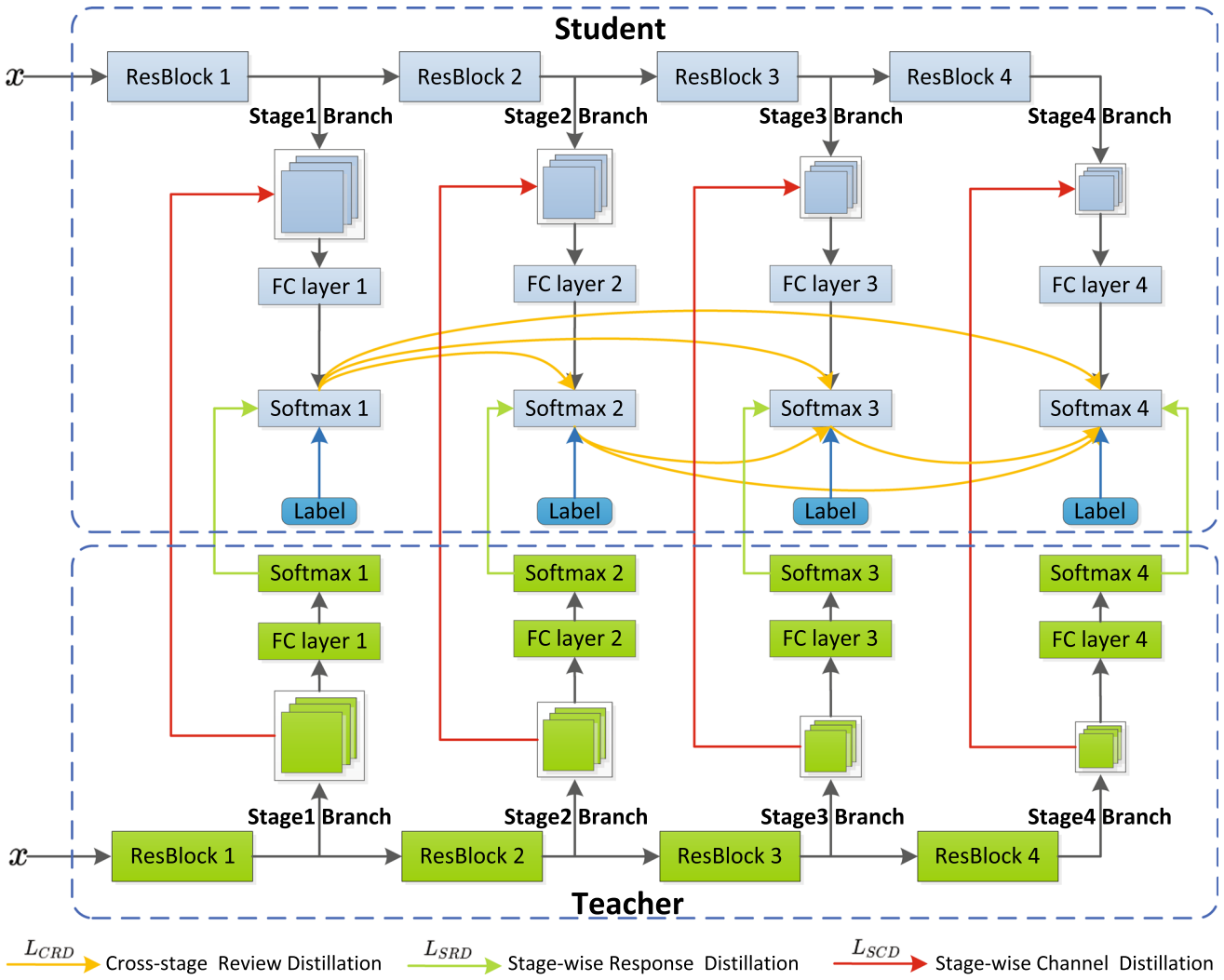


Fig. 1 The main MTKD-SSR framework. Specifically, stage-wise response distillation transfers the logits of multiple stages/blocks as the knowledge from teacher to student; stage-wise channel distillation enables the student to match the channel features of the teacher at each stage/block; cross-stage review distillation explores the self-learning for transferring the shallow blocks' knowledge to the learning of the deep blocks within the student network (Color figure online)

where K is the number of auxiliary classifiers. Note that the temperature used in p_i^s and p_i^t is denoted as T_1 according to Eq. (1). The superscripts s and t represent the student network and the teacher network, respectively. $\mathcal{L}_{KL}(.)$ represents the Kullback–Leibler (KL) divergence, which is defined as:

$$\mathcal{L}_{KL}\left(p_i^s, p_i^t\right) = \sum_{c=1}^{C} p_i^t(c) \log \frac{p_i^t(c)}{p_i^s(c)} . \tag{4}$$

By using the multi-stage offline teacher-student knowledge distillation, we allow each student's shallow/deep block to mimic the corresponding teacher's shallow/deep block via learning from teacher logit knowledge.

3.3 Cross-Stage Review Distillation

Existing teacher-student knowledge distillation methods mainly focus on how to make full use of the teacher network to transfer more knowledge to the student network. In contrast, student self-distillation methods use a pre-trained student network as a teacher to teach a same randomly initialized student network, and transfer knowledge from one part of student network to guide the learning of other parts of the same student network. Generally, self-distillation methods always adopt the outputs of the last layer of the pre-trained student network as the regularization to guide the learning of the same student network (Yuan et al., 2020), and they also explore the knowledge from the subsequent layers of the student network to its preceding layers (Zhang et al., 2019). It

is argued that the subsequent layers contain more compact semantic information but less redundant useless information, so that it is naturally assumed that distilling the knowledge from subsequent layers to the preceding ones can be better. However, we empirically found (shown in Sect. 5) that the self-distillation performance can be improved by distilling the knowledge from the preceding layers of the student network to its subsequent layers. The possible reason is that the preceding layers contain more dark knowledge (Hinton et al., 2015) that is constructive for self-distillation. From the student perspective, the knowledge from the preceding layers to its subsequent ones can be regarded as the previously learned knowledge, and the corresponding self-distillation is the self-reflection via reviewing the previously learned knowledge from the preceding layers. Thus, the student performance can be improved by consolidating the previously learned knowledge via self-distillation.

Here, we propose a new student self-distillation method called knowledge review distillation based on the real-world learning practice, known as self-reflection. It is noteworthy that our knowledge review distills the logit information from the shallow layers to the deep layers, which is in the reverse direction of standard self-distillation. Particularly, it can also be shown in Fig. 1. In such an approach, the student network is able to continuously review the knowledge that has been gained so far at each stage, allowing itself to reflect on its learning. Our proposed method makes up for the deficiency of using student network's last block to improve self performance. The i-th shallow block's knowledge review can be formulated as:

$$\mathcal{L}_{CRD}(i) = \sum_{j=1}^{i-1} \mathcal{L}_{KL}\left(p_j^s, p_i^s\right) . \tag{5}$$

That is, the i-th shallow block reviews all previous $i-1$ shallow blocks' logit knowledge. The temperature in p_j^s and p_i^s is denoted as T_2 according to Eq. (1). The total student knowledge review loss can be formulated as:

$$\mathcal{L}_{CRD} = \sum_{i=2}^{K} \mathcal{L}_{CRD}(i) . \tag{6}$$

Notably, cross-stage review distillation contains different student self-reflections during the multi-stage learning.

3.4 Stage-Wise Channel Distillation

Besides the teacher network's logit knowledge, we also consider the channel knowledge transferred from teacher network to student at multiple blocks/stages, in order to further enhance student's performance. Here, we adapt channel-wise attention (Zhou et al., 2006), which can enhance the important channels and weaken the insignificant ones, to characterize the knowledge from each channel. The channel knowledge contained in channel maps will be transferred at each stage of training process.

The channel distillation between the teacher and the student networks at i-th shallow block can be formulated as:

$$\mathcal{L}_{SCD}(i) = \frac{\sum_{n=1}^{B}\sum_{j=1}^{D}\left(W_{nj}^t - W_{nj}^s\right)^2}{B \times D} , \tag{7}$$

where the W_{nj}^t and W_{nj}^s indicate the weights of the j-th channel for the teacher and student networks due to the n-th sample, respectively. D represents the number of channels and B is the mini-batch size. Each channel's weight can be evaluated as (Hu et al., 2018; Zhou et al., 2006):

$$W_l = \frac{1}{H \times Y}\sum_{i=1}^{H}\sum_{j=1}^{Y} A(i, j), \tag{8}$$

where W_l denotes the weight of the l-th channel, both H and Y indicate the spatial dimensions of the feature maps, and $A(\cdot)$ is the activation function. The channel's weights indicate the channel-wise statistics, which can be obtained by squeezing the features, and contain global spatial information. In most circumstances, the teacher network is more accurate than the student network, thus the former can often have a better estimated weight distribution in feature map. The channel knowledge distillation will help the student network to mimic teacher's weight distribution in channel feature maps to improve the shallow blocks and also strengthen the method of student knowledge review. Lastly, if the channel numbers of the teacher and the student networks are mismatch, we will use an additional 1×1 convolution layer. The overall channel distillation loss at different stages can be rewritten as

$$\mathcal{L}_{SCD} = \sum_{i=1}^{K} \mathcal{L}_{SCD}(i) . \tag{9}$$

Through the stage-wise channel distillation, the student network can learn more important information of channel feature maps at each block/stage from the teacher network. The intuitive diagram of our designed stage-wise channel distillation method is shown in Fig. 2.

