Dual-Level Knowledge Distillation via Knowledge Alignment and Correlation

Fei Ding, Yin Yang, *Member, IEEE*, Hongxin Hu, *Member, IEEE*, Venkat Krovi, *Senior Member, IEEE*, and Feng Luo, *Senior Member, IEEE*

***Abstract*— Knowledge distillation (KD) has become a widely used technique for model compression and knowledge transfer. We find that the standard KD method performs the knowledge alignment on an individual sample indirectly via class prototypes and neglects the structural knowledge between different samples, namely, knowledge correlation. Although recent contrastive learning-based distillation methods can be decomposed into knowledge alignment and correlation, their correlation objectives undesirably push apart representations of samples from the same class, leading to inferior distillation results. To improve the distillation performance, in this work, we propose a novel knowledge correlation objective and introduce the dual-level knowledge distillation (DLKD), which explicitly combines knowledge alignment and correlation together instead of using one single contrastive objective. We show that both knowledge alignment and correlation are necessary to improve the distillation performance. In particular, knowledge correlation can serve as an effective regularization to learn generalized representations. The proposed DLKD is task-agnostic and model-agnostic, and enables effective knowledge transfer from supervised or self-supervised pretrained teachers to students. Experiments show that DLKD outperforms other state-of-the-art methods on a large number of experimental settings including: 1) pretraining strategies; 2) network architectures; 3) datasets; and 4) tasks.**

***Index Terms*— Convolutional neural networks, dual-level knowledge, knowledge distillation (KD), representation learning, teacher–student model.**

I. Introduction

DEEP neural networks have recently achieved remarkable success in computer vision [1] and natural language processing [2]. However, they usually require high computation and memory demand, which limits their deployment in practical applications. Knowledge distillation (KD) provides a promising solution to build lightweight models by transferring knowledge from high-capacity teachers to smaller students [3], [4]. There are two key points when performing KD: distillation location and objective. The standard KD method [4] minimizes the Kullback–Leibler (KL)-divergence objective between the probabilistic outputs (final logits) of teacher and student networks. The logit distillation actually transfers the dark knowledge, i.e., the relative probabilities assigned to incorrect classes. Recently, contrastive representation distillation (CRD) [5] has achieved superior results on various tasks by using the contrastive objective to transfer knowledge on feature representation (penultimate layer) instead of logits. The main difference between KD and CRD lies in the distillation location and objective, but it remains unclear whether these two methods share common functionalities and whether they can complement each other.

To uncover the relationships between existing distillation methods, we reformulate the standard KD and CRD objectives and identify distillation methods as knowledge alignment or knowledge correlation according to whether the transferred knowledge comes from an individual sample or across samples. We find that standard KD indirectly performs knowledge alignment through the class prototypes, while CRD applies a distillation objective similar to self-supervised contrastive loss [6]–[8] which can be decomposed into knowledge alignment and correlation. Therefore, both KD and CRD include the knowledge alignment objective and CRD has an extra correlation objective. However, we find that the knowledge correlation objective of CRD aims to distribute the negative samples (samples from different instances) more uniformly, which undesirably pushes apart samples from the same class and results in inferior distillation performance. Thus, it is necessary to propose a novel knowledge correlation objective. Besides, the standard KD method relies too much on specific pretraining strategies and network architectures, which requires a more general distillation solution to effectively combine knowledge alignment and correlation together.

In this work, we extract the common part of the existing distillation methods and propose a $L2$-based knowledge alignment objective. We find that a spindle-shaped transformation plays a pivotal role in knowledge alignment. Then, we introduce an effective knowledge correlation objective to capture structural knowledge of the teacher. Both of our alignment and correlation objectives focus on the feature representation. Therefore, our method is independent of the specific pretraining tasks or architectures, which provides a more flexible KD. We demonstrate that knowledge alignment and correlation are necessary to improve the distillation performance. In particular, knowledge correlation can serve as an effective regularization to enable the student to learn generalized representations. We identify the proposed method

Manuscript received 30 August 2021; revised 4 May 2022; accepted 7 July 2022. Date of publication 14 July 2022; date of current version 6 February 2024. The work of Feng Luo was supported in part by the U.S. National Science Foundation (NSF) under Grant ABI-1759856, Grant MRI-2018069, and Grant MTM2-2025541. *(Corresponding author: Feng Luo.)*

Fei Ding, Yin Yang, and Feng Luo are with the School of Computing, Clemson University, Clemson, SC 29634 USA (e-mail: feid@clemson.edu; yin5@clemson.edu; luofeng@clemson.edu).

Hongxin Hu is with the Department of Computer Science and Engineering, University at Buffalo The State University of New York, Buffalo, NY 14260 USA (e-mail: hongxinh@buffalo.edu).

Venkat Krovi is with the Department of Automotive Engineering and the Department of Mechanical Engineering, Clemson University, Clemson, SC 29634 USA (e-mail: vkrovi@clemson.edu).

Digital Object Identifier 10.1109/TNNLS.2022.3190166

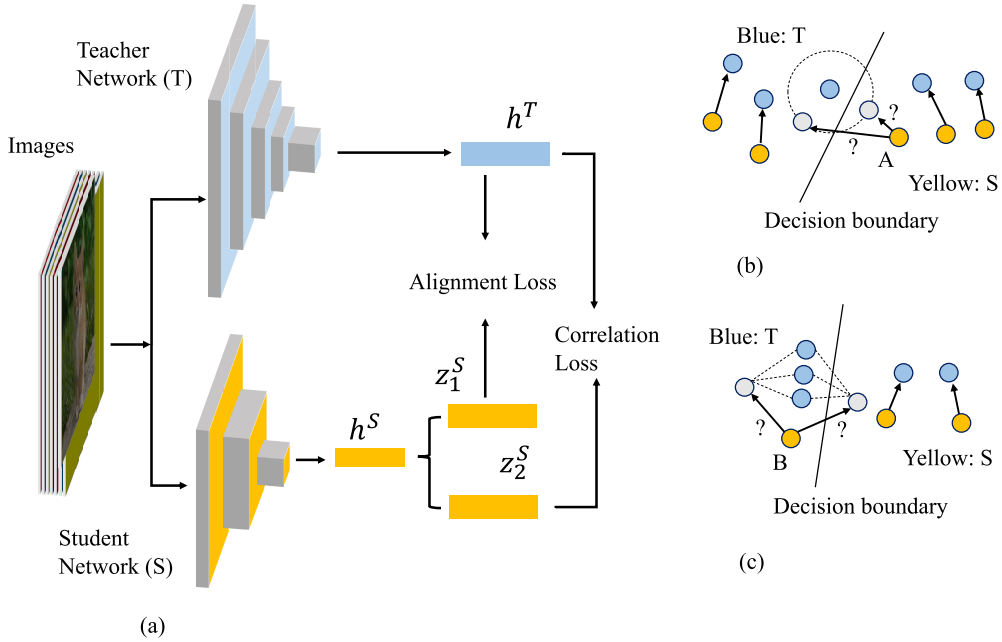


Fig. 1. Overview of knowledge alignment and correlation. (a) Our distillation framework: $\mathbf{h}^T$ and $\mathbf{h}^S$ indicate representations of the teacher and student. $\mathbf{z}_1^S$ and $\mathbf{z}_2^S$ are two different transformations for distillation. (b) Knowledge alignment focuses on direct feature matching, and (c) knowledge correlation captures relative relationship between samples. The blue (the teacher) and yellow (the student) circles represent different samples. ? indicates that A and B samples could be mapped to different locations (gray circles). Given the decision boundary, different mappings lead to different classification results. The dotted circle in (b) indicates possible feature alignment results and dotted lines in (c) indicate that two different mappings share the same relationship between samples. (b) and (c) Illustrate the necessity of knowledge alignment and correlation. It could not achieve the optimal distillation via one single objective.

as dual-level KD (DLKD) (DLKD) to emphasize that it effectively combines both knowledge alignment and correlation, as shown in Fig. 1. Besides, we introduce an optional supervised distillation objective by leveraging the labels, which can indirectly transfer the category-wise structural knowledge between networks. To summarize, our main contributions are as follows.

