Visual Intelligence

RESEARCH Open Access

Semantic-aware knowledge distillation with parameter-free feature uniformization

Guangyu Guo[1], Longfei Han[2], Le Wang[3], Dingwen Zhang[1,4*] and Junwei Han[5]

Abstract

Knowledge distillation aims to distill knowledge from teacher networks to train student networks. Distilling intermediate features has attracted much attention in recent years as it can be flexibly applied in various fields such as image classification, object detection and semantic segmentation. A critical obstacle of feature-based knowledge distillation is the dimension gap between the intermediate features of teacher and student, and plenty of methods have been proposed to resolve this problem. However, these works usually implement feature uniformization in an unsupervised way, lacking guidance to help the student network learn meaningful mapping functions in the uniformization process. Moreover, the dimension uniformization process of the student and teacher network is usually not equivalent as the mapping functions are different. To this end, some factors of the feature are discarded during parametric feature alignment, or some factors are blended in some non-parametric operations. In this paper, we propose a novel semantic-aware knowledge distillation scheme to solve these problems. We build a standalone feature-based classification branch to extract semantic-aware knowledge for better guiding the learning process of the student network. In addition, to avoid the inequivalence of feature uniformization between teacher and student, we design a novel parameter-free self-attention operation that can convert features of different dimensions into vectors of the same length. Experimental results show that the proposed knowledge distillation scheme outperforms existing feature-based distillation methods on the widely used CIFAR-100 and CINIC-10 datasets.

Keywords: Knowledge distillation, Feature-based, Semantic-aware, Parameter-free, Self-attention, Feature uniformization

1 Introduction

Convolutional neural networks have achieved great success in different fields such as computer vision [1–6] and speech recognition [7, 8]. However, modern deep networks have improved in accuracy with high computational costs and heavy models. A variety of approaches have been proposed to relieve this problem, such as model pruning [9, 10], knowledge distillation [11–16], and model quantization [17]. Among those methods, knowledge distillation (KD) has been widely used in such diverse areas as classification [18, 19], object detection [20–23], semantic segmentation [24, 25], and transfer learning [26].

Knowledge distillation aims to guide the training process of a (compact) student network by extracting information from a (complex) teacher network. Previous knowledge distillation methods can be divided into two categories based on the information they employed. The first category exploits the logits predicted by the teacher to guide the student. Since the output spaces of the teacher and student have the same dimension (number of categories), this category of knowledge distillation is easy to implement by using Kullback-Leibler divergence [11, 12]. The other category of methods aims to extract knowledge from the intermediate features of the teacher. Since features of teacher and student have different channel numbers and spatial sizes in most cases, an extra mapping module is needed

*Correspondence: zhangdingwen2006yyy@gmail.com
[1]Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an, China
[4]Xijing Hospital, The Fourth Military Medical University, Xi'an, China
Full list of author information is available at the end of the article

in feature distillation to unify the dimensions of those features. To be specific, previous works have proposed fruitful ways to implement feature uniformization, such as introducing extra convolution layers [27–30], downsampling [31, 32], selecting and matching [33], and mini-batch level feature multiplication [34].

There are two issues for these previous feature-based distillation approaches. First, since previous feature uniformization methods mainly implement the feature uniformization processes in an unsupervised or class-agnostic manner, the distilled representation does not contain all key semantic information from the teacher. Moreover, as the capacity of the teacher network is larger than that of the student network, some retained information in the distilled knowledge may be redundant for the student network. Second, the feature uniformization process of the teacher and student is not equivalent. As depicted in Fig. 1(a)–1(c), the inequivalence is usually produced during the data selection or feature transformations. Early works utilize different mapping functions for teacher and student [28, 30], or only transform features for only one network in teacher and student [27–29, 31]. Some works apply the same mapping function to teacher and student, but some factors of the original features are discarded in the process of data selection [33] or Taylor series expansion [32] (Fig. 1(b)). Similarity-perceiving [34] multiplies the feature of a mini-batch to its transposed feature to make the dimension uniform, which leads to a mixture of data from different samples (Fig. 1(c)). The existence of those inequivalences will make the learning process unaligned: knowledge transferred from the teacher and what is received by the student is not in the same data space.

In this paper, we propose a novel semantic-aware distillation scheme with a parameter-free uniformization operation (SPU) to resolve the aforementioned two issues. The SPU can transfer semantic-aware information from teacher to student with equivalent transformations (Fig. 1(d)). As shown in Fig. 2(a), the SPU aggregates features of different group layers together and then feeds this aggregated feature to a classification module. This feature-based classification module is supervised by the ground truth labels to extract semantic-aware information from intermediate features. The feature aggregation operation requires all the features to be transformed into the same dimension, so we introduce the self-attention mechanism to convert features of different dimensions into vectors of the same length. As demonstrated in Fig. 2(b), the proposed feature uniformization operation mixes features from different samples. Moreover, since this novel feature uniformization module is parameter-free, and the weights of the classification module are frozen when training the student, SPU avoids introducing inequity between the knowledge distilling of the teacher and the learning of the student. To demonstrate the effectiveness of the proposed methods, we conduct experiments on two widely used datasets CIFAR-100 and CINIC-10. Experimental results demonstrate the superiority of the proposed method.

The main contributions of this paper are summarized as follows:

1) We propose a novel knowledge distillation scheme SPU, which can extract semantic-aware knowledge from the intermediate features to guide the training process of the student network.
2) We propose a parameter-free self-attention operation that can achieve equivalent feature uniformization between teacher and student. This operation can convert features of different dimensions into vectors of the same length without information loss caused

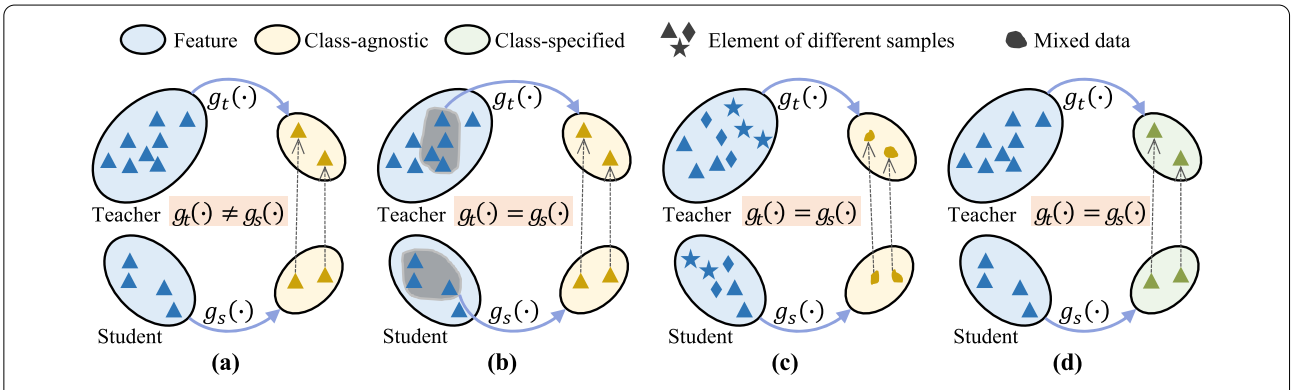

Figure 1 Illustration of different ways to unify the dimensions of teacher-student feature pairs. Ellipses of different colors represent different types of data space. Different shapes ▲, ♦, ★ represent elements of different samples, and the irregular shape represents a mixture of data from different samples. (**a**) The features of teacher and student are mapped by different transformations. (**b**) The features of teacher and student are first selected and then mapped by the same transformations. (**c**) Teacher and student features are mapped by the same transformation, but information from different samples is mixed. (**d**) Our SPU uses an equivalent transformation to the teacher and student, and the distillation occurs between the same sample. Moreover, SPU distills semantic-aware information from the teacher to guide the student

