Pay Attention to Your Positive Pairs: Positive Pair Aware Contrastive Knowledge Distillation

Zhipeng Yu
SEECE, UCAS
yuzhipeng21@mails.ucas.ac.cn

Qianqian Xu*
IIP, ICT, CAS
xuqianqian@ict.ac.cn

Yangbangyan Jiang
SKLOIS, IIE, CAS
SCS, UCAS
jiangyangbangyan@iie.ac.cn

Haoyu Qin
SenseTime Group Limited
qinhaoyu1@sensetime.com

Qingming Huang*
SCST, UCAS
IIP, ICT, CAS
BDKM, CAS
Peng Cheng Laboratory
qmhuang@ucas.ac.cn

ABSTRACT

Deep neural networks have achieved impressive success on various multimedia applications in the past decades. To reach a higher performance on real-world resource-constrained devices with large models that are already learned, knowledge distillation, which aims at transferring representational knowledge from a large teacher network into a small student network, has attracted increasing attention. Recently, contrastive distillation methods have achieved superior performance in this area, due to the powerful representability brought by contrastive/self-supervised learning. These models often transfer knowledge through individual samples or inter-class relationships, while ignoring the correlation lying among intra-class samples, which convey abundant information. In this paper, we propose a Positive pair Aware Contrastive Knowledge Distillation (PACKD) framework to extend the contrastive distillation with more positive pairs to capture more abundant knowledge from the teacher. Specifically, it pulls together features of pairs from the same class learned by the student and teacher while simultaneously pushing apart those of pairs from different classes. With a positive-pair similarity weighting strategy based on optimal transport, the proposed contrastive objective is able to improve the feature discriminability between positive samples with large visual discrepancies. Experiments on different benchmarks demonstrate the effectiveness of the proposed PACKD.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

Knowledge Distillation, Neural Networks, Contrastive Learning

MM '22, October 10–14, 2022, Lisbon, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548256

ACM Reference Format:
Zhipeng Yu, Qianqian Xu, Yangbangyan Jiang, Haoyu Qin, and Qingming Huang. 2022. Pay Attention to Your Positive Pairs: Positive Pair Aware Contrastive Knowledge Distillation . In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548256

1 INTRODUCTION

Recently, deep neural networks bring tremendous success on a wide range of multimedia tasks, such as image classification [10, 17], cross-modal retrieval [40] and visual question answering [15]. Although for these problems, deeper and wider models usually exhibit better performance, their requirements for greater computational resources are usually prohibitive in real-world applications. Correspondingly, how to achieve the trade-off between accuracy and speed has become an emergency problem in resource-limited scenarios. As a popular way of model compression for this case, knowledge distillation is proposed to learn a student model under the guidance of an off-the-shelf teacher model, which is usually equipped with a more complicated architecture or trained on additional data not available for the student. Namely, the goal of knowledge distillation is to transfer the "knowledge" learned by the teacher to the student and then narrow the performance gap between the teacher and student as much as possible.

A straightforward way to distill the knowledge is to minimize the discrepancy between the output probability distribution of the teacher and student [12]. As the prediction offers more specific guidance than the one-hot labels, such "dark knowledge" has been proved beneficial for students. However, considering that the probability distribution contains very limited representation information, subsequent methods turn to mimic the intermediate features of the teacher [11, 24, 36]. Recently, more and more attention has been further attracted to capture the knowledge lying in the correlation among different instances [22]. As a typical framework, CRD [30] learns the student representations in a contrastive way where not only features extracted by the student and teacher for the same instance should be close, but also those for different instances should be distant. Then SSKD [33] introduces an auxiliary self-supervised

[*] Corresponding author.

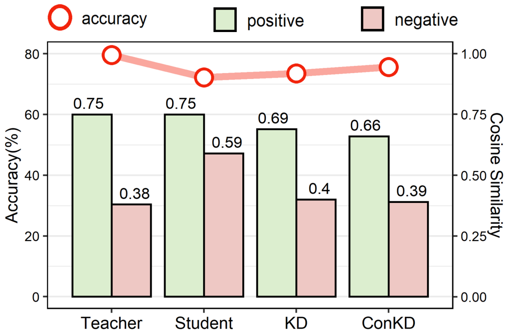


Figure 1: Average cosine similarity between penultimate features of positive/negative pairs and Top-1 accuracy of different models on CIFAR100 (with ResNet32x4 as the teacher and ResNet8x4 as the student).

task for better extracting the knowledge. With cross-sample correlation, these contrastive distillation methods have made great performance improvements over conventional KD models.

Nevertheless, no matter whether there are negative instances, existing KD methods essentially align the knowledge based on each (positive) instance. Indeed, such instance-wise information and inter-class correlation could improve the discriminability of the learned features. On the contrary, the intra-class correlation, a piece of important information in classification, is ignored by these models. Merely pushing apart features of inter-class samples might not be sufficient to exploit the rich knowledge learned by the teacher.

As shown in Fig. 1., after visualizing the average similarity scores of both positive and negative pairs for different models, we can observe that the original student network achieves a very close positive-pair similarity score to the teacher, but also exhibits a much larger negative-pair similarity. With the help of KD or Contrastive KD which mainly focuses on instance-wise information or inter-class correlation, the negative-pair similarity is effectively decreased to a similar level as the teacher, so the model’s accuracy could be significantly improved. However, it also needs to be noticed that they also result in a distinct degradation over the positive similarity. And such an incomplete knowledge transfer might be one of the causes of the remaining performance gap between teacher and distilled student models. Therefore, it is necessary to simultaneously take the intra- and inter-class correlation into account in the distillation process.

Motivated by this, we propose a Positive pair Aware Contrastive Knowledge Distillation (PACKD) framework to simultaneously pull intra-class sample features from the student and teacher together and push inter-class ones away. As its name suggests, here, we take a step further to incorporate more positive sample pairs of each class in the contrastive scheme. Specifically, for each positive sample, we include many other positives that are drawn from either augmentations or samples in the same class to form a positive anchor set. Then features in the positive set learned by the student and teacher are required to be relatively close, while features of positive and negative samples are forced to be distant.