1) We introduce a novel KD method, DLKD (DLKD), which provides a general and model-agnostic solution to transfer richer representational knowledge between networks.
2) We define a general knowledge quantification metric to measure and evaluate the consistency of visual concepts in the learned representation.
3) We show that knowledge alignment and correlation can provide effective supervisory signals for KD, and allow students to learn more generalized representations.
4) We demonstrate that DLKD consistently outperforms state-of-the-art methods over a large set of experiments including different pretraining strategies (supervised and self-supervised), network architectures (vgg, ResNets, WideResNets, MobileNets, and ShuffleNets), datasets (CIFAR-10/100, STL10, ImageNet, and Cityscapes) and tasks (classification, segmentation, and self-supervised learning).

II. Related Work

A. Knowledge Distillation

Hinton *et al.* [4] first propose KD to transfer dark knowledge from the teacher to the student. The softmax outputs encode richer knowledge than one-hot labels and can provide extra supervisory signals. Softmax regression representation learning (SRRL) [9] performs KD by leveraging the teacher's projection matrix to train the student's representation via $L2$ loss. However, these works rely on a supervised pretrained teacher (with logits), and they may not be suitable for self-supervised pretrained teachers. Self-supervision knowledge distillation (SSKD) [10] is proposed to combine the self-supervised auxiliary task and KD to transfer richer dark knowledge, but it cannot be trained in an end-to-end training way. Similar to logits matching, intermediate representation [11]–[15] are widely used for KD. FitNet [11] proposes to match the whole feature maps, which is difficult and may affect the convergence of the student in some cases. Attention transfer [12] utilizes spatial attention maps as the supervisory signal. In flow-based distillation [13], interlayer flow matrices of the teacher are computed to guide the learning of the student. Activation boundaries (AB) [15] proposes to learn the activation boundaries of the hidden neurons in the teacher. Similarity-preserving (SP) [14] focuses on transferring the similar (dissimilar) activations between the teacher and student. However, most of these works depend on certain architectures, such as convolutional networks. Since these distillation methods involve knowledge matching in an individual sample, they are related to knowledge alignment. Our work also includes the knowledge alignment objective, but does not rely on pretraining strategies or network architectures.

B. Knowledge Alignment and Self-Supervised Learning

Self-supervised learning [6]–[8], [16], [17] focuses on learning low-dimensional representations by the instance discrimination, which usually requires a large number of negative samples. Recently, Bootstrap Your Own Latent (BYOL) [18] and self-DIstillation with NO labels (DINO) [19] utilize the momentum encoder to avoid collapse without negatives. The momentum encoder can be considered as the mean teacher [20], which is built dynamically during the stu-

dent training. For distillation, the teacher is pretrained and fixed during distillation. Although different views (augmented images) are passed through networks in self-supervised learning, they are from the same original sample for feature alignment. These self-supervised methods perform knowledge alignment between the student and the momentum teacher during each iteration. In particular, DINO focuses on local-to-global knowledge alignment based on multicrop augmentation.

C. Relational KD

Besides knowledge alignment, another research line of KD focuses on transferring relationships between samples. DarkRank [21] utilizes cross-sample similarities to transfer knowledge for metric learning tasks. Also, relational knowledge distillation (RKD) [22] transfers distance-wise and angle-wise relations of different feature representations. Recently, CRD [5] is proposed to apply contrastive objective for structural KD. However, it randomly draws negative samples and inevitably selects false negatives, hence leading to a suboptimal solution. SElf-SupErvised Distillation (SEED) [23] is another contrastive distillation method to transfer relational knowledge between different samples from a self-supervised pretrained teacher. It only considers knowledge correlation between the sample and a queue. But due to the use of a large queue, it cannot effectively transfer knowledge between different semantic samples. Our work proposes an effective knowledge correlation objective.

III. Dual-Level KD

A. Reformulating KD and CRD

Given a pair of teacher and student networks, $f_\eta^T(\cdot)$ and $f_\theta^S(\cdot)$, the distillation methods train the student via extra supervisory signals from the supervised or self-supervised pretrained teacher. $f_\eta^T(\cdot)$ and $\mathbf{h}^T$ denote the feature extractor and representation vector of the teacher. Take the supervised teacher as an example, besides $f_\eta^T(\cdot)$, there is also a projection matrix $\mathbf{W}^T \in \mathbb{R}^{D\times K}$ to map the feature representation to K category logits, where D is the feature dimensionality. We denote by $s(\cdot)$ the softmax function, and the standard KD loss [4] can be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} &= -\sum_{k=1}^{K} s\left(\mathbf{W}_k^T\mathbf{h}^T\right)\log s\left(\mathbf{W}_k^S\mathbf{h}^S\right)\\ &= -\sum_{k=1}^{K} s\left(\mathbf{W}_k^T\mathbf{h}^T\right)\left[\log s\left(\mathbf{W}_k^S\mathbf{h}^S\right)+\log s\left(\mathbf{W}_k^T h_\varphi\left(\mathbf{h}^S\right)\right)\right.\\ &\qquad\left.-\log s\left(\mathbf{W}_k^T h_\varphi\left(\mathbf{h}^S\right)\right)\right]\\ &= -\sum_{k=1}^{K} s\left(\mathbf{W}_k^T\mathbf{h}^T\right)\log s\left(\mathbf{W}_k^T h_\varphi\left(\mathbf{h}^S\right)\right)\\ &\quad+\sum_{k=1}^{K} s\left(\mathbf{W}_k^T\mathbf{h}^T\right)\log\frac{s\left(\mathbf{W}_k^T h_\varphi\left(\mathbf{h}^S\right)\right)}{s\left(\mathbf{W}_k^S\mathbf{h}^S\right)}\end{aligned}\tag{1}$$

where $h_\varphi(\cdot)$, $\mathbf{h}^S$, and $\mathbf{W}_k^S$ are trainable, and $\mathbf{h}^T$ and $\mathbf{W}_k^T$ are frozen. $h_\varphi(\cdot)$ represents a feature transformation function of aligning the student's representation to the teacher's representation. We observe that when $\mathbf{h}^T = h_\varphi(\mathbf{h}^S)$, the first loss item achieves the optimal solution, and the second loss item becomes the KL divergence between softmax distributions. In other words, the standard KD objective is related to knowledge alignment, and can minimize the discrepancy between networks' outputs indirectly through the class prototypes $\mathbf{W}^T$ and $\mathbf{W}^S$. Recently, CRD shows that indirect learning of the teacher's knowledge is not sufficiently effective and proposes the contrastive representation distillation. Inspired by Wang and Isola [24], the softmax formulation of CRD's objective can be reformulated into two parts

$$\begin{aligned}\mathcal{L}_{\mathrm{CRD}} &= -\mathbf{z}_i^S\mathbf{z}_i^T/\tau\\ &\quad+\log\left(\exp\left(\mathbf{z}_i^S\mathbf{z}_i^T/\tau\right)+\sum_{j=1}^{N}\exp\left(\mathbf{z}_i^S\mathbf{z}_j^T/\tau\right)\right)\end{aligned}\tag{2}$$

where $\mathbf{z}_i^S$ and $\mathbf{z}_i^T$ are the positive representation pair of the teacher (T) and student (S) from the sample x_i. τ is the temperature parameter, N indicates the total number of negative samples, and j indicates the jth($j \neq i$) negative sample of $\mathbf{z}_i^S$. Intuitively, the first term encourages the outputs of the teacher and student for the same sample to be similar (alignment), while the second term encourages representations of samples from negatives to be more dissimilar (correlation). However, because negative samples usually are randomly chosen as long as they are different from x_i, the second term causes many negative samples from the same class (false negatives) be undesirably pushed apart in the representation space.