3.5 MTKD-SSR Framework

Our MTKD-SSR contains three different target knowledge distillation modules that function together to realise the learning principles that are multi-stage learning from teacher to student via offline distillation and student self-reflection from the preceding layers of the student network to its subsequent

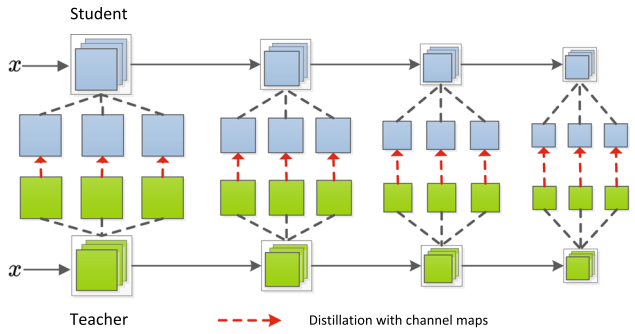


Fig. 2 Stage-wise channel distillation in our MTKD-SSR. In the SCD, the channel maps at multi-layers of the teacher network are modelled as the distilled knowledge, and the student matches the channel knowledge from teacher at different stages

layers via self-distillation. The two multi-stage learning modules allow the student network to learn two different types of knowledge, i.e., logits and channel features, at different levels, and the student self-reflection allows the student to continuously improve itself via learning form the past. The overall loss function of MTKD-SSR including those three target knowledge distillation modules is thus defined as follows:

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{SRD} + \beta \mathcal{L}_{CRD} + \gamma \mathcal{L}_{SCD} + \mathcal{L}_{CE} \,, \tag{10}$$

where α, β, and γ are used to balance different distillation losses. $\mathcal{L}_{CE}$ indicates the cross entropy loss for all the stages of student as follows:

$$\mathcal{L}_{CE}\left(y, p^s\right) = \sum_{i=1}^{K} \sum_{c=1}^{C} -y(c) \log p_i^s(c) \,, \tag{11}$$

where y is the one-hot label vector for an input x.

3.6 Discussion

In this subsection, we highlight the contributions of the proposed MTKD-SSR framework via its differences with the following related knowledge distillation methods (Chen et al., 2021; Heo et al., 2019; Sun et al., 2019; Tung & Mori, 2019; Yim et al., 2017; Zagoruyko & Komodakis, 2017). Specifically, knowledge transfer in the proposed method is formed by both stage-wise and cross-stage distillation in a unified framework: (1) inspired by the multi-stage human learning scheme, multi-stage distillation transfers knowledge from the stage-wise logits and stage-wise channel feature maps; and (2) inspired by the student self-reflection, cross-stage distillation transfers knowledge from shallow layers to deep layers within the same network.

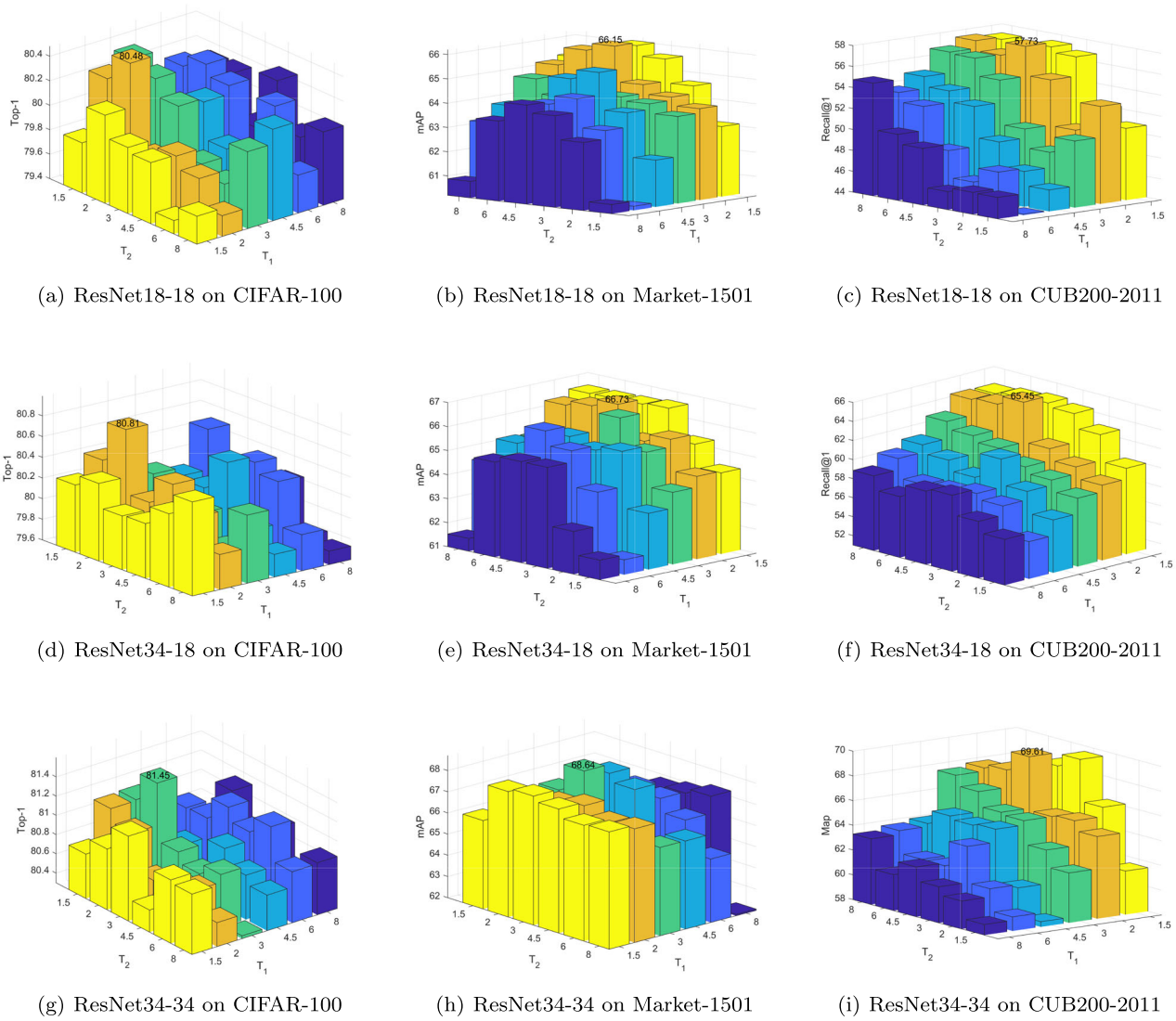
Firstly, the multi-stage knowledge has been used in some existing knowledge distillation methods: Zagoruyko and Komodakis (2017) proposed a stage-wise attention transfer (AT) method via distilling the spatial attention maps of each layer from teacher to student; Yim et al. (2017) proposed a feature distillation method that first designs the flow of solution procedure (FSP) matrix between layers and then the layer-wise FSP of the teacher as the knowledge guides the student; Tung and Mori (2019) extended AT by a similarity-preserving knowledge distillation (SP) method that models the layer-wise attention-based similarity between samples as the knowledge. Heo et al. (2019) proposed a comprehensive overhaul of feature distillation (OFD), that is the layer-wise feature distillation using marginal ReLU. In summary, AT, FSP, and OFD are the multi-stage feature knowledge distillation, while SP is the multi-stage relational knowledge distillation. Different from the above-mentioned methods, MTKD-SSR contains two types of multi-stage distillation schemes, stage-wise response distillation (SRD) and stage-wise channel distillation (SCD).

Secondly, the cross-stage knowledge transfer has been explored in a few knowledge distillation methods (Chen et al., 2021; Sun et al., 2019). As a strategy to train deep model, Sun et al. (2019) proposed deeply-supervised knowledge synergy (DKS) with three pairwise knowledge matching methods. Though the cross-stage review distillation (CRD) in our MTKD-SSR is structured in a similar way to one of the knowledge matching methods in DKS, our CRD aims to mimic the self-reflection ability in human learning and thus form a new self-distillation method for model compression. Recently, Chen et al. (2021) proposed the offline knowledge distillation via knowledge review (DKR), which is the cross-stage distillation from different layers of the teacher to the last layer of the student. Compared to DKR, our CRD as self-distillation is the cross-stage distillation from shallow layers to deep layers with the same network.