However, noticing that there could be a large visual discrepancy between some positive samples, forcing these positives too close to each other may also compromise the model’s representation ability. To alleviate this issue, in our contrastive objective, we further weigh the similarity between features from the student and teacher networks among the positive set, which is based on optimal transport, to pull together intra-class samples in a more reasonable way. The contributions of this paper are summarized as follows:

- We propose a positive pair aware contrastive distillation framework called *PACKD*. For each positive sample, a positive set is constructed for the contrastive objective to take the intra-class correlation into account to enhance the distillation quality.
- We further present a positive-pair similarity weighting strategy based on optimal transport for the contrastive objective to discriminate positive samples with large visual discrepancies.
- Extensive evaluation on several popular benchmarks shows that the proposed method could outperform existing models.

2 RELATED WORK

Model compression. To develop efficient deep models, recent works usually focus on the following aspects. A straightway is to design a lightweight but powerful network. For example, depthwise convolution is used to replace standard convolution[25]. Pointwise group convolution and channel shuffle are proposed to reduce the burden of computation[38] while maintaining high accuracy. Another way is network pruning, since the original convolution network has many redundant parameters, pruning extra neurons or filters can boost the inference time[8]. Quantization, widely used in industrial applications, seeks to use low-precision bits to store the model’s weights or activation outputs[7]. Knowledge distillation(KD) [12], which is mainly concerned with this paper, aims at transferring knowledge from a cumbersome network to a small network.

Knowledge Distillation. The standard approach to achieving KD is minimizing the Kullback-Leibler (KL) divergence between the probabilistic outputs of a teacher and a student. Sometimes, the student network is trained by the classifier output of an ensemble of teacher networks [12]. Such logit-based distillation methods mainly rely on effective regularization and optimization techniques rather than learning schemes or architectures. DML [39] proposes a mutual learning manner to train students and teachers simultaneously. TAKD [18] introduces an intermediate-sized network named “teacher assistant” to bridge the gap between teachers and students. Besides, several works also aim at interpreting the classical KD method [23, 28]. To further resist the performance degradation in teacher-student transfer, leveraging more information from the pretrained teacher model, especially the intermediate layers becomes a more popular solution. For example, FITNET [24] proposes to transfer the knowledge using not only final outputs but also intermediate features. In AT [36], the authors propose an attention-based method to match the activation-based and gradient-based spatial attention maps. Another line of knowledge distillation methods focuses on transferring the relationship between features, rather than the actual features themselves. In SP [32], CC [22], RKD [20], the relationships across samples are employed to guide student learning high-level representations. VID [1], PKT [21] reformulate

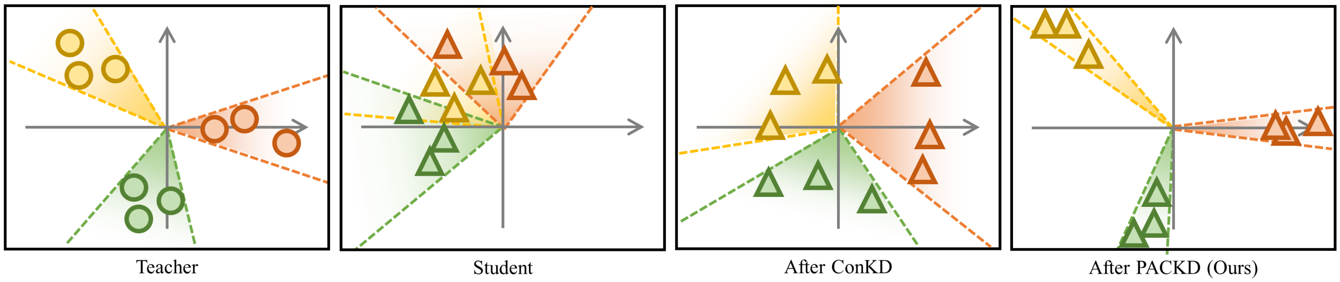


Figure 2: Illustration of feature space for different models. Teacher models usually have more separable feature space with smaller inter-class cosine similarities than students. Existing knowledge distillation methods like ConKD can help the student to reduce the inter-class similarity, but may not capture the correlation among intra-class samples. By contrast, the proposed PACKD pulls together the intra-class samples effectively while pushing away inter-class samples at the same time.

knowledge distillation as a procedure of maximizing the mutual information between the teacher and the student networks. Based on research for classification, knowledge distillation on other multimedia tasks is further studied [13, 15, 16, 29]. Trustworthy distillation methods are also explored. For example, [9] is a data-free distillation method to protect users' privacy in multimedia applications.

Contrastive Knowledge Distillation. Contrastive learning has recently received interest due to its success in self-supervised learning tasks [3, 19, 35]. The key idea is to encourage positive pairs to be close while contrasting negative pairs in a latent space. By applying contrastive learning to representation space, recent knowledge distillation methods such as CRD [30] has achieved great performance improvements over conventional KD models. To name a few, WCoRD [2] leverages both primal and dual forms of Wasserstein distance for global and local knowledge transfer. CRCD [41] distills the inter-sample relation via relation contrastive loss. Besides, SSKD [33] and HSAKD [34] introduce extra self-supervision (SS) signals to extract richer knowledge from the teacher by adding an SS-module for both teacher and student. In this paper, we extend existing contrastive distillation schemes with more positive pairs to explicitly utilize the intra-class correlation. Moreover, using a positive-pair similarity weighting strategy based on optimal transport, PACKD is able to better transfer discriminative information from teacher to student.

3 METHODOLEGY

3.1 Preliminaries

Given a set of n samples denoted by $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the input of i-th image, y_i is the corresponding label. We refer teacher network and student network as $f^T(\cdot)$ and $f^S(\cdot)$, ϕ_i represents the penultimate layer output extracted by $f(x_i)$, z_i denotes the representation feature mapped by a linear projection head $l(\cdot)$, i.e., $z_i = l(\phi_i)$. The predictive class probability distribution $p(x_i;\tau) = \sigma(c(\phi_i)/\tau)$ where $c(\cdot)$ is the classifier, σ is a softmax function, τ is referred as the temperature parameter.