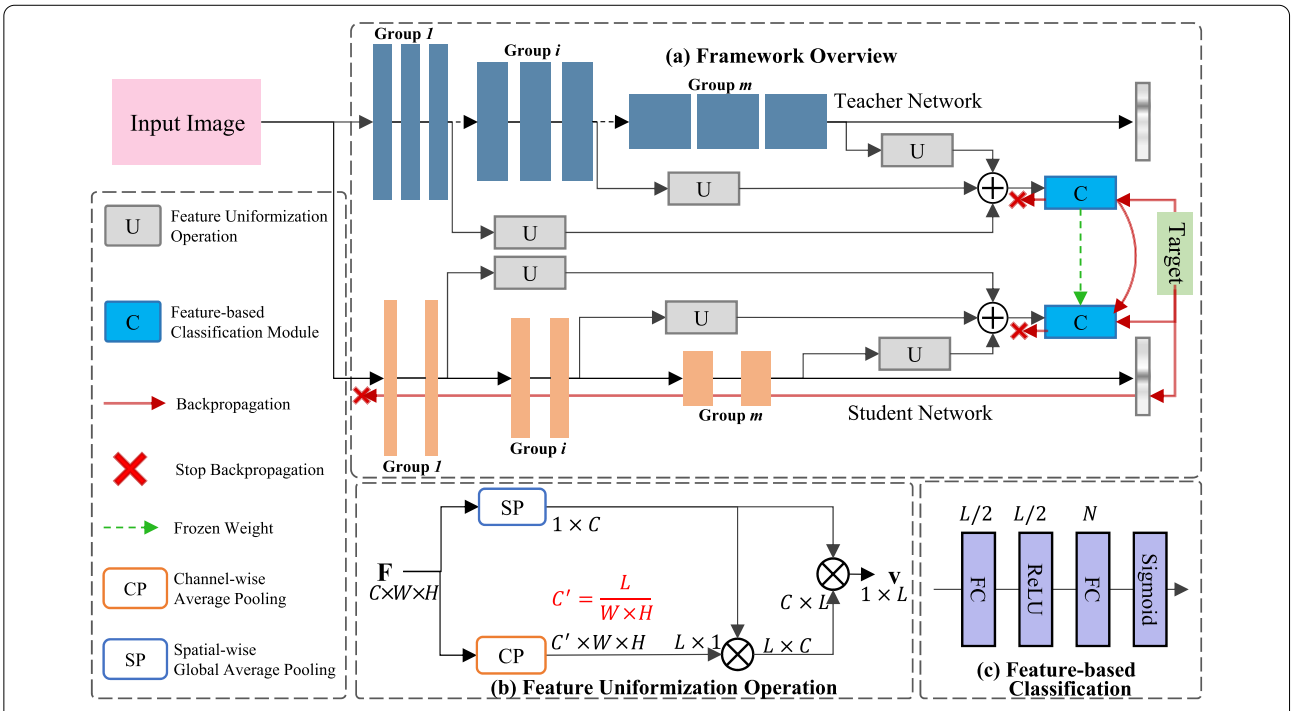

Figure 2 Overview of the framework. (**a**) Illustration of the framework for the proposed feature-based knowledge distillation scheme SPU. For every single network (teacher or student), we first convert its intermediate features into vectors of fixed dimension by a parameter-free feature transformation operation. Then, all the vectors are fed to the classification module. The feature-based distillation branch of the teacher is supervised by the ground truth and its loss does not backpropagate to the backbone of the teacher. Its output in the evaluation process is used to guide the feature-based distillation branch of the student. ⊕ denotes element-wise sum. (**b**) Detailed implementation of the feature uniformization operation. ⊗ denotes the matrix inner-product operation. (**c**) Architecture of the feature-based classification module

by the parametric feature alignment process or blending data.

3) Comprehensive experiments were conducted on the CIFAR-100 and CINIC-10 datasets to demonstrate the effectiveness of the proposed SPU scheme.

2 Related works

2.1 Knowledge distillation

Buciluǎ et al. [35] first introduced the notion of training a smaller model (i.e., student) by learning the cumbersome models (i.e., teacher). Recently, due to the rapid development of deep learning, similar approaches have been proposed in the field of convolutional neural networks, and are usually named knowledge distillation [11, 12]. The early knowledge distillation methods used the predicted logits of the teacher to guide student training. An essential way is to regard the predicted logits of the teacher as the soft target of the student [11, 12]. Park et al. [36] transferred mutual relations between different samples rather than the output of individual samples. Yu et al. [37] successfully applied knowledge distillation in metric learning.

In addition to the predicted logits, plenty of works have been proposed to guide the student by the intermediate representations of the teacher [29, 31, 38–40]. Compared to distilling knowledge from logits, distilling intermediate features is much more helpful to guide the student in more complex tasks, such as object detection [20, 21] and semantic segmentation [24]. The fundamental problem of transforming intermediate representations is that the dimension of the teacher feature is different from that of the student feature. Some previous works [26, 27, 29, 40] overcame these obstacles by building an adaptive module between hidden layers of teacher and student, which makes the student unable to directly learn the intermediate representations. Apart from that, Kim et al. [28] proposed an unsupervised reconstruction method to align the dimension of factors used for distillation. Similarity preserving (SP) [34] unifies the dimensions at the mini-batch level by matrix operation. DFA [30] introduces network architecture search [41] to learn the adaptive module between hidden layers of teacher and student. MFD [33] selects elements from the features of teacher and student to achieve knowledge distillation by matching the selected student feature to that of the teacher. In addition to the feature, the loss function is designed, including activation transfer loss with boundaries formed by hidden neurons [42], Jacobian [32], and instance relationship graph [43]. Moreover, different works aim to minimize the similarity be-

tween the features of the teacher and student. Passalis et al. [31] downsampled the teacher features by t-SNE [44] and transferred probabilistic features to the student. However, the distilled information of these aforementioned methods may not be accurate enough as their feature uniformization modules are usually unsupervised, and inequivalence will also be introduced between the knowledge distilling of the teacher and the learning of the student.