Lastly, all above-mentioned recent methods (AT, FSP, SP, OFD, DKS, and DKR) transfer knowledge using only one kind of distillation scheme with one type of knowledge. Different from this, the proposed MTKD-SSR employs a multiple knowledge transfer scheme with offline and self distillation to simultaneously distill both logits and channel knowledge.

4 Experiments

In all experiments, we adopt ResNet (He et al., 2016) and SENet (Hu et al., 2018) as the backbone network to form the teacher-student architecture and evaluate the proposed MTKD-SSR method on the five popular visual recognition datasets, CIFAR-100 (Krizhevsky & Hinton, 2009), Market-1501 (Zheng et al., 2015), CUB200-2011 (Wah et al., 2011), Pascal VOC (Everingham et al., 2010), and ImageNet (Deng et al., 2009). We compare MTKD-SSR with



(a) ResNet18-18 on CIFAR-100 (b) ResNet18-18 on Market-1501 (c) ResNet18-18 on CUB200-2011

(d) ResNet34-18 on CIFAR-100 (e) ResNet34-18 on Market-1501 (f) ResNet34-18 on CUB200-2011

(g) ResNet34-34 on CIFAR-100 (h) ResNet34-34 on Market-1501 (i) ResNet34-34 on CUB200-2011

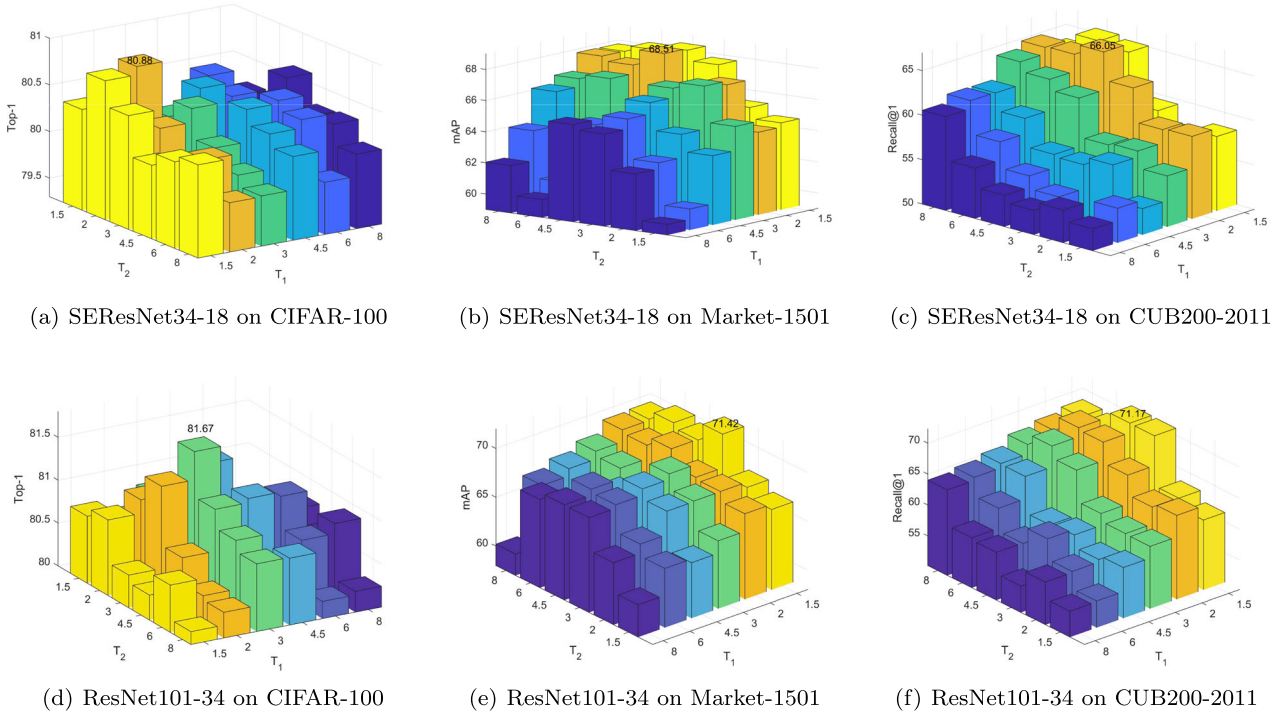
Fig. 3 Results on CIFAR-100, the mAP accuracy (%) on Market-1501, and the Recall@1 accuracy (%) on CUB200-2011 with varying temperatures T_1 and T_2 using ResNet18-18, ResNet34-18 and ResNet34-34 as the teacher-student architectures

the recent state-of-the-art knowledge distillation methods that are the standard knowledge distillation (KD) (Hinton et al., 2015), self-distillation (SD) (Zhang et al., 2019), hybrid attention transfer (H-AT) (Qu et al., 2020), knowledge distillation via channel correlation structure (CCS) (Li et al., 2021), task-oriented feature distillation (TOFD) (Zhang et al., 2020), distillation via knowledge review (DKR) (Chen et al., 2021), and variational information distillation (VID) (Ahn et al., 2019). We describe the teacher-student architectures as follows: ResNet34-18 means that ResNet34 is used as the teacher network and ResNet18 is used as the student network; other architectures are defined in a similar way. For semantic segmentation, we compare our MTKD-SSR to SKD (Liu et al., 2020), IFVD (Wang et al., 2020) and CWD (Shu et al., 2021), where PSPNet (Zhao et al., 2017) with different ResNet backbone networks are used for the teacher-student architecture.

4.1 Datasets and Experimental Setups

We perform extensive experiments on the following five visual recognition datasets:

– **CIFAR-100** (Krizhevsky & Hinton, 2009). It has 60,000 images collected from 100 classes, including 50,000 training samples and 10,000 test samples, the size of which is 32 × 32 pixels. All the models were trained with batch size 128 and 200 epochs. The initial learning rate is 0.1 and then is further divided by 10 at the 60-th, 120-th, and 160-th epochs.



(a) SEResNet34-18 on CIFAR-100 (b) SEResNet34-18 on Market-1501 (c) SEResNet34-18 on CUB200-2011

(d) ResNet101-34 on CIFAR-100 (e) ResNet101-34 on Market-1501 (f) ResNet101-34 on CUB200-2011

Fig. 4 Results on CIFAR-100, the mAP accuracy (%) on Market-1501, and the Recall@1 accuracy (%) on CUB200-2011 with varying temperatures T_1 and T_2 using SEResNet34-18 and ResNet101-34 as the teacher-student architectures

- **Market-1501** (Zheng et al., 2015). It contains 32,217 images of 1,501 pedestrians captured by 6 cameras, each of which consists of 128 × 64 pixels. We set the initial learning rate to 0.1, which is then divided by 10 at the 40-th epoch, batch size to 32 and the number of epochs to 60.
- **CUB200-2011** (Wah et al., 2011). It has 11,788 bird images collected from 200 bird sub-categories. It has been divided into the training set with 5,994 images and the test set with 5,794 images. With batch size 32, all the models were trained for 90 epochs. The initial learning rate is 0.1 and then is multiplied by 0.1 at the 30-th, 60-th and 80-th epochs.
- **ImageNet** (Deng et al., 2009). It has 1,000 image categories. Its training set has 1.2 million images, including 50,000 validation images and 100,000 test images. Similarly, we trained all the models with 128 batch size for 90 epochs. The initial learning rate is 0.1 and then is further divided by 10 at the 30-th, 60-th and 80-th epochs.
- **Pascal VOC** (Everingham et al., 2010). It contains Pascal VOC 2012 and SBD. It has been divided into the training set with 10,582 images and the test set with 1,449 images. With batch size 16, all the model were trained for 100 epochs. The initial learning rate is 0.01 and is then further divided by 10 at the 30-th, 60-th, and 90-th epochs.

4.2 Analysis

In this section, we explore the effect of different temperatures on distillation performance with different teacher-student architectures. In MTKD-SSR, the temperature T_1 in Eq. (3) is used to soften the transferred knowledge on the teacher-student architecture during the stage-wise response distillation, and T_2 in Eq. (5) is used to soften the knowledge within the student network during the cross-stage review distillation. As stated in (Chen et al., 2021; Jafari et al., 2021), a proper temperature makes the transferred knowledge informative for improving the performance. Figures 3 and 4 show the Top-1 accuracy of our MTKD-SSR on CIFAR-100, the mAP accuracy on Market-1501, and the Recall@1 accuracy on CUB200-2011 with different temperatures T_1 and T_2 in $\{1.5, 2.0, 3.0, 4.5, 6.0, 8.0\}$. We see that the temperature has a large impact on the distillation performance. Specifically, we use the temperatures as $T_1 = 2.0$ or 3.0 and $T_2 = 2.0$ on CIFAR-100, $T_1 = 1.5, 2.0$ or 3.0 and $T_2 = 3.0$ or 4.5 on Market-1501, and $T_1 = 1.5$ or 2.0 and $T_2 = 4.5$ or 6.0 on CUB200-2011. The temperature can not only form the soften informative knowledge during the stage-wise distillation, but also adjust the reviewed knowledge within the student structure during cross-stage distillation. This also implies that the knowledge at each stage of student plays a different role for student self-reflection. The possible reason

for the cross-stage review distillation is that the introduction of the temperature parameter can further well distill the dark knowledge from the preceding layers of the student network to its subsequent layers. Through the introduction of two temperatures in the proposed MTKD-SSR, the dark knowledge contained in both teacher and student networks can be fully explored for learning the student.