The distinction of knowledge alignment and correlation provides a novel viewpoint to analyze different distillation methods by reformulating their objectives. From the above analysis, we find that both KD and CRD contain the knowledge alignment objective. We also find that although CRD considers transferring the relationship between samples, it is not optimal due to the problem of false negatives. Here, we propose a novel knowledge correlation objective to capture structural knowledge of samples. And we apply two independent objectives to perform knowledge alignment and correlation, respectively. Both of the proposed objectives are calculated at the feature level, which allows our method to be extended to new pretraining strategies and architectures.

B. Knowledge Alignment

A well-trained teacher already encodes excellent representational knowledge, i.e., categorical knowledge. The stronger supervision is necessary for better matching between the teacher's representation ($f_\eta^T(x)$) and the transformation of the student's representation ($h_\varphi(f_\theta^S(x))$). To meet the requirement of knowledge alignment ($\mathbf{h}^T = h_\varphi(\mathbf{h}^S)$), we propose an $L2$-based knowledge alignment objective

$$\mathcal{L}_{\mathrm{Align}} = \mathbb{E}_x\left[\left\|h_\varphi\left(f_\theta^S(x)\right)-f_\eta^T(x)\right\|_2^2\right].\tag{3}$$

This objective forces the student to directly mimic the teacher's representation, thus can provide stronger supervisory signals of interclass similarities than the standard KD loss [4]. Equation (3) applies the feature representation (penultimate layer) to perform knowledge alignment. Our method is better than previous FitNet loss which matches whole feature

maps and may cause training to become difficult or even fail when $h_\varphi(\cdot)$ is only regarded as dimensionality matching. In Section VI, we confirm that appropriate representation capability of $h_\varphi(\cdot)$ plays a key role in knowledge alignment.

The knowledge alignment can be further expressed as

$$\mathcal{L}_{\varphi,\theta} = \mathbb{E}_x\left[l\left(h_\varphi\left(f_\theta^S(x)\right), g_\phi\left(f_\eta^T(x)\right)\right)\right] \tag{4}$$

where $l(\cdot,\cdot)$ loss function is used to penalize the difference between networks in different outputs. This is a generalization of existing KD objectives [4], [9], [11]–[13]. For example, Hinton *et al.* [4] calculate KL-divergence between f^T and f^S in which the linear functions h_φ and g_ϕ map representations to logits. SRRL [9] utilizes the teacher's pretrained projection matrix W^T to enforce the teacher's and student's feature to produce the same logits via the $L2$ loss. These methods rely on the logits of the classification task. In contrast, our method is task-agnostic. Although knowledge alignment is the common part of the existing distillation methods, it does not ensure that the teacher's knowledge is fully transferred, as it neglects the structural knowledge between different samples.

C. Knowledge Correlation

The pretrained teacher also encodes the knowledge of rich relationships between samples, and knowledge correlation allows the student to learn a structure of the representation space similar to the teacher. Here, we propose a novel knowledge correlation objective to capture structural knowledge from the teacher. To be specific, we calculate the relational scores for each $(N+1)$-tuple samples as the cross-sample relational knowledge. The correlation objective can be expressed as

$$\mathcal{L}_{\text{Corr}} = \sum_{i=1}^{N} l\Big(\psi\left(f_\eta^T(\tilde{x}_i), f_\eta^T(x_1), \ldots, f_\eta^T(x_N)\right), \psi\left(f^S(\tilde{x}_i), f^S(x_1), \ldots, f^S(x_N)\right)\Big) \tag{5}$$

where N is the batch size, ψ is the relational function that measures the relational scores between the augmented $\tilde{x}_i$ and samples $\{x_i\}_{i=1:N}$. $l(\cdot,\cdot)$ is a loss function. The samples in each batch have different semantic similarities, and ψ needs to assign higher scores to samples with similar semantic meaning and lower relational scores otherwise. Here, we apply the cosine similarity to measure the semantic similarity between representations, and transform them to softmax distribution for the knowledge correlation objective. All similarities between $\{\tilde{x}_i\}_{i=1:N}$ and $\{x_i\}_{i=1:N}$ can be written as matrix $\mathcal{A}$. For the teacher network, $\mathcal{A}_{i,j}$ is calculated by the representations $\mathbf{h}^S$. For the student network, we also apply a transformation function to the representation $\mathbf{z}^S$ for loss calculation.

We apply the softmax function as the relational function ψ and KL-divergence loss as $l(\cdot,\cdot)$ to transfer these relationships from the teacher to the student

$$\mathcal{L}_{\text{Corr}} = \sum_{i}^{N}\sum_{j}^{N} -\frac{\exp\left(\mathcal{A}_{i,j}/\tau\right)}{\sum_j \exp\left(\mathcal{A}_{i,j}/\tau\right)} \cdot \log \frac{\exp\left(\mathcal{A}_{i,j}/\tau\right)}{\sum_j \exp\left(\mathcal{A}_{i,j}/\tau\right)} \tag{6}$$

where τ is the temperature parameter to soften peaky distributions and $f(\cdot)$ is the teacher or student network.

We also compare our knowledge correlation objective with other relational distillation objectives. RKD [22] proposes distance-wise and angle-wise losses for relational KD. The former has a significant difference in scales and makes training unstable. The latter utilizes a triplet of samples to calculate angular scores $\mathrm{O}(N^3)$ complexity. Our KL-based solution achieves high-order property with $\mathrm{O}(N^2)$ complexity. SEED [23] is proposed to transfer knowledge from a self-supervised pretrained teacher by leveraging similarity scores between a sample and a queue. However, the large queue results in sparse softmax outputs due to lots of dissimilar samples, which makes it not effective to transfer knowledge between different semantic samples. We directly calculate mutual relationships in each batch and utilize KL divergence loss, which does not require additional queue and large-size batch, thus has high computational efficiency.

D. Supervised KD

Both above objectives are related to feature representations and therefore independent of specific pretraining tasks. Here, we also propose an additional distillation objective for supervised pretrained teachers based on the InfoNCE loss. We overcome the false negative problem in CRD by leveraging the true labels to construct positives from the same category and negatives from different categories. There are two kinds of anchors (teacher and student anchor) in distillation

$$\mathcal{L}_{\text{Sup}}^{T/S} = -\frac{1}{C}\sum_{i=1}^{N}\sum_{j=1}^{2N} \mathbb{1}_{i\neq j} \cdot \mathbb{1}_{\mathbf{y}_i=\mathbf{y}_j} \cdot \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_j/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i\neq k} \cdot \exp(\mathbf{z}_i \cdot \mathbf{z}_k/\tau)} \tag{7}$$

where $C = 2N_{y_i} - 1$ and N_{y_i} is the number of images with the label y_i in the minibatch. The feature vectors $\mathbf{z}$ are transformed from $\mathbf{h}^T$ or $\mathbf{h}^S$ via multilayered perceptron (MLP) heads. $\mathbf{z}_i$ is the anchor representation of the teacher or student. $\mathbf{z}_j$ and $\mathbf{z}_k$ represent positive and negative features, respectively. When $\mathbf{z}_i$ is from the teacher, $\mathbf{z}_j$ and $\mathbf{z}_k$ are from the student, vice versa. This objective provides categorical similarities to encourage a student to map samples from the same category into close representation space and samples from different categories be far away. Our formulation is similar to the supervised contrastive loss [25], with the difference that our objective requires fixed anchors for knowledge transfer.