In contrast to previous works, we propose a novel feature-based knowledge distillation scheme to transfer semantic-aware information from the teacher to the student. At the same time, a novel method is proposed to achieve feature uniformization without introducing inequivalence between the teacher and student. TOFD [45] is also a feature-based method to distill semantic-aware information, but it belongs to the class where the features of teacher and student are mapped by different transformations as TOFD uses different convolutional layers for features from teacher and student, respectively. Different from TOFD, our proposed SPU transforms the features of the teacher and student into the same dimension by a parameter-free operation. In addition, for the fully connected layer in the feature distillation branch of the student that is used to learn semantic information, its weight is directly transferred from the teacher and is frozen through the training process, which ensures that the knowledge extracted from the teacher is easier for the student to learn. Moreover, as illustrated in Fig. 3, our proposed SPU has explored different ways of feature aggregation and proves that it is better to aggregate features of different layers before learning semantic information.

2.2 Self-attention

Attention mechanisms have been widely explored and used in recent years [46–48], and have been rapidly adopted in various tasks such as natural language processing [46], visual dialog [49, 50], and semantic segmentation [51, 52]. Different from the traditional attention mechanism that occurs between the source data and target data [46, 53], self-attention mechanism only uses the input information itself [47, 48].

Vaswani et al. [48] proposed the Transformer that only consists of self-attention modules. Transformer has demonstrated its potential in the field of language understanding [54–56], and many Transformer-based methods have been proposed to resolve computer vision problems such as image classification [57] and object detection [58]. Currently, simplifying the Transformer and reducing its computational cost has received much attention [59, 60].

In addition to the Transformer and its following works, the self-attention mechanism has also been explored to build the inner-relationship of an input feature or special data processing [61–64]. Non-Local [61] proposed a self-attention module to model the cross-frame relationship for a video clip. Wu et al. [65] further extended the non-local mechanism to untrimmed video recognition. Zhang et al. [63] introduced the self-attention mechanism in the anchor of the detector to achieve global respective fields. Zhu et al. [66] proposed self-attention methods to address the long-tailed visual recognition problem. In this paper, we introduce the self-attention mechanism into feature-based knowledge distillation to avoid the inequivalence problem in the feature uniformization process. To be specific, we propose a novel parameter-free self-attention operation that can convert input features of different dimensions into vectors of the same dimension.

3 Methods

In this paper, we propose a *s*emantic-aware feature distillation scheme with *p*arameter-free *u*niformization (SPU). First, classical feature distillation methods are briefly reviewed. Then, a detailed description of the proposed semantic-aware distillation scheme and parameter-free feature uniformization operation is given. Finally, some details of the implementation are provided.

3.1 Overview of feature distillation

We first provide a review of previous distillation approaches. Given an image $\mathbf{I}$, both the teacher and student network can generate a series of intermediate features:

$$\mathbf{F}^{\mathrm{t}} = f^{\mathrm{t}}\left(\mathbf{W}^{\mathrm{t}}, \mathbf{I}\right), \qquad \mathbf{F}^{\mathrm{s}} = f^{\mathrm{s}}\left(\mathbf{W}^{\mathrm{s}}, \mathbf{I}\right), \tag{1}$$

where $f^{\mathrm{t}}(\cdot)$, $f^{\mathrm{s}}(\cdot)$ denote the inference operation of the teacher network and student network, respectively. $\mathbf{W}^{\mathrm{t}}$ and $\mathbf{W}^{\mathrm{s}}$ are their weights. $\mathbf{F}^{\mathrm{t}} = \{\mathbf{F}^{\mathrm{t}}_1, \mathbf{F}^{\mathrm{t}}_2, \ldots, \mathbf{F}^{\mathrm{t}}_m\}$, $\mathbf{F}^{\mathrm{s}} = \{\mathbf{F}^{\mathrm{s}}_1, \mathbf{F}^{\mathrm{s}}_2, \ldots, \mathbf{F}^{\mathrm{s}}_m\}$, and m denotes the number of groups.

Feature-based knowledge distillation methods can be generalized and defined by the following knowledge loss function:

$$\mathcal{L}_{\mathrm{fkd}} = \sum_{i=1}^{m} \left\| g^{\mathrm{t}}\left(\mathbf{F}^{\mathrm{t}}_i\right), g^{\mathrm{s}}\left(\mathbf{F}^{\mathrm{s}}_i\right) \right\|_p, \tag{2}$$

where knowledge loss used for each feature pair is usually $\mathcal{L}_1$ or $\mathcal{L}_2$ loss (i.e., $p \in [1, 2]$), or some variants of them such as partial L_2 loss [29, 33]. $g^{\mathrm{t}}(\cdot)$ and $g^{\mathrm{s}}(\cdot)$ denote the mapping functions of teacher and student, respectively. In general, the goal of $g^{\mathrm{t}}(\cdot)$ and $g^{\mathrm{s}}(\cdot)$ is to unify the dimensions of features extracted from teacher and student network. In Fig. 1(a)–1(c), we illustrate how previous works resolve the dimension gap problem and bring up two problems that need to be solved. First, the feature uniformization modules of previous feature distillation methods are always unsupervised or trained in a class-agnostic manner, which would lead to inaccurate distilled knowledge. Second, the way of distilling knowledge from the teacher is not equivalent to the learning process of the student, which is caused by different mapping functions, or discarding and mixing data.

Algorithm 1: The training process of the teacher and student networks in SPU.

Input: The training images and the groud truth $\{\mathcal{I}, \mathcal{Y}\}$;

Output: The learned student network $\mathbf{W}^{\mathrm{s}}$;

Optimize $\mathbf{W}^{\mathrm{t}}$ by $\mathcal{L}_{\mathrm{b}}^{\mathrm{t}}$ and $\mathcal{Y}$;
Freeze $\mathbf{W}^{\mathrm{t}}$;
Optimize $\mathbf{W}^{\mathrm{fcls}}$ by $\mathcal{L}_{\mathrm{fcls}}$ and $\mathcal{Y}$;
Freeze $\mathbf{W}^{\mathrm{fcls}}$;
for *i in 1 ... N* **do**
    Feed image I into the teacher, get $\mathbf{x}_{\mathrm{f}}^{\mathrm{t}}$ and $\mathbf{p}^{\mathrm{t}}$;
    Optimize $\mathbf{W}^{\mathrm{s}}$ student by $\{y, \mathbf{p}^{\mathrm{t}}\}$ and $\mathcal{L}^{\mathrm{s}}$.
return The learned student network $\mathbf{W}^{\mathrm{s}}$.

3.2 Semantic-aware feature distillation

In previous feature distillation works, the knowledge is always distilled from the teacher network in a class-agnostic or unsupervised way. Even though most of these methods can enhance the similarity between teacher and student features, but the distilled knowledge from the teacher may still miss some key semantic information and contain some redundant information to student. To resolve this problem, we propose a novel feature distillation scheme to extract semantic-aware information from the teacher network.