Vanilla knowledge distillation. Since KD is based on the principle that student will benefit from imitating the behaviors of teachers, the goal is often achieved by aligning the output scores between the student and teacher under the guidance of KL-divergence loss as follows [12]:

$$\mathcal{L}_{kd}^{i} = \tau^2 KL(p^T(x_i;\tau) \parallel p^S(x_i;\tau)) \tag{1}$$

Contrastive distillation. Previous work like [3] has proved the effectiveness of the contrastive loss in representation learning. Let $\tilde{z}$ denote the feature extracted from $\tilde{x}$, the augmented variant of the instance x. For the i-th instance, the augmented variant $\tilde{z}_i$ is regarded as the anchor point. Then the feature of the original instance denoted by z_i is used as the positive sample, while some other instance features are sampled to construct a negative sample set $\mathcal{N}_i = \{z_j^-\}_{j=1}^{K}$. A commonly used contrast objective is defined as:

$$\mathcal{L}_{contrast}^{i} = -\log \frac{\exp(sim(\tilde{z}_i, z_i)/\tau)}{\sum_{z_j \in \mathcal{N}_i \bigcup \{z_i\}} \exp(sim(\tilde{z}_i, z_j)/\tau)} \tag{2}$$

where $sim(\cdot)$ is a similarity function. The success might mainly come from that, this process is equivalent to maximizing a lower bound on the mutual information of representations among different augmentations [31]. Recent methods like CRD and SSKD further employ contrastive leaning into KD. Specifically, the anchor feature is extracted by the student, while the positive and negative sample features are obtained from the teacher. Accordingly, a general contrastive distillation objective can be formulated as:

$$\mathcal{L}_{conkd}^{i} = -\log \frac{\exp(h(z_i^S, z_i^T))}{\exp(h(z_i^S, z_i^T)) + \sum_{z^T \in \mathcal{N}_i} \exp(h(z_i^S, z^T))} \tag{3}$$

where

$$h(z_i^S, z_j^T) = cosine(z_i^S, z_j^T)/\tau. \tag{4}$$

Compared with the original contrastive loss Eq. (2), we see that the contrastive distillation objective not only aligns z^T and z^S from the same instance while pushing apart z^T from z^S belonging to different samples. Such a procedure can be viewed as maximizing the mutual information between the teacher and student. Note that the negative samples may be chosen from those z_k^T with $i \neq k$ or with $y_i \neq y_k$, *i.e.*, instances from different instances or classes.

3.2 The Proposed Method

The formulation of previous contrastive distillation objectives reflects that they are mainly based on the instance-wise information or inter-class correlation among instances. However, as we discussed before, they could help the student model to maintain the negative-pair similarity well, while often decreasing the positive-pair similarity at the same time, since the intra-class correlation contained in the label information is not utilized. As shown in Fig. 2,

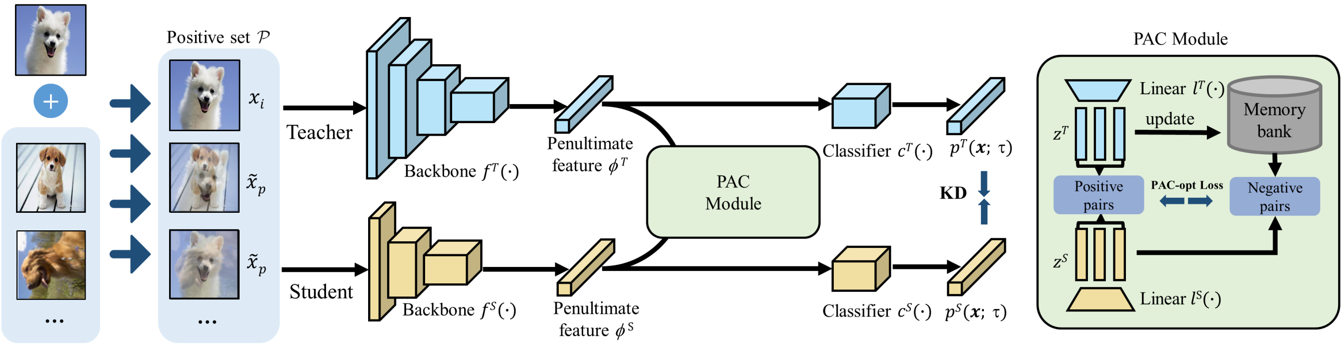


Figure 3: Overview of our proposed PACKD (take the image classification task as an example). In the student training stage, knowledge transfer on both probability and representation space is achieved by vanilla KD loss and the PAC module, respectively. In the PAC module, representation feature z^T and z^S from the same positive set construct multiple positive pairs with a total size $|\mathcal{P}_i|^2$. The PAC loss will encourage these positive pairs to be close, while contrasting negative pairs constructed by z^S and K negative features selected from the memory bank.

besides pushing apart features of inter-class samples, it is possible to incorporate more positive pairs to improve the intra-class similarity learning and further exploit the rich knowledge learned by the teacher. To this aim, we extend the only positive sample in existing contrastive distillation models (*i.e.*, z_i^T in Eq. (3)) to **a positive set** $\mathcal{P}_i$, consisting of several samples from the same class together with the original sample itself, for each anchor image.

The workflow of the proposed PACKD framework is shown in Fig. 3. Both teacher and student models consist of three components: 1) a backbone $f(\cdot)$ for feature extraction; 2) a classifier $c(\cdot)$ for the main classification task; 3) a projection head $l(\cdot)$ to map penultimate features into a latent space with the same dimension, which is the key module for contrastive distillation. In this framework, the teacher and student network are trained separately.

Training the Teacher Network. Following SSKD, the training of teacher network has two steps. In the first step, we train $f^T(\cdot)$ and $c^T(\cdot)$ on the main classification task using a cross-entropy (CE) loss without involving data transformation. In the second step, with fixed $f^T(\cdot)$ and $c^T(\cdot)$, the extra projection head $l^T(\cdot)$ will be learned on a contrastive task using the loss in Eq. (2). The second step aims at adapting the projection head to transform penultimate features ϕ^T from the existing backbone into positive-correlation-aware representations z^T. It is noteworthy that the first step can be replaced by existing pretrained teachers and the second step is highly efficient given that the projection head is usually small.