E. DLKD Objective

The total distillation objective for any pretraining teacher is a linear combination of knowledge alignment and correlation objectives

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{Align}} + \lambda_2\mathcal{L}_{\text{Corr}} \tag{8}$$

where λ_1 and λ_2 are balancing weights. For the supervised pretrained teacher, we also add the above supervised distillation loss $\mathcal{L}_{\text{Sup}}$ and the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ with different balancing weights. This objective forces a student

network to learn multiple facets of representational knowledge from a teacher, as shown in Fig. 1.

IV. Knowledge Quantification Metric

To evaluate the distillation performance, it is necessary to understand the representation knowledge by quantifying the knowledge encoded in networks. Cheng *et al.* [26] proposed to quantify the visual concepts of networks on foreground and background, which requires annotations of the object bounding box. However, these kinds of ground-truth bounding boxes are not always available. Here, we define more general metrics to explain and analyze the knowledge encoded in networks based on the conditional entropy.

Let $\mathbf{X}$ denote a set of input images. The conditional entropy $H(\mathbf{X}|\mathbf{z} = f(x))$ measures how much information from the input image x to the representation $\mathbf{z}$ is discarded during the forward propagation [26], [27]. A perturbation-based method [27] is proposed to approximate $H(\mathbf{X}|\mathbf{z})$. The perturbed input $\tilde{x}$ follows Gaussian distribution with the assumption of independence between pixels, $\tilde{x} \sim \mathcal{N}(x, \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2))$, where n denotes the total number of pixels. Therefore, the image-level conditional entropy $H(\mathbf{X}|\mathbf{z})$ can be decomposed into pixel-level entropy H_i ($H(\mathbf{X}|\mathbf{z}) = \sum_{i=1}^{n} H_i$), where $H_i = \log \sigma_i + (1/2)\log(2\pi e)$. High pixel-wise entropy H_i indicates that more information is discarded through layers. The pixels with low pixel-wise entropy are more related with the representation, thus the low-entropy pixels can be considered reliable visual concepts.

We define two general quantification metrics from the view of knowledge quantification and consistency: average and Intersection over Union (IoU). The average entropy $\bar{H} = (1/n)\sum_i H_i$ of the image indicates how much information is discarded in the whole input. A smaller $\bar{H}$ indicates that the network utilizes more pixels to compute feature representation from the input. However, more visual concepts do not always lead to the optimal feature representation, which might result in the over-fitting issue [28]. Ideally, a well-learned network is supposed to encode more robust and reliable knowledge. Thus, we measure the knowledge consistency by the IoU metric, which quantifies the consistency of visual concepts between two views of the same image, i.e., two augmented images x_1 and x_2

$$\text{IoU} = \mathbb{E}_{x\in\mathcal{X}}\left[\frac{\sum_{i\in x_1\cap x_2}\left(S^1_{\text{concept}}(x_i)\cap S^2_{\text{concept}}(x_i)\right)}{\sum_{i\in x_1\cap x_2}\left(S^1_{\text{concept}}(x_i)\cup S^2_{\text{concept}}(x_i)\right)}\right] \quad \text{where } S_{\text{concept}}(x) = \mathbb{1}\left(\bar{H} > H_i\right) \tag{9}$$

where $\mathbb{1}$ is the indicator function, and $S_{\text{concept}}(x)$ denotes the set of visual concepts (pixels with lower entropy than $\bar{H}$). $i \in x_1 \cap x_2$ denotes the same pixels of two augmented images. These same pixels are supposed to obtain similar visual concepts and keep a good consistency between augmented images. We choose the ratio between number of visual concepts overlap and number of visual concepts union (IoU) to measure the knowledge consistency of the learned representations. Our IoU metric meets the requirements of generality and coherency [26], and can be used to quantify and analyze the visual concepts without relying on specific architectures, tasks, or datasets.

V. DLKD and Mutual Information Bound

Considering the representations of teacher and student in terms of T and S ($T = f_\eta^T(x)$, $S = f_\theta^S(x)$), we define a distribution q with binary variable C to denote whether a pair of representations $(f_\eta^T(x_i), f_\theta^S(x_j))$ is drawn from the joint distribution $p(T, S)$ or the product of marginals $p(T)p(S)$: $q(T, S|C = 1) = p(T, S)$, $q(T, S|C = 0) = p(T)p(S)$. The joint distribution indicates positive pairs from close representation space, and the product of marginals indicates negative pairs from far representation space. CRD only considers the same input provided to $f_\eta^T(\cdot)$ and $f_\theta^S(\cdot)$ as the positives, and samples drawn randomly from the training data as the negatives, which leads to sampling bias problem [29].

Given N_p positive samples and N_n negative samples, we consider the positives in T and S from $p(T, S)$ are empirically related and semantically similar, e.g., representations of the same sample, augmented sample, and samples from the same category, and the negative samples are drawn empirically from different categories. The contrastive-based distillation methods aim to encourage student's representations to be close to teacher's representations in positives, and those of negatives to be more orthogonal. Then, the priors can be written as: $q(C = 1) = N_p/(N_p + N_n)$ and $q(C = 0) = N_n/(N_p + N_n)$. According to the Bayes' rule, the posterior $q(C = 1|T, S)$ can be written as

$$q(C = 1|T, S) = \frac{p(T, S)}{p(T, S) + p(T)p(S)(N_n/N_p)} \tag{10}$$

$$\log q(C = 1|T, S) = -\log\left(1 + (N_n/N_p)\frac{p(T)p(S)}{p(T, S)}\right) \leq -\log(N_n/N_p) + \log\frac{p(T, S)}{p(T)p(S)}. \tag{11}$$

Taking expectation over both sides with respect to $q(T, S|C = 1)$, we have the mutual information bound as follows:

$$I(T; S) \geq \log(N_n/N_p) + \mathbb{E}_{q(T,S|C=1)}\log q(C = 1|T, S). \tag{12}$$

The first term $\log(N_n/N_p)$ is constant for the given dataset. Previous studies [5] suggest that a larger batch size can obtain a better lower bound. But our analysis indicates that the influence factor is the ratio of negative and positive samples, which depends on the training data. The second term is to maximize the expectation with respect to the student parameters to increase the lower found. But the true distribution $q(C = 1|T, S)$ is intractable. We note that this equation is similar to the InfoNCE loss [6], which provides a tractable estimator.

When the teacher's representation $\mathbf{z}_i^T$ and the student's representation $\mathbf{z}_i^S$ form a positive pair, we can relate our knowledge alignment objective to the dot product of positive samples in the InfoNCE through (13), where we maximize the similarity of teacher and student's representations via knowledge alignment

$$\mathcal{L}_{\text{Align}} = -\frac{\mathbf{z}_i^S \cdot \mathbf{z}_i^T}{\|\mathbf{z}_i^S\| \cdot \|\mathbf{z}_i^T\|} = \frac{1}{2}\cdot\|\mathbf{z}_i^S - \mathbf{z}_i^T\|_2^2 - 1. \tag{13}$$

For the knowledge correlation objective, it does not directly align representations between networks. Instead, it considers the relationship between an anchor $\mathbf{z}_i^T$ and the jth sample $\mathbf{z}_j^T$ in the teacher by the softmax function

$$\psi\left(\mathbf{z}_i^T, \mathbf{z}_j^T\right) = \frac{\exp\left(\mathbf{z}_i^T \mathbf{z}_j^T / \tau\right)}{\sum_{k=1}^{N} \exp\left(\mathbf{z}_i^T \mathbf{z}_k^T / \tau\right)}. \tag{14}$$

In practice, we convert the relationships between all samples in the batch to the softmax distribution. Then we apply KL-divergence loss to transfer the relationships from the teacher to the student. Because the teacher already encodes the relational knowledge between samples, our knowledge correlation objective encourages the student to learn the similar relationships between samples. Thus, it enables the student to map samples from the same category to be closer, and indirectly models the binary classification problem, which is related to $q(C = 1|T, S)$. Because the objectives for knowledge alignment and correlation do not rely on an explicit definition of positives/negatives, it is applicable in supervised/self-supervised pretrained teachers.