Suppose the ground truth as $\mathbf{y} = [y_1, y_2, \ldots, y_N]$, where N is the number of categories. The learning process of the proposed scheme is shown in Algorithm 1. We first train the backbone of the teacher network whose weights can be represented by $\mathbf{W}^{\mathrm{t}}$. The loss for the backbone of the teacher is the cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}}^{\mathrm{t}}\left(\mathbf{x}_{\mathrm{b}}^{\mathrm{t}}, \mathbf{y}\right) = -\sum_{y \in \mathbf{y}, x \in \mathbf{x}_{\mathrm{b}}^{\mathrm{t}}} y \log x, \tag{3}$$

where $\mathbf{x}_{\mathrm{b}}^{\mathrm{t}}$ denotes the logits predicted by the backbone of the teacher. As shown in Fig. 2(a), we first use the proposed parameter-free feature uniformization operation for all intermediate features, and obtain a series of vectors from the teacher and student, represented by $\{\mathbf{v}_1^{\mathrm{t}}, \mathbf{v}_2^{\mathrm{t}}, \ldots, \mathbf{v}_m^{\mathrm{t}}\}$ and $\{\mathbf{v}_1^{\mathrm{s}}, \mathbf{v}_2^{\mathrm{s}}, \ldots, \mathbf{v}_m^{\mathrm{s}}\}$, respectively. These vectors are of the same length L. For each network, we add these vectors together:

$$\mathbf{h}^{\mathrm{t}} = \sum_{i=1}^{m} \mathbf{v}_i^{\mathrm{t}}, \qquad \mathbf{h}^{\mathrm{s}} = \sum_{i=1}^{m} \mathbf{v}_i^{\mathrm{s}}, \tag{4}$$

$\mathbf{h}^{\mathrm{t}}$ and $\mathbf{h}^{\mathrm{s}}$ are fed to the feature-based classification module to obtain the classification score:

$$\mathbf{x}_{\mathrm{f}}^{\mathrm{t}} = f^{\mathrm{cls}}\left(\mathbf{W}^{\mathrm{fcls}}, \mathbf{h}^{\mathrm{t}}\right) \qquad \mathbf{x}_{\mathrm{f}}^{\mathrm{s}} = f^{\mathrm{cls}}\left(\mathbf{W}^{\mathrm{fcls}}, \mathbf{h}^{\mathrm{s}}\right), \tag{5}$$

where f^{cls} and $\mathbf{W}^{\mathrm{fcls}}$ denote the inference operation and weights of the feature-based classification module, respectively. Then, we freeze $\mathbf{W}^{\mathrm{t}}$ and train the feature-based classification module:

$$\mathcal{L}_{\mathrm{fcls}}\left(\mathbf{x}_{\mathrm{f}}^{\mathrm{t}}, \mathbf{y}\right) = -\sum_{y \in \mathbf{y}, x \in \mathbf{x}_{\mathrm{f}}^{\mathrm{t}}} y \log x, \tag{6}$$

since the feature-based classification branch of the teacher is trained by the ground truth, the predicted classification probability will be semantic-aware and more appropriate to be tranferred to the student.

After training the feature-based classification module, we freeze its weights and then train the student network. The backbone of the student is supervised by the ground truth and trained by the cross-entropy loss:

$$\mathcal{L}_{\mathrm{cls}}^{\mathrm{s}}\left(\mathbf{x}_{\mathrm{b}}^{\mathrm{s}}, \mathbf{y}\right) = -\sum_{y \in \mathbf{y}, x \in \mathbf{x}_{\mathrm{b}}^{\mathrm{s}}} y \log x, \tag{7}$$

where $\mathbf{x}_{\mathrm{b}}^{\mathrm{s}}$ is predicted by the backbone of the student. To distill the knowledge from the teacher, we follow the traditional logits based knowledge distillation [12] and train the feature-based classification of the student by using the softened predicted class scores of the teacher network:

$$\mathcal{L}_{\mathrm{feat}}^{\mathrm{s}} = (1-\alpha)\mathcal{L}_{\mathrm{fcls}}^{\mathrm{s}} + \alpha \mathcal{L}_{\mathrm{kd}}, \tag{8}$$

where

$$\mathcal{L}_{\mathrm{fcls}}^{\mathrm{s}}\left(\mathbf{x}_{\mathrm{f}}^{\mathrm{s}}, \mathbf{y}\right) = -\sum_{y \in \mathbf{y}, x \in \mathbf{x}_{\mathrm{f}}^{\mathrm{s}}} y \log x, \tag{9}$$

$$\mathcal{L}_{\mathrm{kd}} = -\tau^2 \sum_{i=1} \mathbf{p}_i^{\mathrm{t}} \mathbf{p}_i^{\mathrm{s}}, \tag{10}$$

where $\mathbf{p}^{\mathrm{s}} = \log(\mathrm{softmax}(\mathbf{x}_{\mathrm{f}}^{\mathrm{s}}))$ and $\mathbf{p}^{\mathrm{t}} = \mathrm{softmax}(\frac{\mathbf{x}_{\mathrm{f}}^{\mathrm{t}}}{\tau})$, and $\mathbf{x}_{\mathrm{f}}^{\mathrm{s}}$ is predicted by the feature based classification branch of the student. τ is a temperature parameter, and α is a balance hyper-parameter for balancing the classification loss $\mathcal{L}_{\mathrm{fcls}}$ and the Kullback Leibler divergence loss $\mathcal{L}_{\mathrm{kd}}$. The weights of the feature-based classification module is frozen to ensure the equivalence between $g^{\mathrm{t}}(\cdot)$ and $g^{\mathrm{s}}(\cdot)$.

The total loss for the student is:

$$\mathcal{L}^{\mathrm{s}} = \mathcal{L}_{\mathrm{cls}}^{\mathrm{s}} + \mathcal{L}_{\mathrm{feat}}^{\mathrm{s}}. \tag{11}$$

3.3 Parameter-free feature uniformization

As illustrated in Fig. 2(a), to train the feature-based classification module in the proposed SPU scheme, we have to convert all the intermediate features of the network into the same dimension. Moreover, we also want to keep the feature uniformization process of teacher and student equivalent, i.e., $g^{\mathrm{t}}(\cdot) = g^{\mathrm{s}}(\cdot)$. To satisfy the above two

conditions, we introduce the self-attention mechanism to achieve parameter-free feature uniformization operation.

Given an arbitrary feature $\mathbf{F}$ from $[\mathbf{F}_i^s, \mathbf{F}_i^t]$, suppose it is of spatial size $W \times H$ and channel number C, the goal of the proposed uniformization operation is to convert $\mathbf{F}$ into a vector $\mathbf{v} \in \mathbb{R}^{1\times L}$, where L is a pre-defined hyper-parameter and remains constant for all features.

Figure 2(b) provides the framework of the proposed feature uniformization operation. We first apply a channel pooling to $\mathbf{F}$ and obtain a feature map of dimension $C' \times W \times H$, where $C' = \frac{L}{W\times H}$. Channel pooling is a average pooling with kernel size C' and stride C'. Then, we reshape this feature map to a vector $\mathbf{v}_c \in \mathbb{R}^{L\times 1}$. In addition, we perform a global averge pooling opration to the spatial dimension of $\mathbf{F}$, and obtain a vector $\mathbf{v}_s \in \mathbb{R}^{1\times C}$. Then, we can calculate attention $\mathbf{A} \in \mathbb{R}^{L\times C}$ by Equation (12):

$$\mathbf{A} = \mathbf{v}_c \otimes \mathbf{v}_s, \tag{12}$$

where $\otimes$ denotes matrix inner-product operation. After the attention map is obtained, we can generate a fixed dimension vector $\mathbf{v} \in \mathbb{R}^{L\times C}$ by Equation (13):

$$\mathbf{v} = \mathbf{v}_s \otimes \mathbf{A}^T. \tag{13}$$

We observe that the proposed self-attention operation can convert features of indefinite dimensions into vectors of fixed dimensions without information discarding caused by the parametric feature alignment process or disturbing the original data. Moreover, the proposed feature uniformization operation is also parameter-free, so the equivalence of $g^t(\cdot)$ and $g^s(\cdot)$ will not be broken due to the feature uniformization.