Training the Student Network. In the student training stage, all the network modules including $f(\cdot), c(\cdot), l(\cdot)$ are trained in an end-to-end manner under the supervision of the learned teacher network. For each input image x_i, the score $p^T(x_i;\tau)$ obtained by the student classifier is used to calculate the CE loss using the corresponding ground-truth label y_i, together with the KD loss in Eq. (1) using the output of the teacher classifier. Meanwhile, its penultimate features ϕ_i^T are fed into a Positive pair Aware Contrastive (PAC) module to carry out the contrastive distillation with multiple positive pairs. In this module, we first choose some other features to construct the corresponding positive and negative sample sets according to their ground-truth labels. Then the following positive pair aware contrastive loss based on Eq. (3) takes both the positive and negative samples as input for contrastive distillation.

Positive pair Aware Contrastive Loss. From the perspective of metric leaning in [27], Eq. (3) can be written as (K+1)-triplet formulation in Eq. (5)

$$\mathcal{L}_{conkd}^{i} = \log(1 + \sum_{z^T \in \mathcal{N}_i} \exp(h(z_i^S, z^T) - h(z_i^S, z_i^T))) \tag{5}$$

Thus, minimizing $\mathcal{L}_{conkd}$ naturally maximizes the distance between positive and negative pairs. Yet Eq. (5) only involves each instance itself as the positive sample, which is far from enough to transfer the intra-class feature relationship. A straightforward way to incorporate sufficient positive pairs into Eq. (5) for each input can be written as:

$$\mathcal{L}_{pac}^{i} = \log\left(1 + \frac{1}{|\mathcal{P}_i|} \sum_{z_p^S \in \mathcal{P}_i^S} \sum_{z^T \in \mathcal{N}_i} \exp(h(z_p^S, z^T) - h(z_p^S, z_i^T))\right) \tag{6}$$

where the superscript S in $\mathcal{P}_i^S$ indicates that this set contains all the features of positive samples in $\mathcal{P}_i$ extracted by the student network, and $|\mathcal{P}_i|$ denotes the cardinality. In this way, student is able to not only imitate the behaviors of teacher, but also reduce the intra-class variance of each class to boost classification performance. Taking all positive pairs within a positive set into account, this loss can be reformulated as follows:

$$\mathcal{L}_{pac}^{i} = \log\left(1 + \frac{1}{|\mathcal{P}_i|^2} \sum_{z_p^S \in \mathcal{P}_i^S,\, z_q^T \in \mathcal{P}_i^T} \sum_{z^T \in \mathcal{N}_i} \exp(h(z_p^S, z^T) - h(z_p^S, z_q^T))\right) \tag{7}$$

Here, similar to $\mathcal{P}_i^S$, $\mathcal{P}_i^T$ consists of all the positive sample features extracted by the teacher network.

However, this objective treats each positive pair equally, ignoring the different importance of positive pairs. Furthermore, when there are large intra-class discrepancies or aggressive augmentations, the obtained positive pairs are not supposed to be pulled together too much. Forcing these samples too close to each other might compromise the discriminability of the features, and thus degrade the performance. To deal with this problem, we reweigh the positive pairs based on optimal transport (OT) according to their importance, which can effectively improve the intra-class similarity and avoid the influence of discrepancies between positive samples.

OT aims at finding a plan to transport mass from a probability distribution to another distribution with a minimum cost. In our case, for each input image, we have two probability vectors $\boldsymbol{u} \in \mathbb{R}^{|\mathcal{P}_i^S|}$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{P}_i^T|}$ as the source and target distributions, which are supported on the student and teacher embeddings of positive samples, *i.e.*, points in $\mathcal{P}_i^S$ and $\mathcal{P}_i^T$, respectively. The cost matrix of moving the points in $\mathcal{P}_i^S$ to those in $\mathcal{P}_i^T$ is denoted by $\boldsymbol{C}$. We hope to find the optimal transport plan $\boldsymbol{\pi} \in \mathbb{R}_+^{|\mathcal{P}_i| \times |\mathcal{P}_i|}$ via solving the following linear optimization problem:

$$
\begin{aligned}
&OT(\boldsymbol{u}, \boldsymbol{v}; C) = \min_{\boldsymbol{\pi}} \langle \boldsymbol{\pi}, \boldsymbol{C} \rangle, \\
&s.t. \sum_{j=1}^{n} \boldsymbol{\pi}_{ij} = u_i, \quad \sum_{i=1}^{n} \boldsymbol{\pi}_{ij} = v_j,
\end{aligned} \tag{8}
$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of matrices that $\langle \boldsymbol{\pi}, \boldsymbol{C} \rangle = \sum_{j=1}^{|\mathcal{P}_i|} \sum_{k=1}^{|\mathcal{P}_i|} \pi_{jk} \cdot C_{jk}$. In practice, the transport cost C_{jk} is calculated by the distance $1 - h(z_j^S, z_k^T)$, which could measure the discrepancy between positive samples across the teacher and student. The probability vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are simply set as $\frac{1}{|\mathcal{P}_i|} \mathbf{1}_{|\mathcal{P}_i|}$, since initially we do not have any prior knowledge about the weights for each data point. The optimal transport plan $\boldsymbol{\pi^*}$ is obtained by the Sinkhorn algorithm originally designed for the entropy-regularized OT problems [5], which could also provide an approximate solution for unregularized OT problems through fast iterations with a very small regularization coefficient (*e.g.*, 10^{-7}). Although this algorithm is with a computational complexity of $O(m^2)$, in our model $m = |\mathcal{P}_i|$ is usually small hence it would not introduce much computational burden.