4.3 Results

To evaluate the proposed MTKD-SSR method, we conduct comparative experiments on different tasks, CIFAR-100 and ImageNet for image classification, Market-1501 for person re-identification, CUB200-2011 for image retrieval, and Pascal VOC for segmentation. In our experiments, we adopt five teacher-student architectures, ResNet34-34, ResNet34-18, ResNet18-18, SEResNet34-18, and ResNet101-34. Note that since the compared SD is a self-distillation method with the same teacher and student networks, there is no performance reported on ResNet34-18, SEResNet34-18, and ResNet101-34. In each table, the ↑ indicates the improvement over the baseline.

Results on CIFAR-100. In Table 1, we see that our MTKD-SSR significantly outperforms all compared methods. Specifically, the accuracy improvements of our MTKD-SSR over SD (Zhang et al., 2019) with the second highest accuracy are 1.42% on ResNet34-34 and 1.60% on ResNet18-18, and its improvements over TOFD (Zhang et al., 2020) with the second highest accuracy are 2.17% on ResNet34-18, 1.97% on SEResNet34-18, and 0.99% on ResNet101-34. Moreover, the student trained via our method outperforms its teacher even if both teacher and student share the same structures. For example, the accuracy improvements of student over its teacher are 2.95% on ResNet34-18 and 3.09% on SEResNet34-18. Notably, SD (Zhang et al., 2019) transfers knowledge from the deep-level to the shallow-level classifiers via self-distillation and obtains better performance than KD, H-AT, TOFD, CCS, DKR and VID. This implies that the student can improve itself by self-reflective learning without the capacity gap between the teacher and student. Compared to SD, the proposed MTKD-SSR method distills the knowledge from the shallower layers to the deeper ones via self-reflective learning. Therefore, our MTKD-SSR using cross-stage review distillation is a promising knowledge distillation method to explore the dark knowledge from shallow to deep.

Furthermore, we visualize the feature representations of the student ResNet18 trained by our MTKD-SSR and the compared methods on CIFAR-100 using t-SNE (Van der Maaten & Hinton, 2008). As shown in Fig. 5, among the compared methods, TOFD, CCS, DKR and SD are very related to our proposed method. CCS designs the channel feature knowledge transfer, DKR reviews the shallower-level features of teacher to guide the learning of the deeper-level features of student, TOFD adopts the block-wise knowledge transfer, and SD transfers the deepest-level knowledge to guide the shallower levels via self-distillation. We can see that our proposed MTKD-SSR has better compact class-specific clusters than the compared methods. Though the related methods (i.e., TOFD, CCS, DKR and SD) are more or less similar to our MTKD-SSR in the types of knowledge and distillation strategies, our MTKD-SSR learns more discriminative representation.

Results on Market-1501. In Table 2, we see that the proposed MTKD-SSR achieves the highest performance among all the compared methods, where our trained student also outperforms its guided teacher. Specifically, the mAP accuracy improvements of our trained student over its teacher are 5.88% on ResNet34-34, 3.97% on ResNet34-18, 4.47% on ResNet18-18, 3.65% on SEResNet34-18, and 5.06% on ResNet101-34. This fact means the proposed MTKD-SSR can distills more informative knowledge contained in both teacher and student networks for the student learning. Interestingly, when teacher and student have the same model structures, student surpasses teacher with a large margin. The possible reason is that no teacher-student capacity gap exists for favorable performance and our cross-stage self-distillation enhance the student learning. We also find that TOFD with block-wise distillation, SD with self-learning and DKR with knowledge review always perform better than KD, H-AT, CCS and VID.

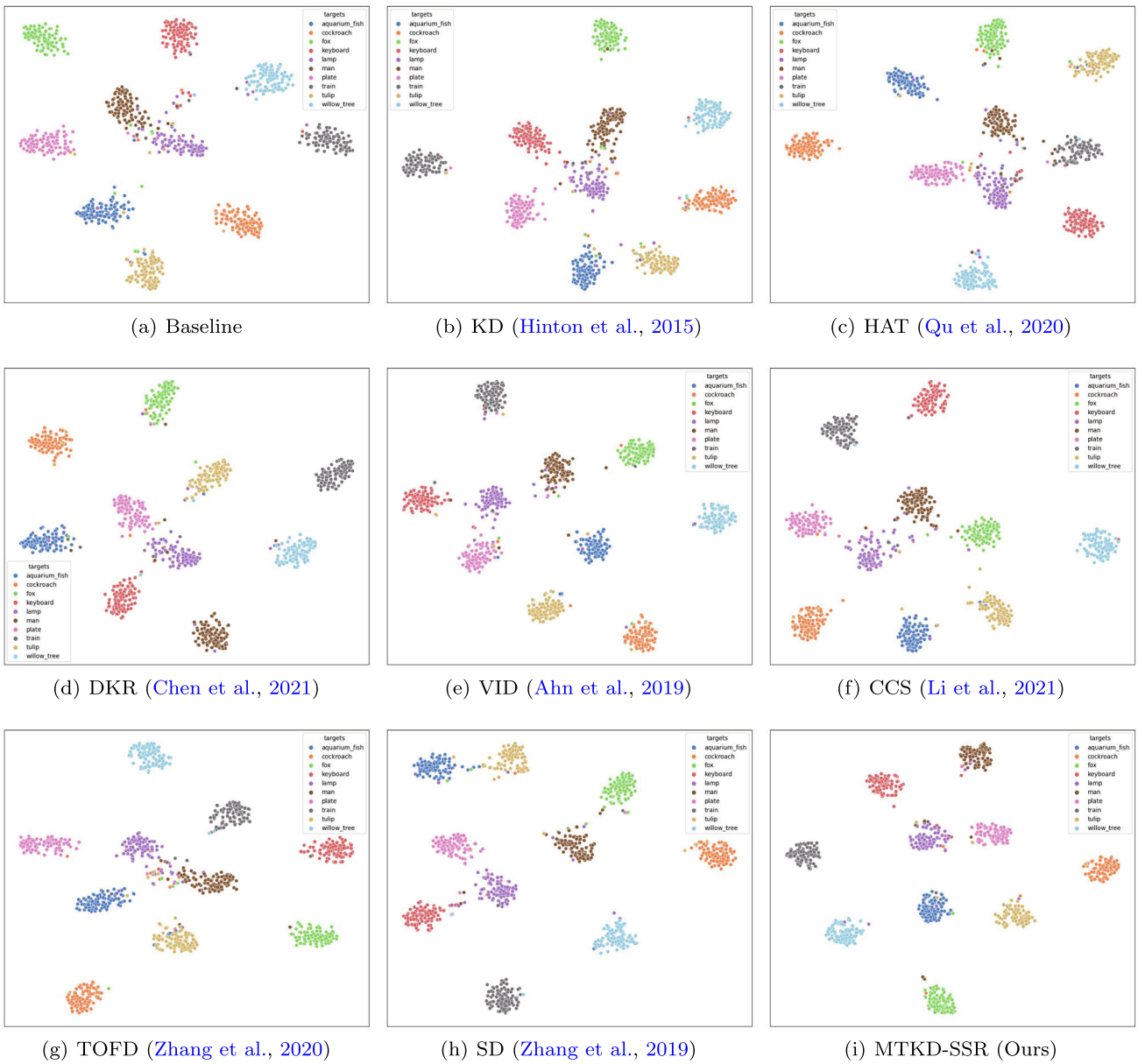
Results on CUB200-2011. Table 3 shows the Recall@1 accuracy of all the compared methods. We can see that the proposed MTKD-SSR significantly outperforms all the other compared methods with a large margin. Especially, our trained student performs very better than its guided teacher. Meanwhile, DKR with knowledge review and CCS with channel feature knowledge always obtain the better performance than KD, H-AT, SD, VID, and TOFD. This somewhat implies that the channel knowledge used in our stage-wise channel distillation and self-review used in our cross-stage review distillation can well improve the knowledge distillation performance.