Many kinds of self-attention modules have been proposed and embedded inside convolutional neural networks. The proposed feature uniformization operation is similar to some previous works such as non-local mechanisms [61], or some space-time memory methods for video recognition [65, 67], but it is designed for a different purpose as it aims to unify the dimensions of features without introducing any parameters. Moreover, even though the work [38] also uses attention from the teacher to guide the student, it requires the features to have the same spatial size, while our method does not impose any restrictions on the dimension of input features.

3.4 Implementation details

As demonstrated in Fig. 2(c), the feature-based classification module comprises two fully connected layers, where the ReLU activation function follows the first layer, and the sigmoid activation function follows the second layer. Following [34, 38], we set the temperature $\tau = 4$, and $\alpha = 0.9$. The length of the converted vectors l is set to 4096 in the CIFAR-100 experiments and 2048 in the CINIC-10 experiments. Random horizontal flip and random crop are used as the data augmentation. Experiments are conducted on a GTX 1080ti GPU and Intel Xeon E5-2698 v4 CPU, and implemented by Python and PyTorch [68].

4 Experiments

4.1 CIFAR-100

CIFAR-100 [69] is a widely used visual recognition dataset that consists of 60,000 32×32 color images in 100 classes. Each class contains 500 training images and 100 testing images. In the training process, we use SGD as the optimizer with a weight decay of 0.0005 and momentum of 0.9. The teacher and student networks of all the methods are trained for 200 epochs. The initial learning rate is set to 0.1 and decayed by 0.2 at 60, 120, and 160 epochs. The feature-based classification module is trained with 20 epochs, and other parameters remain unchanged.

For experiments on the CIFAR-100 dataset, following previous works [30, 34], a wide residual network (WRN) [70] is used. The detailed network architecture is provided in Table 1. The performance of our method is reported on three teacher-student pairs, and detailed teacher-student settings and their accuracy are provided in Table 2. Our method is compared with the logits-based knowledge distillation method KD [12], and ten feature-based knowledge distillation methods: FitNets [27], AT [38], Jacobian [32], FT [28], AB [42], SP [34], Margin [29], MFD [33], TOFD [45], and SemCKD [71].

As summarized in Table 3, our SPU achieves state-of-the-art accuracy on all three teacher-student pairs. Specifically, for the teacher-student pair (1) that varies in width, SPU outperforms the five feature-based knowledge distillation methods by 0.06% to 1.67%. For the teacher-student pair (2) that varies in depth, SPU outperforms the five feature-based knowledge distillation methods by 0.31% to 2.14% and surpasses KD by 0.99%. The proposed SPU also

Table 1 The configuration of the WideResNet networks used in the CIFAR-100 experiments. The network architecture is denoted as WRN-*j*-*k*, *j* is the depth and *k* is the channel multiplication factor. The standard Wide ResNet blocks are used in each layer group

Group	Block Type	Kernel Size	Channel Number	Block Number	Output Size
1	Conv	3	16	1	32×32
2	WRN block	3	$16k$	$(j-4)/6$	16×16
3	WRN block	3	$32k$	$(j-4)/6$	16×16
4	WRN block	3	$64k$	$(j-4)/6$	8×8
5	AvgPool-FC	1	100	1	1×1

Table 2 Network architectures and accuracy (%) of teacher and student on the CIFAR-100

	Compression type	Teacher	Accuracy	Size	Student	Accuracy	Size	Compress ratio
1	Depth	WRN-28-4	78.91	5.87M	WRN-16-4	77.38	2.77M	47.2%
2	Channel	WRN-28-4	78.91	5.87M	WRN-28-2	75.70	1.47M	25.0%
3	Depth & channel	WRN-28-4	78.91	5.87M	WRN-16-2	73.33	0.7M	11.9%

Table 3 Experimental results in accuarcy (%) on the CIFAR-100 dataset. The best results are illustrated in bold

	KD	FitNets	AT	Jacobian	FT	AB	SP	Margin	MFD	TOFD	SemCKD	Ours
1	78.76	77.49	77.57	77.45	78.65	78.14	78.51	78.64	78.85	79.06	78.62	**79.12**
2	76.53	76.16	75.38	76.11	76.21	76.38	76.41	76.40	77.21	77.15	77.43	**77.52**
3	74.71	73.52	73.87	74.42	74.83	74.68	74.36	73.89	75.26	74.98	75.95	**76.36**

Table 4 The configuration of the ShuffleNetV2 used in the CINIC-10 experiments. Standard ShuffleNetV2 blocks are used in each layer group. k is the channel multiplication factor

Group	Block Type	Kernel Size	Channel	Number Block Number	Output Size
1	Conv	3	24	1	32×32
2	Sh.Net2 block	3	$116k$	4	16×16
3	Sh.Net2 block	3	$232k$	8	8×8
4	Sh.Net2 block	3	$464k$	4	4×4
5	Sh.Net2 block	3	$1024 \max(1, k)$	1	4×4
6	AvgPool-FC	1	10	1	1×1

Table 5 The detailed teacher-student settings and experimental results of accuracy (%) on the CINIC-10 dataset. The best results are illustrated in bold

	Teacher	Size	Acc	Student	Size	Acc	KD	AT	SP	Margin	MFD	TOFD	SemCKD	Ours
1	Sh.NetV2-2.0	5.37M	85.64	Sh.NetV2-1.0	1.27M	83.35	84.35	84.73	84.65	84.90	84.97	85.27	84.86	**85.42**
2	Sh.NetV2-2.0	5.37M	85.64	Sh.NetV2-0.5	0.36M	76.72	79.58	80.29	80.21	79.34	80.59	79.91	79.72	**81.21**
3	Sh.NetV2-1.0	1.27M	83.35	Sh.NetV2-0.5	0.36M	76.72	78.64	77.21	77.62	78.92	79.11	78.75	79.10	**79.17**

performs best on the third teacher-student pair and outperforms the second-best by 0.41%. These experiments demonstrate the universality of the proposed SPU scheme.

4.2 CINIC-10

CINIC-10 [72] is a large classification dataset augmented from CIFAR-100. It contains all the images from CIFAR-100, and 210,000 images from the ImageNet database [73]. All the images are downscaled to 32×32. All 270,000 images are equally split into three subsets: training, validation, and testing. Compared to CIFAR datasets, experimental results on CINIC-10 are more convincing as more numbers and various images are contained. On CINIC-10, we set batch size 96, and the teacher and student networks of all methods are trained with 140 epochs. The learning rate is initially set to 0.01, and then decayed by a factor of 0.1 at epochs 100 and 120. Similar to the experiments on CIFAR-100, we use the SGD optimizer with a momentum of 0.9 and weight decay of 0.0005. The feature-based classification module is trained with 30 epochs, and other parameters remain unchanged.