Then with the optimal weighted similarity $\langle \boldsymbol{\pi}^*, \boldsymbol{h} \rangle$, the enhanced PAC objective is formulated as:

$$
\mathcal{L}_{pac-opt}^{i} = -\log \frac{\exp(\langle \boldsymbol{\pi}^*, \boldsymbol{h} \rangle)}{\exp(\langle \boldsymbol{\pi}^*, \boldsymbol{h} \rangle) + \sum\limits_{z_p^S \in \mathcal{P}_i^S} \sum\limits_{z^T \in \mathcal{N}_i} \exp(h(z_p^S, z^T))} \tag{9}
$$

Positive Sample Generation. Another way to ease the difficulty of learning from intra-class samples is designing proper positive sample generation methods. In PACKD, we adopt a class-invariant mixup technique to further expand the sampling space from the limited training points into a locally linear class-invariant region, denoted as **positive-mix**. Specifically, for a sample x_i, we first randomly select another positive sample x_j from the same class. Then, aggressive augmentations such as large-angle rotation are applied on x_j to make the contradistinction much harder. Finally, we mix the augmented positive sample $\tilde{x}_j$ with x_i by a random ratio $\gamma \in (0, 1)$ as the final positive sample used to construct the $\mathcal{P}_i$:

$$
\tilde{x}_p = \gamma x_i + (1 - \gamma) \tilde{x}_j, \quad s.t. \; y_i = y_j \tag{10}
$$

Negative Sample Generation. Following CRD, we implement a memory bank to preserve the teacher features z^T for each sample from previous batches for a more comprehensive negative sampling. In the student training stage, we construct the negative sample set $\mathcal{N}_i$ from the memory bank by randomly sampling K negative features z^T, whose input labels are different from x_i's.

Overall Loss. In addition to the aforementioned CE loss and KD loss applied to the original data x_i, we also encourage the student's classifier output to be close to that teacher's on the corresponding positive samples $\tilde{x}_p$ generated by the mixup-aug policy. This loss defined on the positive set could be written as:

$$
\mathcal{L}_P^i = \tau^2 \sum_{\tilde{x}_p \in \mathcal{P}_i} KL(p^T(\tilde{x}_p; \tau) \parallel p^S(\tilde{x}_p; \tau)) \tag{11}
$$

Putting all these together, we obtain the overall loss:

$$
\mathcal{L}_{PACKD} = \sum_{i}^{n} \mathcal{L}_{ce}^i + \lambda_1 \mathcal{L}_{kd}^i + \lambda_2 \mathcal{L}_{pac-opt}^i + \lambda_3 \mathcal{L}_P^i \tag{12}
$$

where the λ_is are the balancing hyperparameters.

3.3 Discussion

Previous studies [19, 31] have proven that minimizing the objective $\mathcal{L}_{conkd}$ is equivalent to maximizing the lower bound on the mutual information (MI). For each sample, it holds that:

$$
I(z_i^S, z_i^T) \geq \log(K) - \mathcal{L}_{conkd}^i \tag{13}
$$

where K is the number of negative pairs in the negative sample set, $I(z_i^S, z_i^T)$ represents the mutual information of teacher and student for the same input x_i. Since $\mathcal{L}_{conkd}^i$ ignores the correlation lying among intra-class samples, the proposed objective $\mathcal{L}_{pac}^i$ addresses this limitation by taking all positive pairs in the $\mathcal{P}$ into consideration. Combined with Eq. (13), it shows that $\mathcal{L}_{pac}^i$ actually aims at maximizing the mutual information $I(\mathcal{P}_i^S, \mathcal{P}_i^T)$. Taking a step further, $\mathcal{L}_{pac-opt}$ considers the differences between positive pairs by optimizing the optimal similarity $\langle \boldsymbol{\pi}^*, \boldsymbol{h} \rangle$. Note that according to [19, 31], the optimal $h(\cdot)$ in $\mathcal{L}_{conkd}$ is proportional to the pointwise mutual information, i.e., the density ratio between the joint distribution $p(z_i^S, z_i^T)$ and the product of marginals $p(z_i^S)p(z_i^T)$. Hence, the optimal similarity $\langle \boldsymbol{\pi}^*, \boldsymbol{h} \rangle$ can be regarded as a convex combination of the density ratio, which helps improve the discriminability by adjusting the weights of different positive pairs adaptively based on their cosine distance.

4 EXPERIMENTS

4.1 Experiment Settings

Datasets We use four popular benchmark datasets for evaluation. In these datasets, CIFAR-100 and ImageNet are used for basic performance comparison, while STL-10 and TinyImageNet are for the evaluation of transferability, following [30]. CIFAR-100 [14] is a widely used classification dataset that contains 50K training images with 0.5K images per class and 10K test images of size 32×32. ImageNet [6] is a much larger dataset that provides 1.2 million images from 1K classes for training and 50K for validation. STL-10 [4] consists of a training set of 5K labeled images from 10 classes and 100K unlabeled images and a test set of 8K images. TinyImageNet [6] has 200 classes, each with 500 training images and 50 validation images.

Competitors We compare our method with several representative knowledge distillation methods, including (1) logits based methods: KD[12]; (2) feature based methods: FitNet[24], AT[36], AB [11]; (3) relation based methods: SP[32], CC[22], VID[1], RKD[20], PKT[21]; (4) contrastive distillation methods: CRD[30], WCoRD[2],

Table 1: Top-1 accuracy (%) on CIFAR-100 comparison of SOTA distillation methods. Teacher and student are in the same architecture style. The numbers in Bold and underline denote the best and the second-best results, respectively. For a fair comparison, all comparative methods are combined with conventional KD loss. All experiments are reported the mean result over 3 runs.