Results on ImageNet: Table 4 shows the Top-1 accuracy of our MTKD-SSR and the compared methods. It should be noted that the top-1 accuracy of SD is obtained by the self-learning of the student ResNet18 without the guidance of the teacher ResNet34. We can see that our method performs better than all the compared methods, and our improvement over DKR (Chen et al., 2021) with the second highest accuracy is 0.64%. Meanwhile, compared to the other compared methods, DKR with review mechanism, TOFD with block-wise distillation and SD with self-learning perform better, which further verifies the feasibility and effectiveness of our MTKD-SSR, which adapts knowledge review, block-wise

Table 1 Results on CIFAR-100

Teacher-student architecture	ResNet34-34	ResNet34-18	ResNet18-18	SEResNet34-18	ResNet101-34
Teacher	77.86	77.86	77.52	77.79	80.21
Student	–	77.52	–	77.43	77.86
KD (Hinton et al., 2015)	77.98 (↑ 0.12)	77.73 (↑ 0.21)	77.64 (↑ 0.12)	77.74 (↑ 0.31)	77.99 (↑ 0.13)
DKR (Chen et al., 2021)	78.74 (↑ 0.88)	78.55 (↑ 1.03)	78.29 (↑ 0.77)	78.23 (↑ 0.80)	78.94 (↑ 1.08)
H-AT (Qu et al., 2020)	78.76 (↑ 0.90)	78.14 (↑ 0.62)	78.01 (↑ 0.49)	77.92 (↑ 0.49)	79.03 (↑ 1.17)
VID (Ahn et al., 2019)	78.95 (↑ 1.09)	78.56 (↑ 1.04)	78.33 (↑ 0.81)	78.33 (↑ 0.90)	79.21 (↑ 1.35)
CCS (Li et al., 2021)	79.18 (↑ 1.32)	78.63 (↑ 1.11)	78.21 (↑ 0.69)	78.47 (↑ 1.04)	79.59 (↑ 1.73)
TOFD (Zhang et al., 2020)	79.28 (↑ 1.42)	78.64 (↑ 1.12)	77.92 (↑ 0.40)	78.91 (↑ 1.48)	80.68 (↑ 2.82)
SD (Zhang et al., 2019)	80.03 (↑ 2.17)	–	78.88 (↑ 1.36)	–	–
MTKD-SSR	**81.45 (↑ 3.59)**	**80.81 (↑ 3.29)**	**80.48 (↑ 2.96)**	**80.88 (↑ 3.45)**	**81.67 (↑ 3.81)**

Bold values indicate the best performance among all the methods



(a) Baseline (b) KD (Hinton et al., 2015) (c) HAT (Qu et al., 2020)

(d) DKR (Chen et al., 2021) (e) VID (Ahn et al., 2019) (f) CCS (Li et al., 2021)

(g) TOFD (Zhang et al., 2020) (h) SD (Zhang et al., 2019) (i) MTKD-SSR (Ours)

Fig. 5 Visualization of the student model features on CIFAR-100 when using ResNet34-18 as the teacher-student architecture

Table 2 Results on Market-1501

Teacher-student architecture	ResNet34-34	ResNet34-18	ResNet18-18	SEResNet34-18	ResNet101-34
Teacher	62.76	62.76	61.68	64.86	66.36
Student	–	61.68	–	55.48	62.76
KD (Hinton et al., 2015)	62.82 (↑ 0.06)	61.89 (↑ 0.21)	61.76 (↑ 0.08)	57.80 (↑ 2.32)	63.38 (↑ 0.62)
DKR (Chen et al., 2021)	64.63 (↑ 1.87)	64.36 (↑ 2.68)	63.04 (↑ 1.36)	64.94 (↑ 9.46)	65.11 (↑ 2.35)
H-AT (Qu et al., 2020)	64.25 (↑ 1.49)	63.76 (↑ 2.08)	62.98 (↑ 1.30)	65.87 (↑ 10.39)	66.72 (↑ 3.96)
VID (Ahn et al., 2019)	63.21 (↑ 0.45)	62.73 (↑ 1.05)	62.31 (↑ 0.63)	60.55 (↑ 5.07)	66.50 (↑ 3.74)
CCS (Li et al., 2021)	62.89 (↑ 0.13)	62.47 (↑ 0.79)	62.34 (↑ 0.66)	59.07 (↑ 3.59)	64.55 (↑ 1.79)
TOFD (Zhang et al., 2020)	65.37 (↑ 2.61)	64.69 (↑ 3.01)	63.63 (↑ 1.95)	66.44 (↑ 10.96)	68.12 (↑ 5.36)
SD (Zhang et al., 2019)	65.19 (↑ 2.43)	–	64.22 (↑ 2.54)	–	–
MTKD-SSR	**68.64 (↑ 5.88)**	**66.73 (↑ 5.05)**	**66.15 (↑ 4.47)**	**68.51 (↑ 13.03)**	**71.42 (↑ 8.66)**

Bold values indicate the best performance among all the methods

Table 3 Results on CUB200-2011

Teacher-student architecture	ResNet34-34	ResNet34-18	ResNet18-18	SEResNet34-18	ResNet101-34
Teacher	48.49	48.49	47.55	49.01	50.51
Student	–	47.55	–	46.64	48.49
KD (Hinton et al., 2015)	50.47 (↑ 1.98)	49.44 (↑ 1.89)	47.81 (↑ 0.26)	50.17 (↑ 3.53)	51.86 (↑ 3.37)
DKR (Chen et al., 2021)	61.52 (↑ 13.03)	58.67 (↑ 11.12)	55.92 (↑ 8.37)	59.47 (↑ 12.83)	64.99 (↑ 16.50)
H-AT (Qu et al., 2020)	61.58 (↑ 13.09)	59.63 (↑ 12.08)	52.01 (↑ 4.46)	53.19 (↑ 6.55)	61.86 (↑ 13.37)
VID (Ahn et al., 2019)	61.07 (↑ 12.58)	56.53 (↑ 8.98)	49.32 (↑ 1.77)	55.52 (↑ 8.88)	61.68 (↑ 13.19)
CCS (Li et al., 2021)	66.13 (↑ 17.64)	63.31 (↑ 15.76)	52.43 (↑ 4.88)	58.63 (↑ 11.99)	66.45 (↑ 17.96)
TOFD (Zhang et al., 2020)	53.01 (↑ 4.52)	52.14 (↑ 4.59)	50.63 (↑ 3.08)	52.63 (↑ 5.99)	63.98 (↑ 15.79)
SD (Zhang et al., 2019)	54.37 (↑ 5.88)	-	51.62 (↑ 4.07)	–	–
MTKD-SSR	**69.61 (↑ 21.12)**	**65.45 (↑ 17.90)**	**57.73 (↑ 10.18)**	**66.05 (↑ 19.41)**	**71.17 (↑ 23.22)**

Bold values indicate the best performance among all the methods

Table 4 Results on ImageNet

Teacher-student architecture	ResNet34-18
Accuracy	Top-1
Teacher	73.73
Student	70.37
KD (Hinton et al., 2015)	70.96 (↑ 0.59)
DKR (Chen et al., 2021)	71.99 (↑ 1.62)
H-AT (Qu et al., 2020)	71.55 (↑ 1.18)
VID (Ahn et al., 2019)	71.15 (↑ 0.78)
CCS (Li et al., 2021)	71.29 (↑ 0.92)
TOFD (Zhang et al., 2020)	71.93 (↑ 1.56)
SD (Zhang et al., 2019)	71.81 (↑ 1.44)
MTKD-SSR	**72.63 (↑ 2.26)**

Bold value indicates the best performance among all the methods

Table 5 Results on Pascal VOC

Algorithm	mIoU
Teacher PSPNET-ResNet101	78.52
Student PSPNET-ResNet18	64.52
SKD (Liu et al., 2020)	68.90 (↑ 4.38)
IFVD (Wang et al., 2020)	69.06 (↑ 4.54)
CWD (Shu et al., 2021)	70.27 (↑ 5.75)
MTKD-SSR	**70.84 (↑ 6.32)**

Bold value indicates the best performance among all the methods

distillation and self-learning into the proposed distillation framework.