Following the recent works [30, 34], we use several variants of ShuffleNetV2 [74] for the experiments on the CINIC-10 dataset. The detailed setting of ShufflenetV2 is provided in Table 4. On the CINIC-10 dataset, we compare our method with the logits-based knowledge distillation method KD [12], and six feature-based knowledge distillation methods: AT [38], SP [34], Margin [29], MFD [33], TOFD [45], and SemCKD [71]. As shown in Table 5, the proposed SPU performs best in all three teacher-student pairs and outperforms the second-best by 0.07%∼0.62%.

4.3 ImageNet

ImageNet [73] is a large-scale classification dataset with 1000 categories that contains 1.2 million training images and 50,000 validation images. All the images are downscaled to 224×224. On ImageNet, we set the batch size as 128, and the teacher and student networks of all methods are trained with 120 epochs. The learning rate is initially set to 0.1 and then decayed by a factor of 0.1 at epochs 60 and 90. We use the SGD optimizer with a momentum of 0.9 and weight decay of 0.0005.

Table 6 Experimental results in accuracy (%) on the ImageNet dataset. The best results are illustrated in bold. The teacher is ResNet34 and the student is ResNet18

	Teacher	Student	KD	AT	SP	Margin	CC	CRD	SPU
Top-1	73.31	69.75	70.66	70.69	70.62	70.81	69.96	71.17	**71.31**
Top-5	91.42	89.07	89.88	90.01	89.80	89.98	89.17	90.13	**90.28**

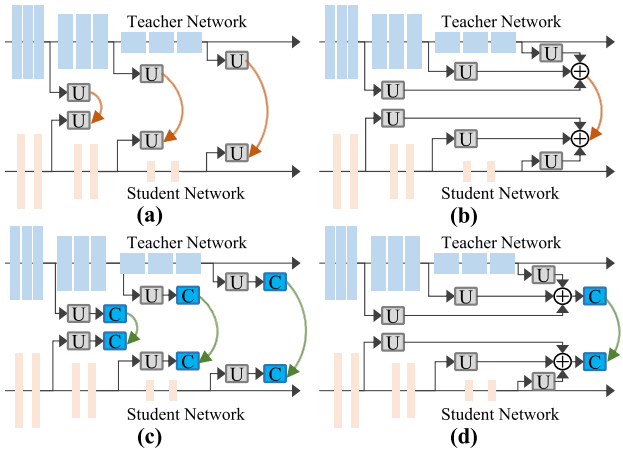

Figure 3 Illustration of different training strategies. **U** denotes the parameter-free feature uniformization operation, and **C** is the feature-based classification module. The orange lines denote the $\mathcal{L}_2$ loss, and the green lines represent the logits-based knowledge distillation, same as $\mathcal{L}^{s}_{feat}$

For experiments on the ImageNet dataset, following previous works [75, 76] to conduct experiments on ResNet [1], the teacher is ResNet34 and the student is ResNet18. Top-1 and Top-5 accuracy results are reported in Table 6. We compare our method with the logits-based knowledge distillation method KD [12], and the feature-based knowledge distillation methods: AT [38], SP [34], Margin [29], CC [75], and CRD [76]. It is observed that our method can achieve the best performance.

4.4 Model analysis

Ablation study about different training strategies As demonstrated in Fig. 3, according to whether features are aggregated and whether feature-based classification modules are adopted, we can obtain four training strategies. It is notable that when the feature-based classification module is not utilized, we use the $\mathcal{L}_2$ loss to train the student referring to previous feature distillation works. In contrast, the student is supervised by the predicted probability of the teacher, and the loss function is the same as the logits-based knowledge distillation loss [12]. For the former, knowledge is transformed by reducing the similarity between the feature pairs. For the latter, knowledge is transformed by reducing the divergence between two class-specified probabilities.

The detailed description of the four training strategies are as follows:

(1) *SS:* features from different group layers are trained *s*eparately, and knowledge is transformed in the form of *s*imilarity. This is the classical setting of feature distillation in previous works.
(2) *AS:* features are *a*ggregated together, and knowledge is transformed in the form of *s*imilarity.
(3) *SP:* features are trained *s*eparately, and knowledge is transformed in the form of class-specified *p*robability.
(4) *AP:* features are *a*ggregated together and are transformed in the form of class-specified *p*robability.

There are two targets with which we compare these four training strategies. First, we aim to explore which form of knowledge is better? Similarity or class-specified probability. Second, should the features from different group layers be aggregated together? We conduct comparison experiments on the CIFAR-100 dataset with three teacher-student pairs. As shown in Table 7, two phenomena are observed:

(1) *AS* outperforms *SS* on all three teacher-student pairs, and *AP* also outperforms *SP*, which indicates that aggregating features from different group layers is better than training each group layer separately. This is because when the features from different group layers are aggregated, each student group layer implicitly receives information from multiple groups of the teacher. Making a single block of the

Table 7 Accuracy (%) of different training strategies on the CIFAR-100 dataset. The best results are illustrated in bold

Strategy	WRN-16-4	WRN-28-2	WRN-16-2
Baseline	77.38	75.70	73.33
SS	77.62	75.74	73.54
AS	77.85	75.95	73.65
SP	78.28	76.56	74.41
AP	79.12	77.52	76.36

student receive knowledge from multiple groups of the teacher has been proven to be useful in MFD [33]. MFD matches each student channel to more than one teacher channel by a pooling operation.

(2) In all three teacher-student pairs, *SP* outperforms *SS*, and *AP* outperforms *AS*. This indicates that transforming the class-specified probability is better than minimizing the similarity between teacher and student features. We believe that this phenomenon is for two reasons. On the one hand, ground-truth supervision would make the predicted logits of the classification module contain more semantic-aware information and less redundant information. On the other hand, because neural networks are usually non-linear due to the adaptive function, and the teacher and student have different architectures, features of teacher and student can be regarded as two matrices generated by two different non-linear functions. Under this circumstance, even though minimizing the similarity between these two matrices will play a role, it is too absolute and can also cause some misguidance to the student.

Sensitivity analysis for hyperparameter L In the proposed SPU scheme, we introduce a novel self-attention feature uniformization operation that can convert features of different dimensions into vectors of the same length L. L is a global hyperparameter and is defined before the training process. As depicted in Fig. 4, we explore how the length of the converted vector affects the performance of SPU. Experiments are conducted on the CIFAR-100 dataset with three teacher-student pairs.

It is noted that the accuracy of SPU first increases when L changes from 1024 to 4096, because when the converted vector is too short, both the information it contains and the compacity of the feature-based classification module will be limited, which will make the student learn insufficient knowledge. In addition, the accuracy will decrease when L changes from 4096 to 16,384. We believe this is because the converted vectors may contain redundant information beyond the compacity of the student. The feature-based classification module is more likely to fall into overfitting as the number of parameters increases.