VI. Experiments

In this section, we first compare our method with state-of-the-art methods in the KD tasks (supervised, structured, and self-supervised KD). Then we conduct an ablation study to verify each loss of DLKD via classification accuracy and knowledge quantification metric. We also perform experiments to evaluate the transferability of representations and the performance under a few-shot scenario.

A. Experimental Settings

1) Network Architectures: We adopt vgg [32] ResNet [33], WideResNet [34], MobileNet [35], and ShuffleNet [36] as teacher–student combinations to evaluate the supervised KD on CIFAR100 dataset [37] and ImageNet dataset [38]. Their implementations are from [5]. For structured KD, we implement DLKD based on [39] and evaluate it on Cityscapes dataset [40]. The teacher model is the PSPNet architecture [41] with a ResNet101 and the student model is set to ResNet18. For self-supervised KD, the teachers are pretrained via MoCo-V2 [42] or SwAV [17] and we directly download the pretrained weights for our evaluation. The student network is set to smaller ResNet networks (ResNet18, 34). We also perform the transferability evaluation of representations on STL10 dataset [43] and TinyImageNet dataset [38], [44].

2) Implementation Details: Our implementation is mainly to verify the effectiveness of DLKD. We follow the same training strategy based on the existing solutions without any tricks. For supervised KD, we use the stochastic gradient descent (SGD) optimizer with the momentum of 0.9 and the weight decay of 5×10^{-4} in CIFAR100. All the students are trained for 240 epochs with a batch size of 64. The initial learning rate is 0.05 and then divided by ten at the 150th, 180th, and 210th epochs. In ImageNet, we follow the official implementation of PyTorch[1] and adopt the SGD optimizer with a 0.9 momentum and 1×10^{-4} weight decay.

[1]https://github.com/pytorch/examples/tree/master/imagenet

TABLE I

Distillation Performance Comparison Between Similar Architectures. It Reports Top-1 Accuracy (%) on CIFAR100 Test Dataset. We Denote the Best and the Second-Best Results by **Bold** and Underline. The Results of All Compared Methods are From [10]. DLKD Consistently Achieves the Best Results on All Comparisons

Teacher	wrn40-2	wrn40-2	resnet56	resnet32×4	vgg13
Student	wrn16-2	wrn40-1	resnet20	resnet8×4	vgg8
Teacher	76.46	76.46	73.44	79.63	75.38
Student	73.64	72.24	69.63	72.51	70.68
KD [4]	74.92	73.54	70.66	73.33	72.98
Fitnets [11]	75.75	74.12	71.60	74.31	73.54
AT [12]	75.28	74.45	<u>71.78</u>	74.26	73.62
FT [30]	75.15	74.37	71.52	75.02	73.42
SP [14]	75.34	73.15	71.48	74.74	73.44
VID [31]	74.79	74.20	71.71	74.82	73.96
RKD [22]	75.40	73.87	71.48	74.47	73.72
AB [15]	68.89	75.06	71.49	74.45	74.27
CRD [5]	<u>76.04</u>	75.52	71.68	75.90	74.06
SSKD [10]	<u>76.04</u>	<u>76.13</u>	71.49	<u>76.20</u>	<u>75.33</u>
DLKD(ours)	**77.20**	**76.74**	**72.34**	**77.11**	**75.40**

The initial learning rate is 0.1 and is decayed by ten at the 30th, 60th, and 90th epoch in a total of 100 epochs. For these two datasets, we apply normal data augmentation methods, such as rotation with four angles, i.e., $0°, 90°, 180°, 270°$. To perform structured KD, the student is trained with an SGD optimizer with the momentum of 0.9 and the weight decay of 5×10^{-4} for 40 000 iterations. The training input is set to 512×512, and normal data augmentation methods, such as random scaling and flipping, are used during the training. The self-supervised KD is trained by an SGD optimizer with the momentum of 0.9 and the weight decay of 1×10^{-4} for 200 epochs. More detailed training information can be found in the compared methods (CRD [5], SKD [39] and SEED [23]). The temperature τ in $\mathcal{L}_{\text{Corr}}$ and $\mathcal{L}_{\text{Sup}}$ is set to be 0.5 and 0.07. For the balancing weights, we set $\lambda_1 = 10$ and $\lambda_2 = 20$ according to the magnitude of the loss value. During supervised KD, we set the weights of $\mathcal{L}_{\text{Sup}}$ and $\mathcal{L}_{\text{CE}}$ loss to be 0.5 and 1.0. All models are trained using Tesla V100 graphics processing units (GPUs) on an NVIDIA DGX2 server.

B. Supervised KD

1) CIFAR100: DLKD is compared with the existing distillation methods, as shown in Tables I and II. Following CRD [5] and SSKD [10], Tables I and II compare teacher–student pairs with similar and different architectures. Our method achieves a large improvement compared with KD and CRD methods, which validates the effectiveness of combination of knowledge alignment and correlation. SSKD is an improved KD method combined with contrast learning, yet only applicable to supervised pretrained teachers for classification tasks, and is more complex which requires two steps. In contrast, our method is simpler, meanwhile still achieve better distillation results and can be applied to supervised and self-supervised pretrained teachers. For similar-architecture comparisons, DLKD increases the performance of the students by an average of 0.66% compared to the other best methods. Taking the teacher resnet32 × 4 as an example, two different

TABLE II

DISTILLATION PERFORMANCE COMPARISON BETWEEN DIFFERENT ARCHITECTURES. IT REPORTS TOP-1 ACCURACY (%) ON CIFAR100 TEST DATASET. WE DENOTE THE BEST AND THE SECOND-BEST RESULTS BY **BOLD** AND UNDERLINE. THE RESULTS OF ALL COMPARED METHODS ARE FROM [10]. DLKD CONSISTENTLY ACHIEVES THE BEST RESULTS ON ALL COMPARISONS

Teacher Student	vgg13 MobileV2	ResNet50 MobileV2	ResNet50 vgg8	resnet32×4 ShuffleV1	resnet32×4 ShuffleV2	wrn40-2 ShuffleV1
Teacher	75.38	79.10	79.10	79.63	79.63	76.46
Student	65.79	65.79	70.68	70.77	73.12	70.77
KD [4]	67.37	67.35	73.81	74.07	74.45	74.83
Fitnets [11]	68.58	68.54	73.84	74.82	75.11	75.55
AT [12]	69.34	69.28	73.45	74.76	75.30	75.61
FT [30]	69.19	69.01	73.58	74.31	74.95	75.18
SP [14]	66.89	68.99	73.86	73.80	75.15	75.56
VID [31]	66.91	68.88	73.75	74.28	75.78	75.36
RKD [22]	68.50	68.46	73.73	74.20	75.74	75.45
AB [15]	68.86	69.32	74.20	76.24	75.66	76.58
CRD [5]	68.49	70.32	74.42	75.46	75.72	75.96
SSKD [10]	71.53	72.57	75.76	78.44	78.61	77.40
DLKD(ours)	**72.52**	**73.18**	**76.15**	**78.89**	**79.54**	**78.01**

TABLE III

TOP-1 AND TOP-5 ERROR RATES (%) ON IMAGENET. WE DENOTE THE BEST AND THE SECOND-BEST RESULTS BY **BOLD** AND UNDERLINE

	Teacher	Student	SP	KD	AT	CRD	SSKD	SRRL [9]	DLKD
Top-1	26.70	30.25	29.38	29.34	29.30	28.83	28.38	28.27	**27.88**
Top-5	8.58	10.93	10.20	10.12	10.00	9.87	9.33	9.40	**9.30**

types of student networks resnet8 × 4 and ShuffleV2 achieve 77.11% and 79.54% performance, respectively. This demonstrates that DLKD can break through the architecture-specific limitation to achieve excellent performance. Notably, we find that DLKD enables the student to obtain better performance than the teacher in three out of five pairs. While comparing the teacher–student pairs with different architectures, DLKD also enables the student to learn better than the teacher.

