DCD: Discriminative and Consistent Representation Distillation

Nikolaos Giakoumoglou
Imperial College London
London, UK, SW7 2AZ
n.giakoumoglou23@imperial.ac.uk

Tania Stathaki
Imperial College London
London, UK, SW7 2AZ
t.stathaki@imperial.ac.uk

Abstract

Knowledge Distillation (KD) aims to transfer knowledge from a large teacher model to a smaller student model. While contrastive learning has shown promise in self-supervised learning by creating discriminative representations, its application in knowledge distillation remains limited and focuses primarily on discrimination, neglecting the structural relationships captured by the teacher model. To address this limitation, we propose ***D****iscriminative and* ***C****onsistent* ***D****istillation (DCD), which employs a contrastive loss along with a consistency regularization to minimize the discrepancy between the distributions of teacher and student representations. Our method introduces learnable temperature and bias parameters that adapt during training to balance these complementary objectives, replacing the fixed hyperparameters commonly used in contrastive learning approaches. Through extensive experiments on CIFAR-100 and ImageNet ILSVRC-2012, we demonstrate that DCD achieves state-of-the-art performance, with the student model sometimes surpassing the teacher's accuracy. Furthermore, we show that DCD's learned representations exhibit superior cross-dataset generalization when transferred to Tiny ImageNet and STL-10. Code is available at* `https://github.com/giakoumoglou/distillers`.

1. Introduction

Knowledge Distillation (KD) has emerged as a prominent technique for model compression, enabling the transfer of knowledge from large, high-capacity teacher models to more compact student models [31]. This approach is particularly relevant today, as state-of-the-art vision models in tasks such as image classification [21, 42], object detection [39, 54], and semantic segmentation [10, 11] continue to grow in size and complexity. While these large models achieve impressive performance, their computational demands make them impractical for real-world applications [23, 36], leading practitioners to seek more efficient alternatives through model compression techniques [5, 53].

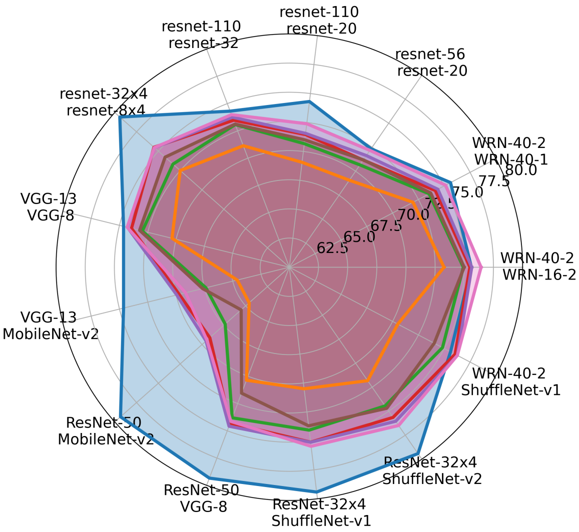


Figure 1. Radar plots comparing top-1 accuracy (%) of different distillation methods with DCD on CIFAR-100. Methods are represented by color as follows: Teacher (blue), Student (orange), KD [31] (green), CRD [60] (red), CRD+KD [60] (purple), DCD (ours) (brown), and DCD+KD (ours) (pink).

The representation learning capabilities of neural networks play a crucial role in their performance [4, 38]. In the context of KD, while the original approach [31] focused on transferring knowledge through logit outputs, subsequent work has emphasized the importance of intermediate feature representations [50, 55, 61, 67]. These intermediate representations capture rich structural information and hierarchical features that are essential for robust model performance [35, 70]. Recent advances in KD have explored various ways to transfer this representational knowledge, including attention transfer [67], correlation congruence [52], and relational knowledge distillation [50].

Contrastive learning has recently revolutionized self-supervised representation learning [13, 28], demonstrating its effectiveness in learning discriminative features without labels. This success has inspired its adoption in KD frameworks [22, 60]. However, existing contrastive distillation approaches face several limitations: they often require large memory banks to store negative samples [60], rely on fixed hyperparameters that limit their adaptability [13], and may not fully preserve the structural relationships captured by the teacher model [61]. Furthermore, the focus on discrimination alone can lead to suboptimal knowledge transfer, as it neglects the importance of maintaining consistent representational patterns between teacher and student models [52, 60].