Moreover, from the overall point of view on three teacher-student pairs, the change in performance is not dramatic, which indicates that the proposed method is not sensitive to L. In addition, another phenomenon is that the changing trend of the three experiments is similar. This is because the classification module is trained based on the features of the teacher network.

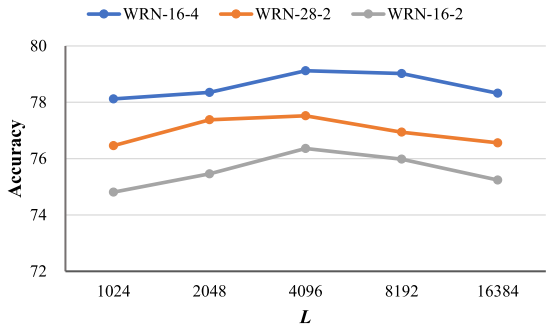

Figure 4 Accuracy(%) with the changes of the hyperparameter L on the CIFAR-100 dataset. We report the three teacher and student pairs in the CIFAR-100 experiments, and the teacher is WRN-28-4 in all three pairs

5 Conclusion

In this paper, a simple by efficient feature-based knowledge distillation method named SPU is proposed. Unlike previous works that unify the teacher-student features in a class-agnostic way, SPU takes the classification into consideration and introduces a dependent feature-based classification module to distill semantic-aware information from the intermediate features of the teacher. As the proposed semantic-aware knowledge distillation requires all the features of the network to be uniform to the same dimension, we present a novel self-attention module to resolve the dimension gap problem between features. Moreover, since the proposed self-attention feature uniformization operation is parameter-free, and the weights of the feature-based classification module are frozen in the training process of the student, the feature transformations of the teacher and student are equivalent, which will help to avoid inaccurate knowledge transfer. Experiments on the CIFAR-100 and CINIC-10 datasets demonstrate the effectiveness of the proposed SPU scheme. In the future, we plan to extend our method to more complex tasks such as object detection or semantic segmentation by distilling task-specified knowledge from the intermediate features.

Funding

This work was supported in part by Key-Area Research and Development Program of Guangdong Province (Grant No. 2021B0101200001), and the National Natural Science Foundation of China (Grant No. 62293543), and the National Natural Science Foundation of China (Grant No. U21B2048), and the Open Research Projects of Zhejiang Lab (Grant No. 2019KD0AD01/010).

Abbreviations
KD, knowledge distillation; SPU, semantic-aware feature distillation with parameter-free uniformization; SP, similarity preserving; WRN, wide residual network.

Availability of data and materials
Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by GG and LH. GG and DZ contributed significantly to the establishment of the methodological framework and manuscript preparation. JH and LW helped perform the analysis with constructive discussions and revised the manuscript. All authors commented on previous versions of the manuscript, and all authors read and approved the final manuscript.

Author details
[1]Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an, China. [2]Beijing Technology and Business University, Beijing, China. [3]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi, China. [4]Xijing Hospital, The Fourth Military Medical University, Xi'an, China. [5]Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, China.

Received: 4 August 2022 Revised: 29 November 2022
Accepted: 23 February 2023 Published online: 08 May 2023

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Los Alamitos: IEEE.
2. Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE conference on computer vision* (pp. 1440–1448). Los Alamitos: IEEE.
3. Wang, W., Shen, J., & Porikli, F. (2015). Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3395–3402). Los Alamitos: IEEE.
4. Yang, L., Han, J., Zhao, T., Lin, T., Zhang, D., & Chen, J. (2021). Background-click supervision for temporal action localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 9814–9829.
5. Guo, G., Han, J., Wan, F., & Zhang, D. (2021). Strengthen learning tolerance for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7403–7412). Los Alamitos: IEEE.
6. Yang, L., Han, J., & Colar, D. Z. (2022). Effective and efficient online action detection by consulting exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3160–3169). Los Alamitos: IEEE.
7. Ravanelli, M., Parcollet, T., & Bengio, Y. (2019). The pytorch-kaldi speech recognition toolkit. In *IEEE international conference on acoustics, speech and signal processing* (pp. 6465–6469). Los Alamitos: IEEE.
8. Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: a systematic review. *IEEE Access*, *7*, 19143–19165.
9. Dong, X., & Yang, Y. (2019). Network pruning via transformable architecture search. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 760–771). Red Hook: Curran Associates.
10. Dong, X., & Yang, Y. (2019). Searching for a robust neural architecture in four GPU hours. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1761–1770). Los Alamitos: IEEE.
11. Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2654–2662). Red Hook: Curran Associates.
12. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint. arXiv:1503.02531.
13. You, S., Xu, C., Xu, C., & Tao, D. (2017). Learning from multiple teacher networks. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1285–1294). New York: ACM.
14. Lan, X., Zhu, X., & Gong, S. (2018). Knowledge distillation by on-the-fly native ensemble. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 7528–7538). Red Hook: Curran Associates.
15. Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., & Li, L.-J. (2017). Learning from noisy labels with distillation. In *Proceedings of the European conference on computer vision* (pp. 1910–1918). Los Alamitos: IEEE.
16. Hyun Cho, J., & Hariharan, B. (2019). On the efficacy of knowledge distillation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4794–4802). Los Alamitos: IEEE.
17. Leng, C., Dou, Z., Li, H., Zhu, S., & Jin, R. (2018). Extremely low bit neural network: squeeze the last bit out with ADMM. In *Proceedings of the AAAI conference on artificial intelligence 2018* (pp. 3466–3473). Menlo Park: AAAI Press.
18. Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10687–10698). Los Alamitos: IEEE.
19. Shi, M., Qin, F., Ye, Q., Han, Z., & Jiao, J. (2017). A scalable convolutional neural network for task-specified scenarios via knowledge distillation. In *IEEE international conference on acoustics, speech and signal processing* (pp. 2467–2471). Los Alamitos: IEEE.
20. Li, Q., Jin, S., & Yan, J. (2017). Mimicking very efficient network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6356–6364). Los Alamitos: IEEE.
21. Chen, G., Choi, W., Yu, X., Han, T. X., & Chandraker, M. (2017). Learning efficient object detection models with knowledge distillation. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 742–751). Red Hook: Curran Associates.
22. Wei, Y., Pan, X., Qin, H., Ouyang, W., & Yan, J. (2018). Quantization mimic: towards very tiny CNN for object detection. In *Proceedings of the European conference on computer vision* (pp. 267–283). Berlin: Springer.
23. Wang, T., Yuan, L., Zhang, X., & Feng, J. (2019). Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4933–4942). Los Alamitos: IEEE.
24. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2604–2613). Los Alamitos: IEEE.
25. Fang, C., Wang, L., Zhang, D., Xu, J., Yuan, Y., & Han, J. (2022). Incremental cross-view mutual distillation for self-supervised medical CT synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 20645–20654). Los Alamitos: IEEE.
26. Yim, J., Joo, D., Bae, J., & Kim, J. (2017). 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). Los Alamitos: IEEE.
27. Romero, A., Ballas, N., Ebrahimi Kahou, S., Chassang, A., Gatta, C., & Bengio, Y. (2015). Fitnets: hints for thin deep nets. [Paper presentation]. *International conference on learning representations*, San Diego, USA.
28. Kim, J., Park, S., & Kwak, N. (2018). Paraphrasing complex network: network compression via factor transfer. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31* (pp. 2760–2769). Red Hook: Curran Associates.
29. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., & Choi, J. Y. (2019). A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1921–1930). Los Alamitos: IEEE.
30. Guan, Y., Zhao, P., Wang, B., Zhang, Y., Yao, C., Bian, K., & Tang, J. (2020). Differentiable feature aggregation search for knowledge distillation. In *Proceedings of the European conference on computer vision* (pp. 469–484). Berlin: Springer.
31. Passalis, N., & Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European conference on computer vision* (pp. 268–284). Berlin: Springer.