2) ImageNet: We further conduct the experiment (teacher: ResNet34, student: ResNet18) on ImageNet. As shown in Table III, our DLKD achieves the best classification performances for both Top-1 and Top-5 error rates, which demonstrate the efficiency and scalability on the large-scale dataset.

C. Structured KD

Semantic segmentation can be considered as a structured prediction problem, with different levels of similarities among pixels. To transfer the structured knowledge from the teacher to the student, it is also necessary to perform the pixel-level knowledge alignment and correlation in the feature space. The former encourages the student to learn similar feature representations for each pixel from the teacher, even though their receptive fields (convolutional networks) are different. The latter focuses on maintaining the similarity between pixels belonging to the same class, and the dissimilarity of pixels between different classes. SKD [39] proposes to transfer pair-wise similarities among pixels in the feature space. Intra-class Feature Variation Distillation (IFVD) [45] proposes to transfer similarities between each pixel and its corresponding class prototype. In contrast, our distillation method can achieve better distillation results than the existing structured KD methods (see Table IV).

D. Self-Supervised KD

We evaluate the self-supervised distillation with the k-NN nearest neighbor classifier ($k = 10$) as in SEED [23], which does not require any hyperparameter tuning, nor augmentation. Table V shows the distillation results from different teacher–student pairs. The results of all compared methods are from [23]. The first two rows show the supervised training and self-supervised (MoCo-V2) training baseline results.

TABLE IV

SEGMENTATION PERFORMANCE COMPARISON ON CITYSCAPES VAL DATASET. TEACHER: RESNET101 AND STUDENT: RESNET18

Method	val mIoU (%)	Params (M)
Teacher	78.56	70.43
Student	69.10	13.07
SKD [39]	72.70	13.07
IFVD [45]	74.54	13.07
DLKD(ours)	**75.73**	13.07

TABLE V

TOP-1 k-NN CLASSIFICATION ACCURACY(%) ON IMAGENET. + AND ∗INDICATES THE TEACHERS PRETRAINED BY MOCO-V2 AND SWAV

Teacher	ResNet18	ResNet34
Supervised	69.5	72.8
Self-supervised	36.7	41.5
R-50$^+$ + SEED	43.4	45.2
R-101$^+$ + SEED	48.6	50.5
R50x2* + SEED	55.3	58.2
R50x2* + Ours	**56.4**	**59.6**

TABLE VI

IMAGENET TEST ACCURACY(%) USING LINEAR CLASSIFICATION. + AND ∗INDICATE THE TEACHERS PRETRAINED BY MOCO-V2 AND SWAV

Methods		ResNet18		ResNet34	
	Top-1	Top-1	Top-5	Top-1	Top-5
Supervised		69.5		72.8	
Self-supervised		52.5	77.0	57.4	81.6
R-50$^+$ + SEED	67.4	57.9	82.0	58.5	82.6
R50x2* + SEED	77.3	63.0	84.9	65.7	86.8
R50x2* + Ours	77.3	**65.8**	**86.5**	**67.9**	**87.7**

TABLE VII

DISTILLATION PERFORMANCE COMPARISON OF DIFFERENT $h_\varphi(\cdot)$ ON THE RESNET32 × 4 AND SHUFFLEV2. IT REPORTS TOP-1 ACCURACY (%) ON CIFAR100 TEST DATASET. IT DENOTES MULTIPLES OF $\dim(\mathbf{z}_T)$

Hidden size	0.25 ×	0.5 ×	1 ×	2 ×	4 ×	8 ×	16 ×	32 ×	64 ×
Top-1	78.54	78.63	78.58	78.62	78.43	78.57	**79.01**	78.81	78.66

TABLE VIII

ABLATION STUDY OF DLKD. IT REPORTS TOP-1 ACCURACY (%) OF TWO TEACHER–STUDENT PAIRS ON CIFAR100 TEST DATASET

Teacher Student	resnet32×4 resnet8×4	resnet32×4 ShuffleV2
$\mathcal{L}_{\text{Align}}$	76.59	79.01
$\mathcal{L}_{\text{Corr}}$	74.94	76.06
$\mathcal{L}_{\text{Sup}}$	74.73	75.98
$\mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Sup}}$	76.99	79.26
$\mathcal{L}_{\text{Corr}} + \mathcal{L}_{\text{Sup}}$	75.90	77.35
$\mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Corr}}$	76.90	79.17
All	**77.11**	**79.54**

TABLE IX

ABLATION STUDY OF DLKD. TOP-1 ACCURACY (%) OF LINEAR EVALUATION ON TWO DATASETS USING LEARNED REPRESENTATION ON CIFAR100 DATASET (TEACHER: RESNET32 × 4, STUDENT: RESNET8 × 4)

Datset	STL10	TinyImageNet
$\mathcal{L}_{\text{Align}}$	75.86	40.50
$\mathcal{L}_{\text{Corr}}$	73.73	36.70
$\mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Corr}}$	77.48	42.17
All	**77.95**	**42.32**

The k-NN accuracy of self-supervised pretrained ResNet-50(R-50) and ResNet-50w2(R50x2) are 61.9% and 67.3% [19]. We apply the same pretrained R50x2 teacher as [23], to train students (ResNet18 and ResNet34) using the same training strategy. The results show that our solution can further improve the classification accuracy of students.

We also evaluate the self-supervised KD by linear classification following previous works in SEED [23]. We apply the SGD optimizer and train the linear classifier for 100 epochs. The weight decay is set to be zero, and the learning rate is 30 at the beginning then reduced to three and 0.3 at 60 and 80 epochs. Table VI reports the Top-1 and Top-5 accuracy and indicates that our method also works well in self-supervised settings.

E. Ablation Study

To verify the importance of the transformation function $h_\varphi(\cdot)$, we apply two-layer MLP, which is widely used in self-supervised learning [8], [18], for $\mathcal{L}_{\text{Align}}$ and $\mathcal{L}_{\text{Corr}}$ on student's output. We set different dimensions for the hidden layer to model different capabilities. Table VII compares different multiples of the student representation's dimension [$\dim(\mathbf{z}_T)$] and shows that the choice of representation's dimension is important to achieve the optimal performance. A spindle-shaped MLP (16 times) can achieve best alignment results. For $\mathcal{L}_{\text{Corr}}$, we have not observed similar trends and directly set all dimensions to $\dim(\mathbf{z}_S)$. For the additional $\mathcal{L}_{\text{Sup}}$ and $\mathcal{L}_{\text{CE}}$ losses, we apply linear projections.