Results on Pascal VOC. Table 5 shows the performance of the proposed MTKD-SSR in terms of the mean intersection over union (mIoU). We use PSPNet (Zhao et al., 2017) as the baseline with ResNet101 in teacher and ResNet18 in student during knowledge distillation. In our experiments, the optimal temperatures $T_1 = 2$ and $T_2 = 3$ are determined by the same way as in Sect. 4.2. Specifically, we find that our method outperforms all other methods, e.g., our improvement over CWD with the second highest accuracy is 0.57%. Meanwhile, CWD with channel-wise distillation outperforms SKD with pixel-wise and pair-wise distillation and

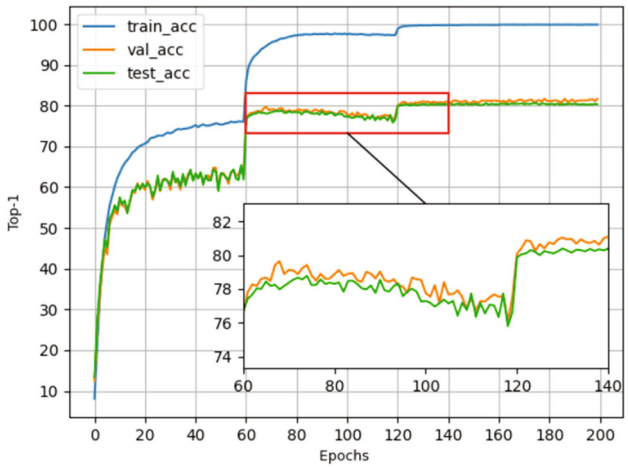


Fig. 6 Accuracy curves on the CIFAR-100 dataset

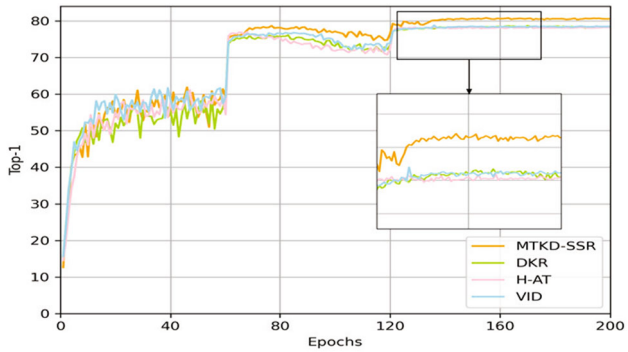


Fig. 7 Accuracy curves comparing with other methods on the CIFAR-100 dataset

IFVD with intra-class feature variation distillation. In Table 5, semantic segmentation experiments further demonstrate the effectiveness of the proposed distillation framework with a good generalization performance on downstream tasks.

4.4 Analysis of Accuracy Curves

To better understand the effectiveness of MTKD-SSR, we report the experimental results at each training epoch on the CIFAR-100 dataset using ResNet34-18 as the teacher-student architecture as follows. We divide CIFAR-100 dataset into the training set with 45,000 samples, the validation set with 5,000 samples, and the test set with 10,000 samples. As illustrated in Fig. 6, train_acc, val_acc and test_acc denote the Top-1 accuracy on the training, validation and test sets, respectively. We find that the classification performance on the training set is very close to 100%, where the validation and testing accuracies share almost the same pattern at each training epoch. Additionally, we compare the accuracy curves of our MTKD-SSR and the related knowledge distillation methods (i.e., DKR, H-AT and VID) in Fig. 7. Specifically, we see that the proposed MTKD-SSR clearly outperforms DKR, H-AT, and VID after 60 epochs.

Table 6 Ablation study on distillation mechanisms using ResNet34-18 on CIFAR-100

Variants	SRD	CRD	SCD	Top-1
MTKD-SSR	✓	✓	✓	80.81
A	✓	✓	×	79.96
B	×	✓	✓	79.51
C	✓	×	✓	79.85
D	✓	×	×	79.65
E	×	✓	×	78.97
F	×	×	✓	79.08

5 Ablation Study

In this section, we further evaluate the proposed MTKD-SSR from the perspective of different distillation mechanisms, different self-reflections, different feature knowledge and different self-distillation methods, to analyze the effectiveness of each applied module.

5.1 Ablation Study on Distillation Mechanisms

The proposed MTKD-SSR model mainly has three distillation mechanisms: stage-wise channel distillation (SCD), stage-wise response distillation (SRD), and cross-stage review distillation (CRD). To explore the impact of the certain distillation on the performance of MTKD-SSR, we conducted the ablation experiments about the following six variants: 1) Variant A with both SRD and CRD formulated as $\alpha\mathcal{L}_{SRD} + \beta\mathcal{L}_{CRD} + \mathcal{L}_{CE}$; 2) Variant B with both CRD and SCD formulated as $\beta\mathcal{L}_{CRD} + \gamma\mathcal{L}_{SCD} + \mathcal{L}_{CE}$; 3) Variant C with both SRD and SCD formulated as $\alpha\mathcal{L}_{SRD} + \gamma\mathcal{L}_{SCD} + \mathcal{L}_{CE}$; 4) Variant D only with SRD formulated as $\alpha\mathcal{L}_{SRD} + \mathcal{L}_{CE}$; 5) Variant E only with CRD formulated as $\beta\mathcal{L}_{CRD} + \mathcal{L}_{CE}$; 6) Variant F only with SCD formulated as $\gamma\mathcal{L}_{SCD} + \mathcal{L}_{CE}$. In Table 6, we have the following remarks: 1) The proposed MTKD-SSR performs better than the variants precluding any one or two distillation mechanisms. 2) Variants with any two distillation mechanisms nearly outperform the ones with any one distillation mechanism. Specifically, we find that:

– Variants with CRD outperform the ones without CRD. For instance, our MTKD-SSR with CRD, SRD and SCD outperforms variant C without CRD, variant A with CRD and SRD outperforms variant D without CRD, and variant B with CRD and SCD outperforms variant F without CRD. The experimental ablations of CRD mean that

our proposed CRD can work well for knowledge distillation. The benefit derived from CRD is due to the possible reason that our cross-stage review distillation as a new self-distillation strategy can provide the layer-wise informative knowledge and reduce the capacity gap cross different layers with the proper temperature.

– Variants with SCD outperform the ones without SCD. For instance, our MTKD-SSR with SCD, SRD and CRD outperforms variant A without SCD, variant B with CRD and SCD outperforms variant E without SCD, and variant C with SCD and SRD outperforms variant D without SCD. The SCD benefit for favourable knowledge distillation is due to the possible reason that our stage-wise channel distillation can effectively transfer the abundant and important channel feature knowledge from each teacher's block to guide the student learning.
– Variants with SRD outperform the ones without SRD. For instance, our MTKD-SSR with SRD, SCD and CRD outperforms variant B without SCD, variant A with SRD and CRD outperforms variant E without SRD, and variant C with SRD and SCD outperforms variant F without SRD. The superior performance derived from for SRD is due to the possible reason that the stage-wise response distillation can fully transfer the layer-wise logits as the knowedge with the appropriate temperature from the teacher to the student.

Through the experimental ablations, it is clear that these distillation mechanisms with different types of knowledge can complement with each other to improve the knowledge distillation performance, and the devised cross-stage review distillation is more prominent among them. In essence, the proposed MTKD-SSR is a multiple knowledge transfer with different types of knowledge (i.e., logit outputs and channel features) and distillation strategies (i.e., offline and self distillation). Therefore, we can conclude from the ablation experiments that our proposed MTKD-SSR framework is feasible and effective for knowledge distillation with the designed multiple knowledge transfer.

5.2 Ablation Study on Self-Reflection

As shown in Table 6, the cross-stage review distillation as a new self-distillation strategy can transfer more informative knowledge for student self-learning. During the cross-stage review distillation process, the student finishes different self-reflections at different stages, shown in Fig. 8. Specifically, three variants with different self-reflections in Fig. 8 are as follows: (1) In Fig. 8a stage 4 reviews stages 1, 2 and 3, stage 3 reviews stages 1 and 2, and stage 2 reviews stage 1; (2) In Fig. 8b stage 4 reviews stages 2 and 3, and stage 3 only reviews stages 2; (3) In Fig. 8c stage 4 only reviews stages 3. To well understand how these self-reflections via

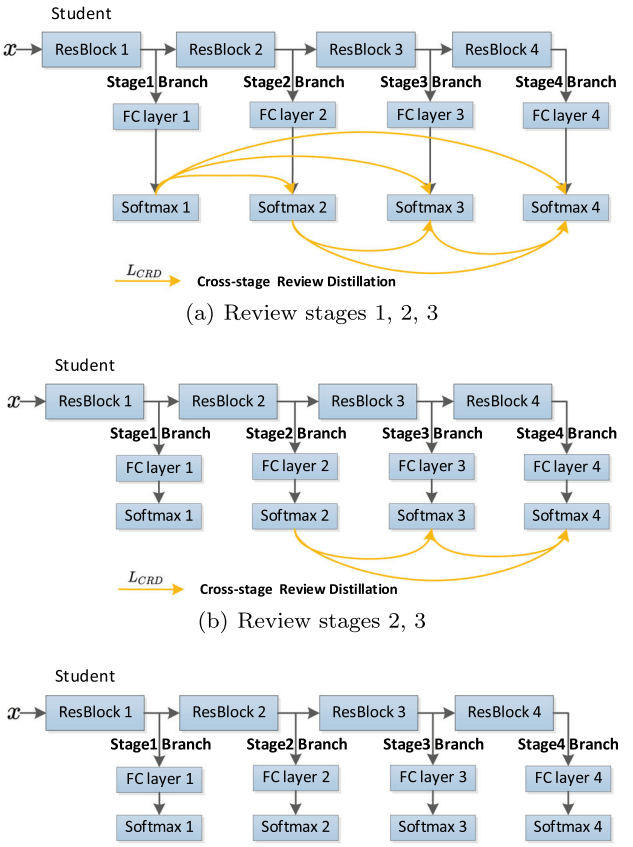


Fig. 8 The diagrams of the different student self-reflections at different stages in the proposed MTKD-SSR

cross-stage review distillation contribute to the performance of our MTKD-SSR, the ablation performance of different self-reflections using two teacher-student architectures (i.e., ResNet34-34 and ResNet34-18) is explored on CIFAR-100 and Market-1501. Note that the variant with different self-reflections in Fig. 8a is used in MTKD-SSR.