32. Srinivas, S., & Fleuret, F. (2018). Knowledge transfer with Jacobian matching. In *Proceedings of the international conference on machine learning* (pp. 4723–4731). PMLR.
33. Yue, K., Deng, J., & Zhou, F. (2020). Matching guided distillation. In *Proceedings of the European conference on computer vision* (pp. 312–328). Berlin: Springer.
34. Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1365–1374). Los Alamitos: IEEE.
35. Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 535–541). New York: ACM.
36. Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE international conference on computer vision* (pp. 3967–3976). Los Alamitos: IEEE.
37. Yu, L., Oguz Yazici, V., Liu, X., van de Weijer, J., Cheng, Y., & Ramisa, A. (2019). Learning metrics from teachers: compact networks for image embedding. In *Proceedings of the IEEE international conference on computer vision* (pp. 2907–2916). Los Alamitos: IEEE.
38. Zagoruyko, S., & Komodakis, N. (2017). Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. [Paper presentation]. *International conference on learning representations*, Toulon, France.
39. Ahn, S., Xu Hu, S., Damianou, A., Lawrence, N. D., & Dai, Z. (2019). Variational information distillation for knowledge transfer. In *Proceedings of the IEEE international conference on computer vision* (pp. 9163–9171). Los Alamitos: IEEE.
40. Koratana, A., Kang, D., Bailis, P., & Lit, M. Z. (2019). Learned intermediate representation training for model compression. In *Proceedings of the international conference on machine learning* (pp. 3509–3518). PMLR.
41. Liu, H., Simonyan, K., & Yang, Y. (2019). DARTS: Differentiable architecture search. [Paper presentation]. *International conference on learning representations*, New Orleans, USA.
42. Heo, B., Lee, M., Yun, S., & Choi, J. Y. (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3779–3787). Menlo Park: AAAI Press.
43. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., & Duan, Y. (2019). Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7096–7104). Los Alamitos: IEEE.
44. Hinton, G., & Roweis, S. T. (2002). Stochastic neighbor embedding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 833–840). Cambridge: MIT Press.
45. Zhang, L., Shi, Y., Shi, Z., Ma, K., & Bao, C. (2020). Task-oriented feature distillation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems 33* (pp. 14759–14771). Red Hook: Curran Associates.
46. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv preprint. arXiv:1409.3215.
47. Lin, Z., Feng, M., Nogueira dos Santos, C., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. arXiv preprint. arXiv:1703.03130.
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Red Hook: Curran Associates.
49. Niu, Y., Zhang, H., Zhang, M., Zhang, J., Lu, Z., & Wen, J.-R. (2019). Recursive visual attention in visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6679–6688). Los Alamitos: IEEE.
50. Fan, H., Zhu, L., Yang, Y., & Wu, F. (2020). Recurrent attention network with reinforced generator for visual dialog. *ACM Transactions on Multimedia Computing Communications and Applications*, *16*(3), 1–16.
51. Sun, G., Wang, W., Dai, J., & Van Gool, L. (2020). Mining cross-image semantics for weakly supervised semantic segmentation. In *Proceedings of the European conference on computer vision* (pp. 347–365). Berlin: Springer.
52. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., & Van Gool, L. (2021). Exploring cross-image pixel contrast for semantic segmentation. arXiv preprint. arXiv:2101.11939.
53. Zhao, B., Li, X., & Lu, X. (2019). Cam-rnn: co-attention model based rnn for video captioning. *IEEE Transactions on Image Processing*, *28*(11), 5552–5565.
54. Devlin, J., Chang, M.-W., Lee, K., & Bert, K. T. (2018). Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805.
55. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the association for computational linguistics* (pp. 2978–2988). Stroudsburg: ACL.
56. Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., & Zhang, Z. (2019). Star-transformer. arXiv preprint. arXiv:1902.09113.
57. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *Proceedings of the international conference on machine learning* (pp. 4055–4064). PMLR.
58. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision* (pp. 213–229). Berlin: Springer.
59. Li, Y., Lin, Y., Xiao, T., & Zhu, J. (2021). An efficient transformer decoder with compressed sub-layers. arXiv preprint. arXiv:2101.00542.
60. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Informer, W. Z. (2021). Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11106–11115). Menlo Park: AAAI Press.
61. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794–7803). Los Alamitos: IEEE.
62. Liu, N., Zhang, N., & Han, J. (2020). Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 13756–13765). Los Alamitos: IEEE.
63. Zhang, W., Wang, B., Ma, S., Zhang, Y., & Zhao, Y. (2021). I2net: mining intra-video and inter-video attention for temporal action localization. *Neurocomputing*, *444*, 16–29.
64. Guo, G., Liu, Z., Zhao, S., Guo, L., & Liu, T. (2021). Eliminating indefiniteness of clinical spectrum for better screening COVID-19. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1347–1357.
65. Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., & Girshick, R. (2019). Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 284–293). Los Alamitos: IEEE.
66. Zhu, L., & Yang, Y. (2020). Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4344–4353). Los Alamitos: IEEE.
67. Wug Oh, S., Lee, J.-Y., Xu, N., & Kim, S. J. (2019). Video object segmentation using space-time memory networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 9226–9235). Los Alamitos: IEEE.
68. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. [Paper presentation]. *The 31st conference on neural information processing systems (NIPS 2017)*, Long Beach, USA.
69. Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.
70. Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *Proceedings of the British machine vision conference* (pp. 87.1–87.12). Durham: BMVA Press.
71. Chen, D., Mei, J.-P., Zhang, Y., Wang, C., Wang, Z., Feng, Y., & Chen, C. (2021). Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 7028–7036). Menlo Park: AAAI Press.
72. Darlow, L.N., Crowley, E. J., Antoniou, A., & Storkey, A. J. (2018). CINIC-10 is not ImageNet or CIFAR-10. arXiv preprint. arXiv:1810.03505.
73. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: a large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255). Los Alamitos: IEEE.
74. Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision* (pp. 116–131). Berlin: Springer.
75. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., & Zhang, Z. (2019). Correlation congruence for knowledge distillation. In *Proceedings of the IEEE international conference on computer vision* (pp. 5007–5016). Los Alamitos: IEEE.
76. Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive representation distillation. [Paper presentation]. *International conference on learning representations*, Addis Ababa, Ethiopia.

Publisher's Note