We also perform the ablation study to examine the effectiveness of each distillation objective, $\mathcal{L}_{\text{Align}}$, $\mathcal{L}_{\text{Corr}}$, and $\mathcal{L}_{\text{Sup}}$. The students are trained via different combinations of these objectives, as shown in Table VIII. We find that combinations of objectives can obtain better results than single objective, indicating that multiple supervisory signals can improve the representation quality of the student. And among these objectives, $\mathcal{L}_{\text{Align}}$ plays a more important role than others in KD. To demonstrate that $\mathcal{L}_{\text{Corr}}$ is also critical in distillation, we compare the transferability of learned representations by using $\mathcal{L}_{\text{Align}}$ and $\mathcal{L}_{\text{Corr}}$, as shown in Table IX. We find that $\mathcal{L}_{\text{Corr}}$ can boost the performance of transfer learning by capturing structural knowledge between samples, which is helpful to learn generalized representations.

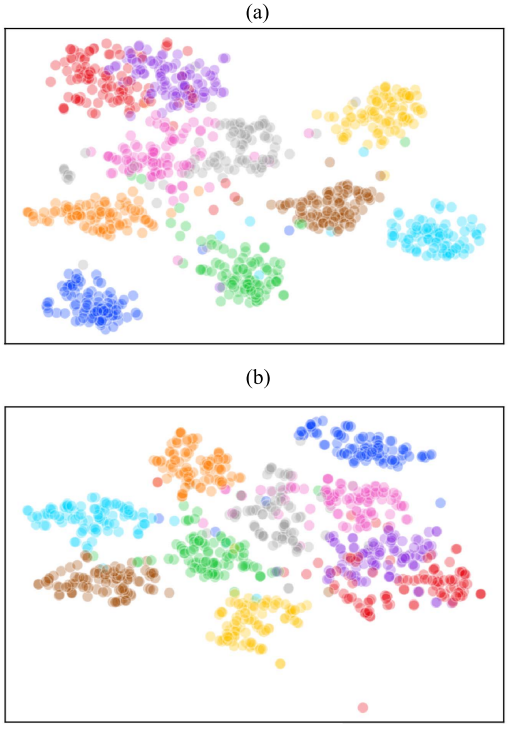


Fig. 2. t-SNE visualization of student's representations. (a) $\mathcal{L}_{\text{Align}}$ loss. (b) $\mathcal{L}_{\text{Corr}}$ loss (teacher: resnet32 × 4, student: resnet8 × 4). $\mathcal{L}_{\text{Align}}$ enables the student to learn representations with the large margin between different classes. $\mathcal{L}_{\text{Corr}}$ enables the student to learn better intraclass structure.

To visually understand the different roles of $\mathcal{L}_{\text{Align}}$ and $\mathcal{L}_{\text{Corr}}$, we perform t-SNE visualization on cifar100 dataset (randomly select ten categories from 100 categories), as shown in Fig. 2. $\mathcal{L}_{\text{Align}}$ tends to make the student learn representations with the large margin between different classes. In contrast, $\mathcal{L}_{\text{Corr}}$ enables the student to capture better intraclass structure for certain classes. It is necessary to combine them to improve the distillation performance.

TABLE X

CLASSIFICATION ACCURACY (%) OF STL10 (TEN CLASSES) AND TINYIMAGENET (200 CLASSES) USING LINEAR EVALUATION ON THE REPRESENTATIONS FROM CIFAR100 TRAINED NETWORKS. WE DENOTE COMPARED RESULTS FROM [10] BY *. WE DENOTE THE BEST AND THE SECOND-BEST RESULTS BY **BOLD** AND UNDERLINE

Dataset	STL10			TinyImageNet		
Teacher	resnet32×4	vgg13	wrn40-2	resnet32×4	vgg13	wrn40-2
Student	resnet8×4	vgg8	ShuffleV1	resnet8×4	vgg8	ShuffleV1
Teacher	70.45	64.45	71.01*	31.92	27.20	31.69
Student	71.26	67.48	71.58*	35.31	30.87	32.43*
KD [4]	71.29	67.81	73.25*	33.86	30.87	32.05*
Fitnets [11]	72.93	67.16	73.77*	37.86	31.20	33.28*
AT [12]	73.46	<u>71.65</u>	73.47*	36.53	33.23	33.75*
FT [30]	74.29	69.93	73.56*	<u>38.25</u>	32.73	33.69*
SP [14]	72.06	68.43	72.28	35.05	31.55	34.74
VID [31]	73.35	67.88	72.56	37.38	31.12	<u>35.62</u>
CRD [5]	73.39	69.20	74.44*	37.13	33.04	34.30*
SSKD [10]	<u>74.39</u>	71.24	<u>74.74</u>*	37.83	<u>34.87</u>	34.54*
DLKD	**77.95**	**74.49**	**77.43**	**42.31**	**38.74**	**42.48**

TABLE XI

QUANTIFICATION OF REPRESENTATIONAL KNOWLEDGE. IT REPORTS AVERAGE SCORES OF TWO STUDENTS TRAINED BY DIFFERENT DISTILLATION METHODS ON CIFAR100 TEST DATASET

Teacher	resnet32×4	resnet32×4
Student	resnet8×4	ShuffleV2
KD	0.4400	0.6307
CRD	0.1460	0.4454
$\mathcal{L}_{\text{Align}}$	0.0934	0.1641
$\mathcal{L}_{\text{Corr}}$	0.2533	0.4288
$\mathcal{L}_{\text{Sup}}$	0.2746	0.3816
DLKD	**0.0887**	**0.1622**

TABLE XII

QUANTIFICATION OF KNOWLEDGE CONSISTENCY. IT REPORTS IOU SCORES (0.0–1.0) OF STUDENTS TRAINED BY DIFFERENT DISTILLATION METHODS ON CIFAR100 DATASET, AND HIGHER IS BETTER

Teacher	resnet32×4	resnet32×4
Student	resnet8×4	ShuffleV2
KD	0.4647	0.2769
CRD	0.7288	0.4612
$\mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Corr}}$	0.7394	0.7449
DLKD	**0.7512**	**0.7528**

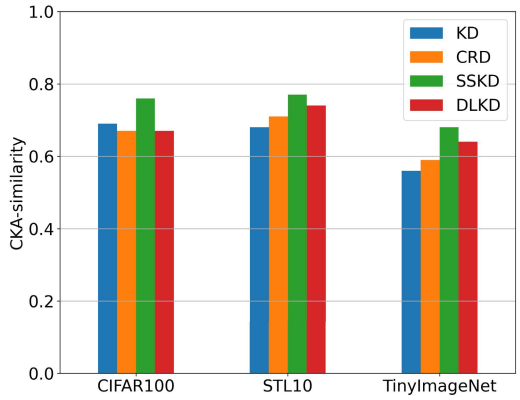


Fig. 3. CKA-similarity between the representations from the teacher (vgg13) and student (vgg8) networks.

F. Transferability of Representations

We also examine whether the representational knowledge learned by DLKD can be transferred to the unseen datasets. We perform six comparisons with three teacher–student pairs. The students are fixed to extract feature representations of STL10 and TinyImageNet datasets (all images resized to 32 × 32). We then compare the quality of the learned representations by training linear classifiers to perform ten-way and 200-way classification. As shown in Table X, DLKD achieves a significant performance improvement compared to multiple baseline methods, demonstrating the superior transferability of learned representations. Notably, most distillation methods improve the quality of the student's representations on STL10 and TinyImageNet. The reason why the teacher performs worse on these two datasets may be that the representations learned by the teacher are biased toward the training dataset and are not generalized well. In contrast, DLKD encourages the student to learn more generalized representations.