Table 7 shows the experimental results of the proposed MTKD-SSR with the variants of different self-reflections. We can see that the student self-reflections via cross-stage self-distillation from the previous stages to the current stage are favourable for improving the knowledge distillation performance. Specifically, the variant in Fig. 8a performs better than the ones in Fig. 8b and c, and the variant in Fig. 8b outperforms the one in Fig. 8c. The possible reason behind the experimental phenomena is that the informative dark knowledge contained in the preceding layers of the student network is well distilled to promote the learning of its subsequent layers within the student network. Besides, this experimental result also implies that the knowledge from each layer of the network benefits self-distillation. Therefore, the proposed cross-stage review distillation mechanism as an important

Table 7 Ablation study on self-reflections in terms of Top-1 accuracy (%) on CIFAR-100 and mAP (%) on Market-1501

Variants	Stage 1	Stage 2	Stage 3	Stage 4	ResNet34-34		ResNet34-18	
					CIFAR-100	Market-1501	CIFAR-100	Market-1501
Fig. 8a	✓	✓	✓	✓	81.45	68.64	80.81	66.73
Fig. 8b	×	✓	✓	✓	80.87	68.38	80.41	66.31
Fig. 8c	×	×	✓	✓	80.49	67.93	79.99	65.88

Table 8 Ablation study on feature knowledge in terms of Top-1 accuracy (%) on CIFAR-100, and mAP (%) on Market-1501

Variants with feature knowledge	ResNet34-34		ResNet34-18	
	CIFAR-100	Market-1501	CIFAR-100	Market-1501
$\mathcal{L}_{KD}$ with Eq. (10)	81.45	68.64	80.81	66.73
$\bar{\mathcal{L}}_{KD}$ with Eq. (13)	80.96	67.58	80.38	66.07

part of our method is very useful to serve as a new and effective self-distillation mechanism.

5.3 Ablation Study on Feature Knowledge

To further verify the superior performance derived from the informative feature knowledge, which is contained in channel features and transferred via our designed stage-wise channel distillation, we compare channel knowledge to the knowledge directly modeled by feature maps (Zhang et al., 2020). The knowledge distillation loss directly using feature maps is formulated as

$$\mathcal{L}_{FD} = \sum_{i=1}^{K} \left\| F_i^s - F_i^t \right\|_2^2, \tag{12}$$

where F_i^t and F_i^s represent the teacher and student's feature maps of the i-th shallow block, respectively. Specifically, using stage-wise feature distillation loss $\mathcal{L}_{FD}$ instead of the proposed stage-wise channel distillation loss $\mathcal{L}_{SCD}$, our full model for knowledge distillation in Eq. (10) is reformulated as

$$\bar{\mathcal{L}}_{KD} = \alpha \mathcal{L}_{SRD} + \beta \mathcal{L}_{CRD} + \gamma \mathcal{L}_{FD} + \mathcal{L}_{CE}\,. \tag{13}$$

The comparative results of the stage-wise distillation using channel maps and feature maps in our proposed model on CIFAR-100 and Market-1501 are shown in Table 8. The teacher-student architectures are ResNet34-34 and ResNet34-18. We see that our proposed method using channel maps clearly outperforms that using feature maps. This is possibly because that the feature knowledge modelled by channel-wise attention tends to focus on those important channel features (Zhou et al., 2006). Therefore, the channel knowledge is more informative in our proposed model and the corresponding stage-wise channel distillation is effective.

5.4 Ablation Study on Self-Distillation

To further verify the effectiveness of the designed CRD, we also show the results by replacing the CRD self-distillation in our MTKD-SSR with vanilla self-distillation, where the knowledge is distilled from the deepest layer to the other preceding layers (Zhang et al., 2019). The comparison between SCD+SRD+CRD (i.e., MTKD-SSR) and SCD+SRD+Self-distillation (Zhang et al., 2019) using the ResNet34-18 teacher-student architecture is shown in Table 9. It is clear that our proposed MTKD-SSR using the CRD self-distillation performs better than that using self-distillation (Zhang et al., 2019). Therefore, this demonstrates that the dark knowledge from the shallow layers can be effectively distilled to the deep layers to improve the self-distillation performance, and our cross-stage review distillation (CRD) with different self-reflections at different stages is very effective for knowledge distillation. Through the above analysis, it can be concluded that the proposed distillation mechanisms and knowledge transfer are prominent and the proposed MTKD-SSR framework is promising for knowledge distillation.

Table 9 Ablation study on self-distillation in terms of Top-1 accuracy (%) on CIFAR-100, and mAP (%) on Market-1501

Datasets	Accuracy	SCD+SRD+CRD (Ours)	SCD+SRD+Self-Distillation (Zhang et al., 2019)
CIFAR-100	Top-1	80.81	80.15
Market-1501	mAP	66.73	65.91

6 Conclusion

Knowledge distillation has been developed for model compression by transferring knowledge in different forms from a large teacher to a small student. Motivated by how a student carries out learning in the real world, our MTKD-SSR framework has moved one step further by integrating the advantages of multi-stage learning and student self-reflection. The former decomposes the learning into multiple stages with different level of difficulties, and the latter further allows a student to reflect on and improve its own learning. Both two learning principles have been naturally unified into the proposed MTKD-SSR framework with three distillation modules, namely stage-wise response distillation, stage-wise channel distillation and cross-stage review distillation. We have conducted extensive experiments on different visual recognition tasks, including image classification, person re-identification, image retrieval, and semantic segmentation, where those experiments together with intensive ablation studies demonstrate that the proposed three modules functioning together can improve the knowledge distillation efficiency and lead to a significantly improved performance. In future, we will explore more efficient and effective design of the student self-reflection for knowledge distillation, and we also would like to generalize it to transformer-based vision/language models.

Acknowledgements This work was partially supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700) and National Natural Science Foundation of China (Grant Nos. 61976107 and 61502208).

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability The used image datasets that support the findings of the current study are Market-1501 (Zheng et al., 2015), CIFAR-100 (Krizhevsky and Hinton., 2009), CUB200-2011 (Wah et al., 2011), ImageNet (Deng et al., 2009), and Pascal VOC (Everingham et al., 2010), and they are publicly available. Our codes generated or used during the study are available from the corresponding author by reasonable request.

References

Ahn, S., Hu, S. X., Damianou, A., Lawrence, N. D., & Dai, Z. (2019). Variational information distillation for knowledge transfer. In *CVPR* (pp. 9163–9171).

Chen, S., Hong, Z., Xie, G. S., Yang, W., Peng, Q., Wang, K., Zhao, J., & You, X. (2022). MSDN: Mutually semantic distillation network for zero-shot learning. In: *CVPR* (pp. 7612–7621).

Chen, W., Li, S., Huang, C., Yu, Y., Jiang, Y., & Dong, J. (2022). Mutual Distillation Learning Network for Trajectory-User Linking. In: *IJCAI*.

Chen, P., Liu, S., Zhao, H., & Jia, J.(2021). Distilling knowledge via knowledge review. In: *CVPR* (pp. 5008–5017).

Chen, D., Mei, J. P., Zhang, H., Wang, C., Feng, Y., & Chen, C. (2022). Knowledge distillation with the reused teacher classifier. In: *CVPR* (pp. 11933-11942).

Chen, J., Chen, Y., Li, W., Ning, G., Tong, M., & Hilton, A. (2021). Channel and spatial attention based deep object co-segmentation. *Knowledge-Based Systems, 211*, 106550.

Chennupati, S., Kamani, M. M., Cheng, Z., & Chen, L. (2021). Adaptive distillation: Aggregating knowledge from multiple paths for efficient distillation. arXiv preprint arXiv:2110.09674.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hier-archical image database. In: *CVPR* (pp. 248–255).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Fan, S., Cheng, X., Wang, X., Yang, C., Deng, P., Liu, M., Deng J., & Liu, M. (2022). Channel Self-Supervision for Online Knowledge Distillation. arXiv preprint arXiv: 2203.11660.

Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., & Liu, Z. (2021). Compressing visual-linguistic model via knowledge distillation. In: *ICCV* (pp. 1428–1438).

Ge, S., Luo, Z., Zhang, C., Hua, Y., & Tao, D. (2019). Distilling channels for efficient deep tracking. *IEEE Transactions on Image Processing, 29*, 2610–2621.