Teacher	KD	Contra	SS-	wrn40-2	wrn40-2	resnet56	resnet32x4	resnet110	vgg13
Student		stive	Module	wrn16-2	wrn40-1	resnet20	resnet8x4	resnet32	vgg8
Teacher				76.00	76.00	72.75	79.55	74.07	74.81
Student				73.03	71.32	68.93	72.22	71.45	70.68
KD [12]	✓			75.47	73.66	71.12	73.53	73.08	73.14
FitNet [24]	✓			75.26	74.65	71.07	74.99	72.97	73.78
AT [36]	✓			75.59	74.34	71.22	75.11	73.84	73.65
VID [1]	✓			75.61	74.09	71.37	74.37	73.45	73.40
SP [32]	✓			75.19	73.6	70.97	73.83	73.53	73.09
CC [22]	✓			75.18	73.67	71.18	74.60	73.90	73.08
RKD [20]	✓			75.20	73.66	70.51	74.47	73.26	73.50
PKT [21]	✓			75.20	74.34	71.55	74.26	73.71	73.37
CRD [30]	✓	✓		75.87	74.38	71.41	75.62	73.95	73.91
WCoRD [2]	✓	✓		76.11	74.72	$\underline{71.92}$	76.15	74.20	74.72
CRCD [41]	✓	✓		75.83	74.34	71.37	75.42	73.84	74.13
SSKD [33]	✓		✓	75.57	$\underline{75.62}$	70.15	75.58	73.50	74.76
ours w/o positive-mix	✓	✓		$\underline{76.30}$	75.41	71.38	76.37	74.34	74.95
ours w/o KD		✓		76.24	75.42	71.79	$\underline{76.9}$	$\underline{74.56}$	$\underline{75.03}$
ours	✓	✓		**77.32**	**77.15**	**72.59**	**77.09**	**75.31**	**75.94**

Table 2: Top-1 accuracy (%) comparison of SOTA distillation methods on ImageNet.

	Teacher	Student	KD [12]	SP [32]	AT [36]	CC [22]	CRD [30]	WCoRD [2]	SSKD [33]	ours
Top-1	73.31	69.75	70.66	70.62	70.70	69.96	71.38	71.56	$\underline{71.62}$	**72.02**
Top-5	91.42	89.07	89.98	89.80	90.00	89.17	90.49	90.55	$\underline{90.67}$	**90.80**

CRCD[41]; (5) self-supervised task methods: SSKD[33]. For a fair comparison, all comparative methods are combined with conventional KD loss by default. To further show the strength of the proposed positive pair aware scheme, we add two ablated variants, our method without KD loss or positive-mix samples, for evaluation. All the results are reproduced using officially released codes except WCoRD, the result of which is copied from the original paper.

Implementation details We set τ in $\mathcal{L}_{kd}$ and $\mathcal{L}_{P}$ to be 4, τ in $\mathcal{L}_{pac}$ and $\mathcal{L}_{pac-opt}$ to be 0.07. We set $\lambda_1 = 1$, $\lambda_2 = 0.8$, $\lambda_3 = 2.0$ in Eq. 12. We use 3 augmented positive samples (*i.e.*, $\tilde{x}_p$) in each $\mathcal{P}_i$ and $K = 16384$ for negative sampling. For CIFAR-100, we train all models for 240 epochs with batch size 64 and initialize the learning rate as 0.05, and decay it by 0.1, respectively, at epoch 150, 180 and 210. For ImageNet, we train all the models for 100 epochs. The initial learning rate is 0.1 and decayed by 10 when the epoch is 30, 60 and 90. The SGD optimizer with 5e-4 weight decay and 0.9 momentum is adopted. We conduct all experiments on one Nvidia GeForce GTX 1080 Ti GPU.

Due to the page limit, we provide visualization results for real samples and additional experiment results on cross-modal transfer and positive set construction in the supplementary materials.

4.2 Comparison with State-Of-The-Arts

Setup We compare our PACKD with SOTA distillation methods on CIFAR-100 and ImageNet with various teacher-student pairs based on different backbones including ResNet [10], Wide ResNet (WRN)[37], MobileNetV2 [25], vgg [26] and ShuffleNet [3].

CIFAR100 We report Top-1 accuracies under the same or different architectures, in Table 1 and 3 respectively, where our PACKD consistently outperforms other competitors by a clear margin. The performance improvement of our full-PACKD denoted as **ours** and the best-performing competing methods are 1.10% (averaged on the six same architecture pairs) and 1.19% (averaged on five different architecture pairs). We see that for some network pairs, the proposed model's performance is even higher than the teacher's. The performance improvement of these student models mainly comes from several factors: (1) Some student networks are intrinsically more compatible with distillation tasks. For example, the student ShuffleV1 can exceed its teacher model wrn40-2 by many distillation methods. (2) On top of the above fact, the student's performance is further enhanced via PACKD's more comprehensive learning objective that takes both intra- and inter-class correlations into account, which could transfer the knowledge from teacher to student better. It is noteworthy that even the ablated variant 'ours w/o

Table 3: Top-1 accuracy (%) on CIFAR-100 comparison of SOTA methods in different architecture styles.

Teacher	KD	Contrastive	SS-Module	ResNet50	ResNet50	resnet32x4	resnet32x4	wrn40-2
Student				MobileV2	vgg8	ShuffleV1	ShuffleV2	ShuffleV1
Teacher				78.87	78.87	79.55	79.55	76.00
Student				64.39	70.68	71.60	72.90	71.60
KD [12]	✓			67.83	73.33	74.20	74.94	75.69
FitNet [24]	✓			66.69	72.99	74.87	75.12	76.00
AT [36]	✓			66.9	73.69	76.15	75.36	76.78
VID [1]	✓			68.86	73.50	75.00	75.80	76.17
SP [32]	✓			68.77	73.99	75.39	75.01	76.18
CC [22]	✓			68.99	73.81	75.35	75.07	75.02
RKD [20]	✓			68.55	73.47	74.58	75.20	75.52
PKT [21]	✓			68.43	73.10	74.30	75.97	76.09
CRD [30]	✓	✓		69.39	73.86	75.40	76.20	76.95
WCoRD [2]	✓	✓		70.12	74.86	75.77	76.48	76.68
CRCD [41]	✓	✓		69.67	73.85	75.35	76.07	76.44
SSKD [33]	✓		✓	71.77	75.95	78.14	78.38	77.17
ours w/o positive-mix	✓	✓		70.07	75.44	76.48	77.59	77.27
ours w/o KD		✓		70.96	75.50	77.77	78.20	77.41
ours	✓	✓		**72.83**	**76.71**	**79.02**	**79.63**	**79.19**

positive-mix' achieves better results than other contrastive distillation methods such as CRD, WCoRD and CRCD, which validates the importance of our $\mathcal{L}_{pac-opt}$ in exploring the intra-class correlation. Nevertheless, SSKD outperforms ours w/o positive-mix and obtains the second-best performance in Table 3. Such an advantage could be attributed to the additional design of SSKD to learn cross-sample knowledge by its self-supervised module. Notably, ours w/o KD (without $\mathcal{L}_{kd}$ and $\mathcal{L}_P$) can acquire comparable performance with other methods with KD loss, which further demonstrates the superiority of our proposed PACKD.