G. Quantification of Knowledge Consistency

Table XII compares the knowledge consistency of student networks trained by different distillation methods. It verifies that representation distillation can learn more reliable knowledge, compared with other distillation methods. Table XI shows the average score $\bar{H}$ of pixel-level conditional entropy as mentioned in Section IV. It indicates that the representation of lower $\bar{H}$ tends to achieve better classification performance. A lower $\bar{H}$ also means that the network focuses on more visual concepts to compute the feature representation. Our method has a lower $\bar{H}$, indicating that the student can learn richer representational knowledge from the teacher. Furthermore, we utilize the IoU score to quantify the knowledge consistency and evaluate the reliability of visual concepts, as shown in Table XII. We show that both of the average and IoU scores can provide additional insights about the KD, in addition to classification accuracy.

H. Teacher–Student Similarity

DLKD can encourage the student to learn richer structured representational knowledge under the dual-level supervisory signals of the teacher. Thus, we conduct the similarity analy-

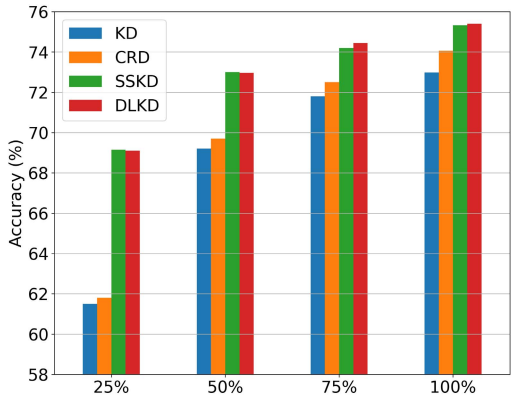


Fig. 4. Top-1 accuracy on CIFAR100 test data under a few-shot scenario. The student network is trained with only 25%, 50%, 75%, and 100% of the available training data.

sis between the teacher's and the student's representations to further understand the contrastive representation distillation. We calculate the centered kernel alignment (CKA)-similarity [46] (radial basis function (RBF) kernel) between the teacher and student networks, as shown in Fig. 3. Combined with Table X, we find that forcing students to be more similar to teachers does not guarantee that students can learn more general representations.

I. Few-Shot Scenario

DLKD enables the student to learn enough representational knowledge from the teacher, instead of relying entirely on labels. It is necessary to investigate the performance of DLKD under limited training data. We randomly sample 25%, 50%, 75%, and 100% images from CIFAR100 train set to train the student network and test on the original test set. The comparisons of different methods (see Fig. 4), show that DLKD maintains superior classification performance in all proportions. As the training set size decreases, dual-level supervisory signals in DLKD serve as an effective regularization to prevent overfitting.

VII. Conclusion

This work provides a novel viewpoint to analyze the existing distillation methods from knowledge alignment and correlation. We investigate their roles and reveal that both of them are necessary to improve distillation performance. In particular, knowledge correlation can serve as an effective regularization to learn generalized representations. We further demonstrate that our solution can increase the lower bound on mutual information between distributions of the teacher and student representations. DLKD is task-agnostic and model-agnostic, and can effectively transfer knowledge from supervised or self-supervised pretrained teachers. Due to the hardware limitation, we have not carried out fully hyperparameter tuning, which can be done in future works to further improve the distillation performance. Furthermore, we plan to apply our method to the domain adaptation task to investigate its adaptation ability in different distributions.

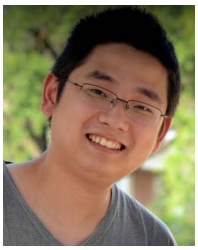
References

[1] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy Student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10687–10698.
[2] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
[3] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 535–541.
[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[5] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," 2019, *arXiv:1910.10699*.
[6] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[7] P. Bachman, R. Devon Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," 2019, *arXiv:1906.00910*.
[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.
[9] J. Yang, B. Marinez, S. A. Center, A. Bulat, and G. Tzimiropoulos, "Knowledge distillation via softmax regression representation learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–14.
[10] G. Xu, Z. Liu, X. Li, and C. Change Loy, "Knowledge distillation meets self-supervision," 2020, *arXiv:2006.07114*.
[11] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.
[12] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.
[13] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4133–4141.
[14] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.
[15] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3779–3787.
[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020, *arXiv:2006.09882*.
[18] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," 2020, *arXiv:2006.07733*.
[19] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," 2021, *arXiv:2104.14294*.
[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," 2017, *arXiv:1703.01780*.
[21] Y. Chen, N. Wang, and Z. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.
[22] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3967–3976.
[23] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, "SEED: Self-supervised distillation for visual representation," 2021, *arXiv:2101.04731*.
[24] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.
[25] P. Khosla *et al.*, "Supervised contrastive learning," 2020, *arXiv:2004.11362*.
[26] X. Cheng, Z. Rao, Y. Chen, and Q. Zhang, "Explaining knowledge distillation by quantifying the knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12925–12935.
[27] C. Guan, X. Wang, Q. Zhang, R. Chen, D. He, and X. Xie, "Towards a deep and unified understanding of deep neural models in NLP," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2454–2463.

[28] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
[29] C.-Y. Chuang, J. Robinson, L. Yen-Chen, A. Torralba, and S. Jegelka, "Debiased contrastive learning," 2020, *arXiv:2007.00224*.
[30] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2760–2769.
[31] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9163–9171.
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[35] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[36] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[37] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.
[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[39] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2604–2613.
[40] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[41] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[42] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.
[43] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[44] S. CS231N. (2015). *Tiny ImageNet Visual Recognition Challenge*. [Online]. Available: https://tiny-imagenet.herokuapp.com
[45] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 346–362.
[46] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, Jun. 2019, pp. 3519–3529.

Fei Ding is currently pursuing the Ph.D. degree with the School of Computing, Clemson University, Clemson, SC, USA.

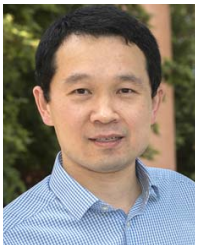
His research interests include deep neural networks, unsupervised learning, and machine learning.



Yin Yang (Member, IEEE) received the Ph.D. degree in computer science from The University of Texas at Dallas, Richardson, TX, USA, in 2013.

He is currently an Associate Professor with the School of Computing, Clemson University, Clemson, SC, USA. His research aims to develop efficient and customized computing methods for challenging problems in graphics, simulation, machine learning, vision, visualization, robotics, medicine, and many other applied areas.

Dr. Yang was a recipient of the National Science Foundation Computer and Information Science and Engineering Research Initiation Initiative (NSF CRII) Award in 2015 and the CAREER Award in 2019.



Hongxin Hu (Member, IEEE) received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in 2012.

He is currently an Associate Professor with the Department of Computer Science and Engineering, University at Buffalo The State University of New York, Buffalo, NY, USA. He has published more than 100 refereed technical articles, many of which appeared in top conferences and journals. His current research interests include security, networking, and machine learning.

Venkat Krovi (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering and applied mechanics from the University of Pennsylvania, Philadelphia, PA, USA, in 1998.

He is currently the Michelin Endowed Chair of vehicle automation with the Department of Automotive Engineering and the Department of Mechanical Engineering, Clemson University, Clemson University, SC, USA. His work has been published in more than 200 journal/conference articles and book chapters, and patents. His research interests include plant-automation, consumer electronics, automobile, defense, and healthcare/surgical simulation arenas.

Feng Luo (Senior Member, IEEE) received the Ph.D. degree in computer science from The University of Texas at Dallas, Richardson, TX, USA, in 2004.

He is currently the Marvin J. Pinson, Jr. '46 Distinguished Professor with the School of Computing, Clemson University, Clemson, SC, USA, and the Founding Director of the Clemson AI Research Institute for Science and Engineering (AIRISE), Clemson University. His research interests include deep learning and application, high throughput biological data analysis, data-intensive bioinformatics, network biology, and computational genomics and genetics.