Gou, J., Sun, L., Yu, B., Du, L., Ramamohanarao, K., & Tao, D. (2022). Collaborative knowledge distillation via multiknowledge transfer. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2022.3212733

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision, 129*(6), 1789–1819.

Guo, S., Alvarez, J. M., & Salzmann, M. (2021). Distilling image classifiers in object detectors. In *NeurIPS* (vol. 34, pp. 1036–1047).

Hagström, L., & Johansson, R. (2021). Knowledge distillation for swedish ner models: A search for performance and efficiency. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 124–134).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019). A comprehensive overhaul of feature distillation. In *ICCV* (pp. 1921-1930).

He, Z., Zhang, L., Gao, X., & Zhang, D. (2022). Multi-adversarial faster-RCNN with paradigm teacher for unrestricted object detection. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-022-01728-z

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

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *CVPR* (pp. 7132–7141).

Huang, Y., Wu, J., Xu, X., & Ding, S. (2022). Evaluation-oriented knowledge distillation for deep face recognition. In *CVPR* (pp. 18740-18749).

Huang, Z., Yang, S., Zhou, M., Li, Z., Gong, Z., & Chen, Y. (2022). Feature map distillation of thin nets for low-resolution object recognition. *IEEE Transactions on Image Processing, 31*, 1364–1379.

Jafari, A., Rezagholizadeh, M., Sharma, P., & Ghodsi, A. (2021). Annealing knowledge distillation. arXiv preprint arXiv: 2104.07163.

Ji, M., Shin, S., Hwang, S., Park, G., & Moon, I. C. (2021). Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In *CVPR* (pp. 10664–10673).

Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., & Patras, I. (2022). DnS: Distill-and-select for efficient and accurate video indexing and retrieval. *International Journal of Computer Vision, 130*(10), 2385–2407.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical Report*.

Li, B., Chen, B., Wang, Y., Dai, T., Hu, M., Jiang, Y., & Xia, S. (2021). Knowledge distillation via channel correlation structure. In: *International conference on knowledge science, engineering and management* (pp. 357–368).

Li, J., Liu, X., Zhang, S., Yang, M., Xu, R., & Qin, F. (2021). Accelerating neural architecture search for natural language processing with knowledge distillation and earth mover's distance. In *ACM SIGIR* (pp. 2091–2095).

Li, C., Peng, J., Yuan, L., Wang, G., Liang, X., Lin, L., & Chang, X. (2020). Block-wisely supervised neural architecture search with knowledge distillation. In *CVPR* (pp. 1989–1998).

Li, Z., Ye, J., Song, M., Huang, Y., & Pan, Z. (2018). Online knowledge distillation for efficient pose estimation. In *ICCV* (pp. 11740–11750).

Liu, L., Huang, Q., Lin, S., Xie, H., Wang, B., Chang, X., & Liang, X. (2021). Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *ICCV* (pp. 8271–8280).

Liu, W., Zhou, P., Wang, Z., Zhao, Z., Deng, H., & Ju, Q. (2020). Fastbert: A self-distilling bert with adaptive inference time. In *ACL* (pp. 6035–6044).

Liu, Y., Shu, C., Wang, J., & Shen, C. (2020). Structured knowledge distillation for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.3001940

Lou, A., & Loew, M.(2021). Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation. In *ICIP* (pp. 1894–1898).

Ma, Z., Luo, G., Gao, J., Li, L., Chen, Y., Wang, S., Zhang, C., & Hu, W. (2022). Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR* (pp. 14074–14083).

Mirzadeh, S. I., Farajtabar, M., Li, A., & Ghasemzadeh, H. (2020). Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. In *AAAI* (pp. 5191–5198).

Mobahi, H., Farajtabar, M., & Bartlett, P. (2020). Self-distillation amplifies regularization in hilbert space. In *NeurIPS* (pp. 3351–3361).

Muhammad, A., Zhou, F., Xie, C., Li, J., Bae, S. H., & Li, Z. (2021). MixACM: Mixup-based robustness transfer via distillation of activated channel maps. In *NeurIPS* (pp. 4555–4569).

Park, D. Y., Cha, M. H., Kim, D., & Han, B. (2021). Learning student-friendly teacher networks for knowledge distillation. arXiv preprint arXiv: 2102.07650.

Peng, Y., Qi, J., Ye, Z., & Zhuo, Y. (2021). Hierarchical visual-textual knowledge distillation for life-long correlation learning. *International Journal of Computer Vision, 129*(4), 921–941.

Phan, M. H., Phung, S. L., Tran-Thanh, L., & Bouzerdoum, A. (2022). Class similarity weighted knowledge distillation for continual semantic segmentation. In *CVPR* (pp. 16866–16875).

Phuong, M., & Lampert C. H. (2019). Distillation-based training for multi-exit architectures. In *ICCV* (pp. 1355–1364).

Qu, Y., Deng, W., & Hu, J.: H-at. (2020). Hybrid attention transfer for knowledge distillation. In *PRCV* (pp. 249–260).

Shen, Y., Xu, L., Yang, Y., Li, Y., & Guo, Y. (2022). Self-Distillation from the last mini-batch for consistency regularization. In *CVPR* (pp. 11943–11952).

Shu, C., Liu, Y., Gao, J., Yan, Z., & Shen, C. (2021). Channel-wise knowledge distillation for dense prediction. In *ICCV* (pp. 5311–5320).

Sun, D., Yao, A., Zhou, A., & Zhao, H. (2019). Deeply-supervised knowledge synergy. In *CVPR* (pp. 6997–7006).

Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *ICCV* (pp. 1365–1374).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*(11), 2579–2605.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wang, Y., Zhou, W., Jiang, T., Bai, X., & Xu, Y. (2020). Intra-class feature variation distillation for semantic segmentation. In *ICCV* (pp. 346–362).

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3349–3364.

Wang, L., & Yoon, K. J. (2022). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(6), 3048–3068.

Wu, G., & Gong, S. (2021). Peer collaborative learning for online knowledge distillation. In *AAAI* (pp. 10302–10310).

Wu, X., He, R., Hu, Y., & Sun, Z. (2020). Learning an evolutionary embedding via massive knowledge distillation. *International Journal of Computer Vision, 128*(8), 2089–2106.

Xu, J., Huang, S., Zhou, F., Huangfu, L., Zeng, D., & Liu, B. (2022). Boosting multi-label image classification with complementary parallel self-distillation. In *IJCAI*.

Yan, H., Zhang, J., Niu, G., Feng, J., Tan, V., & Sugiyama, M. (2021). Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *ICML* (pp. 11693–11703).

Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., & Yuan, C. (2022). Focal and global knowledge distillation for detectors. In *CVPR* (pp. 4643–4652).

Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., & Zhang, Q. (2022). Cross-image relational knowledge distillation for semantic segmentation. In *CVPR* (pp. 12319–12328).

Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR* (pp. 4133–4141).

You, C., Chen, N., & Zou, Y.(2021). Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP* (pp. 7793–7797).

Yuan, F., Shou, L., Pei, J., Lin, W., Gong, M., Fu, Y., & Jiang, D. (2021). Reinforced multi-teacher selection for knowledge distillation. In *AAAI*.

Yuan, L., Tay, F. E., Li, G., Wang, T., & Feng, J. (2020). Revisiting knowledge distillation via label smoothing regularization. In *CVPR* (pp. 3903–3911).

Yu, B., & Tao, D. (2021). Heatmap regression via randomized rounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8276–8289.

Zagoruyko, S., & Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR* (pp. 1–13).

Zhang, S., Liu, H., Hopcroft, J. E., & He, K. (2022). Class-aware Information for Logit-based Knowledge Distillation. arXiv preprint arXiv:2211.14773.

Zhang, L., Shi, Y., Shi, Z., Ma, K., & Bao, C. (2020). Task-oriented feature distillation. In *NeurIPS* (pp. 14759–14771).

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV* (pp. 3713–3722).

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In *CVPR* (pp. 4320–4328).

Zhang, L., Bao, C., & Ma, K. (2022). Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(8), 4388–4403.

Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. In *CVPR* (pp. 11953–11962).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *CVPR* (pp. 2881–2890).

Zhao, T., Han, J., Yang, L., Wang, B., & Zhang, D. (2021). SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning. *International Journal of Computer Vision, 129*(8), 2474–2498.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person reidentification: A benchmark. In *ICCV* (pp. 1116–1124).

Zhou, Z., Zhuge, C., Guan, X., & Liu, W. (2006). Channel distillation: Channel-wise attention for knowledge distillation. arXiv preprint arXiv: 2006.01683

Zhu, X., & Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. In *NeurIPS* (pp. 7517–7527).

Zhu, Y., & Wang, Y.(2021). Student customized knowledge distillation: Bridging the gap between student and teacher. In *ICCV* (pp. 5057–5066).