ImageNet We conduct one teacher-student pair on ImageNet, i.e., ResNet34 as a teacher and ResNet18 as a student. As shown in Table 2, for both Top-1 and Top-5 accuracy, our PACKD obtains the best performance, especially a performance gain of 0.40% compared to SSKD in terms of Top-1 accuracy. These results on ImageNet demonstrate the validity of PACKD on the large-scale dataset.

4.3 Ablation Study

Effect of positive samples. We validate different positive number (number of $\tilde{x}_p$): 0/1/2/3/4 on CIFAR-100 with four teacher-student pairs, including similar and different architectures. They are only trained with $\mathcal{L}_{ce}$ and $\mathcal{L}_{pac-opt}$. As shown in Fig.4(a), We can observe that the performance of our model is constantly improved as the increasing of positive samples. This indicates that intra-class correlations introduced by $\mathcal{L}_{pac-opt}$ benefit knowledge transfer significantly.

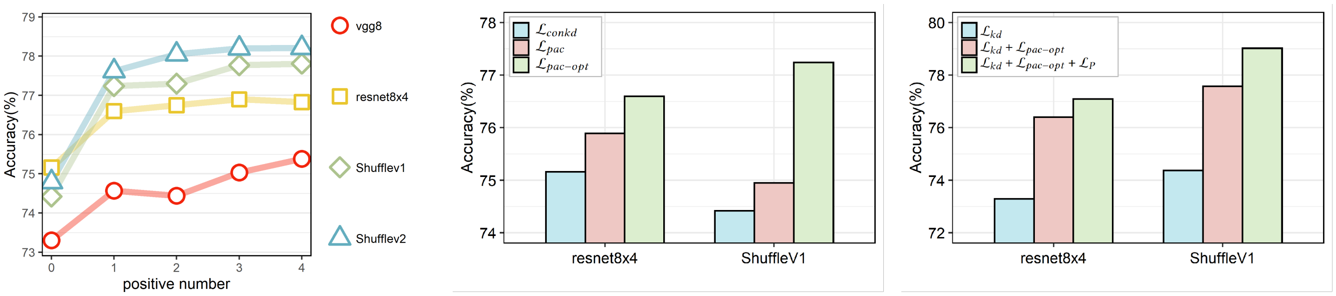
Effect of optimal transport. For $\mathcal{L}_{pac}$ and $\mathcal{L}_{pac-opt}$, positive number is set to 1 and $\mathcal{L}_{conkd}$ is trained with normal data. As shown in Fig.4(b), both $\mathcal{L}_{pac}$ and $\mathcal{L}_{pac-opt}$ surpass the basic $\mathcal{L}_{conkd}$. Meanwhile $\mathcal{L}_{pac-opt}$ substantially outperforms the accuracy upon $\mathcal{L}_{pac}$ (0.71% on resnet8x4 and 2.29% on ShuffleV1). We speculate that the weighting strategy based on optimal transport efficiently reduces intra-class noise in contrastive distillation based on positive pairs.

Effect of loss terms. To explore the effectiveness of designed $\mathcal{L}_{pac-opt}$ and $\mathcal{L}_P$, we conduct the evaluation in three variants, which are conventional KD ($\mathcal{L}_{kd}$), $\mathcal{L}_{kd} + \mathcal{L}_{opt}$ and full setting ($\mathcal{L}_{kd} + \mathcal{L}_{pac-opt} + \mathcal{L}_P$). The results are shown in Fig.4(c), where $\mathcal{L}_{kd}+\mathcal{L}_{pac-opt}$ boost the accuracy by a large margin (more than 2 % in both two setting) compared to the original $\mathcal{L}_{kd}$. Moreover, incorporating $\mathcal{L}_P$ further improves the performance, which means our method can be compatible with traditional logits based distillation methods.

4.4 Further Analysis

Transferability of Learned Representations. Following the linear classification protocol in CRD[30], we freeze the backbone $f^S(\cdot)$ of student network pre-trained on the upstream CIFAR-100, and then train two linear classifiers based on penultimate feature ϕ^S for downstream STL-10 and TinyImageNet, respectively. As shown in Table 4, we can observe that methods acting on representation space usually outperform the basic KD method. Our method PACKD achieves the best performance among all comparative methods, demonstrating that models trained with PACKD can transfer well to other recognition tasks.

Efficiency under Few-shot Scenario. We compare PACKD with KD, CRD and SSKD under few-shot scenarios by retaining 25%, 50%, and 75% training samples while maintaining the original test set. We use the resnet56-resnet20 as the teacher-student pair for evaluation. As shown in Table 5, our method can consistently surpass others by large margins under various few-shot settings. Moreover, it is noteworthy that by using only 25% training samples, our method can achieve comparable accuracy with the baseline trained on the complete set. This is because our method can effectively learn general feature representations from limited data by exploring the intra-class correlation. Previous methods may overfit the limited training data and rarely utilize enough positive pairs. We believe



(a) Effect of positive samples (b) Effect of optimal transport (c) Effect of loss terms

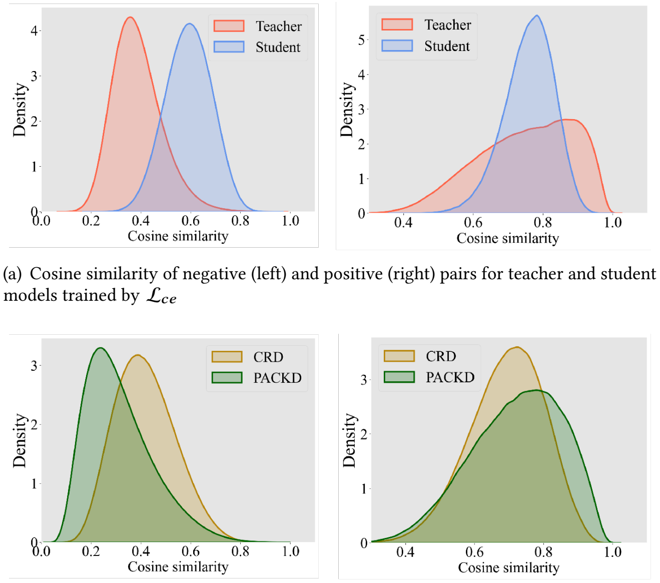
Figure 4: Ablation study on CIFAR100. Student network vgg8 is trained under the pretrained teacher vgg13, while others, i.e., resnet8x4, ShuffleV1, and ShuffleV2, are trained under teacher network resnet32x4.

Table 4: Linear classification Top-1 accuracy (%) of transfer learning on the resnet8x4 pre-trained using the resnet32x4.

Transferred Dataset	Baseline	KD	FitNet	AT	AB	VID	RKD	SP	CC	CRD	SSKD	ours
CIFAR-100⟶ STL-10	69.76	69.56	70.94	70.58	69.47	71.40	71.41	70.96	70.00	70.76	71.89	**72.45**
CIFAR-100⟶ TinyImageNet	34.29	34.77	38.07	37.60	36.40	37.60	38.02	36.06	36.45	38.17	38.56	**38.94**

Table 5: Few shot evaluation on CIFAR100 with different percentage of data.

percentage	KD	CRD	SSKD	ours
25%	64.40	64.71	67.82	**68.63**
50%	68.37	68.90	70.08	**70.73**
75%	69.97	70.86	70.47	**71.70**



(a) Cosine similarity of negative (left) and positive (right) pairs for teacher and student models trained by $\mathcal{L}_{ce}$

(b) Cosine similarity of negative (left) and positive (right) pairs for student models trained by CRD or PACKD

Figure 5: Visualization of similarity score distribution.

learning the abundant information across intra-class samples well compensates for the lack of data.

Visualization. To illustrate how the proposed scheme improves the intra-class correlation, we further visualize the cosine similarity distributions of intra-class and inter-class sample pairs obtained by different methods with the teacher-student architecture pair ResNet32x4-ResNet8x4 on CIFAR-100. Comparing the teacher and vanilla student models (Fig. 5(a)), we see that there is a perceptible difference between their similarity distributions, for both positive and negative pairs. Meanwhile, in Fig. 5(b), it is illustrated that compared with CRD, PACKD not only decreases the negative-pair similarity but also makes a further improvement on the similarity distribution of positive pairs. This again validates the effectiveness of the proposed positive pair aware contrastive distillation scheme.

5 CONCLUSION

Current distillation methods often lead to degradation over positive similarity. Motivated by this, in this paper we propose a contrastive distillation framework named PACKD, which incorporates sufficient positive pairs to extract more discriminative knowledge from the teacher. By adjusting the weight of different positive pairs based on optimal transport, our method surpasses previous CRD and SSKD methods on several classification benchmarks.

We believe our novel view will facilitate the further design of knowledge transfer methods based on contrastive learning. In the future, we plan to extend our method to other multimedia tasks like video understanding and cross-modal retrieval.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102000, in part by National Natural Science Foundation of China: U21B2038, 61931008,6212200758, 61976202, in part by the Fundamental Research Funds for the Central Universities, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000.

REFERENCES

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9163–9171.

[2] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. 2021. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16296–16305.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[4] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 215–223.

[5] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013), 2292–2300.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[7] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[8] Song Han, Jeff Pool, John Tran, and William J Dally. 2015. Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626* (2015).

[9] Zhiwei Hao, Yong Luo, Han Hu, Jianping An, and Yonggang Wen. 2021. Data-Free Ensemble Knowledge Distillation for Privacy-conscious Multimedia Model Compression. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1803–1811.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3779–3787.

[12] G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the Knowledge in a Neural Network. *Computer Science* 14, 7 (2015), 38–39.

[13] Yi Huang, Xiaoshan Yang, and Changsheng Xu. 2021. Multimodal Global Relation Knowledge Distillation for Egocentric Action Anticipation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 245–254.

[14] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[15] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. 2021. From Superficial to Deep: Language Bias driven Curriculum Learning for Visual Question Answering. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3370–3379.

[16] Xiaoqing Liang, Xu Zhao, Chaoyang Zhao, Nanfei Jiang, Ming Tang, and Jinqiao Wang. 2020. Task Decoupled Knowledge Distillation For Lightweight Face Detectors. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2184–2192.

[17] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. 2018. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proceedings of the 26th ACM international conference on Multimedia*. 700–708.

[18] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, and H. Ghasemzadeh. 2020. Improved Knowledge Distillation via Teacher Assistant. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 4 (2020), 5191–5198.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[20] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3967–3976.

[21] Nikolaos Passalis and Anastasios Tefas. 2018. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 268–284.

[22] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. 2019. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5007–5016.

[23] M. Phuong and C. H. Lampert. 2021. Towards Understanding Knowledge Distillation. (2021).

[24] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[27] Kihyuk Sohn. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf

[28] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does Knowledge Distillation Really Work?. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6906–6919.

[29] Jialin Tian, Xing Xu, Zheng Wang, Fumin Shen, and Xin Liu. 2021. Relationship-Preserving Knowledge Distillation for Zero-Shot Sketch Based Image Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5473–5481.

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699* (2019).

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[32] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1365–1374.

[33] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. 2020. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*. Springer, 588–604.

[34] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2021. Hierarchical Self-supervised Augmented Knowledge Distillation. *arXiv preprint arXiv:2107.13715* (2021).

[35] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2022. Mutual contrastive learning for visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3045–3053.

[36] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* (2016).

[37] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[38] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6848–6856.

[39] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. 2017. Deep Mutual Learning. (2017).

[40] L. Zhen, P. Hu, X. Wang, and D. Peng. 2020. Deep Supervised Cross-Modal Retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[41] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. 2021. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9260–9269.