To address these limitations, we propose ***D****iscriminative* and ***C****onsistent* ***D****istillation* (DCD), an approach that combines contrastive learning with consistency regularization to ensure both discriminative power and structural consistency in the student's representations (see Figure 2). Our method eliminates the need for memory banks by leveraging in-batch negative samples and introduces learnable temperature and bias parameters that dynamically adjust during training, enabling more flexible and efficient knowledge transfer. This adaptive approach allows the student to better capture both instance-level discriminative features and global structural patterns from the teacher, leading to more robust and generalizable representations [35].

Our **contributions** are twofold:

1. We propose DCD, an approach that combines contrastive learning with consistency regularization to ensure both discriminative and structurally consistent representations. Our method eliminates the need for memory banks and introduces learnable parameters that dynamically adjust during training, enabling more efficient knowledge transfer than existing approaches.
2. We demonstrate the effectiveness of DCD through extensive experiments on standard benchmarks, showing significant improvements in both accuracy and robustness. DCD outperforms other methods, achieving a 20.31% relative improvement[1] over the original KD. When combined with KD, it shows a 73.87% relative improvement over the original KD.

The rest of this paper is organized as follows. Section 2 reviews related work in KD and contrastive learning. Section 3 details our proposed DCD methodology. Section 4 presents our experimental setup and results, and Section 5 concludes the paper.

2. Related Work

Our method bridges the gap between contrastive learning and knowledge distillation by integrating contrastive learning principles with an explicit consistency regularization.

2.1. Knowledge Distillation

The seminal work on KD [31] introduced the concept of transferring knowledge through softened logit outputs, using temperature scaling in the softmax to better capture the teacher model's knowledge. Similar to the original KD work, our method utilizes temperature scaling, but we extend this concept through our adaptive temperature mechanism. Rather than using a fixed temperature parameter, DCD learns the optimal temperature during training, allowing for more flexible and dynamic knowledge transfer.

Logit-based methods. Following this direction, several works have enhanced logit-based distillation through various techniques. Some recent advancements include label decoupling [72], instance-specific label smoothing [66], probability reweighting [49], and normalizing logits before applying softmax and KL divergence [59]. These methods primarily focus on improving how the teacher's predictions are processed and transferred to guide the student's learning.

Feature-based methods. Another line of research focuses on leveraging intermediate feature representations for knowledge transfer. FitNets [55] pioneered this approach by using intermediate "hints" to guide the student's learning process. Subsequent works have explored various aspects of feature-based transfer: attention transfer [67] aligns attention maps between teacher and student, correlation congruence [52] preserves structural relationships in feature spaces, and Relational Knowledge Distillation (RKD) [50] transfers mutual relations of data examples. Recent advances include cross-stage connection paths [12], direct reuse of teacher's classifier [8], and many-to-one representation matching [41]. Notably, Contrastive Representation Distillation (CRD) [60] leverages contrastive learning to maximize mutual information between teacher and student representations, though it requires large memory buffers for storing negative samples. Like RKD, we preserve relational information through our consistency regularization term that explicitly maintains structural relationships. Moreover, our approach eliminates the need for memory banks through efficient in-batch negative sampling and adds structural consistency preservation, making it both more efficient and effective.

Architecture-aware methods. Recent work has also explored the interaction between network architecture and distillation effectiveness. Methods in this category include differentiable meta-learning for architecture search [19], training-free frameworks for student architecture selection [20], and graph-based methods for architecture adaptation [44]. These approaches aim to optimize not just the distil-

[1] Average relative improvement is calculated as: $\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{Acc}_{\mathrm{DCD}}^{i}-\mathrm{Acc}_{\mathrm{KD}}^{i}}{\mathrm{Acc}_{\mathrm{KD}}^{i}-\mathrm{Acc}_{\mathrm{van}}^{i}}$, where $\mathrm{Acc}_{\mathrm{DCD}}^{i}$, $\mathrm{Acc}_{\mathrm{KD}}^{i}$, and $\mathrm{Acc}_{\mathrm{van}}^{i}$ represent the accuracies of DCD, KD, and vanilla training of the i-th student model, respectively [60].

lation process but also the underlying network structures to achieve better knowledge transfer.

2.2. Contrastive Learning

Contrastive methods in self-supervised learning have proven effective for learning robust representations by maximizing mutual information [32, 62]. These methods transform unsupervised learning into a classification problem, building on foundational work in metric learning [16, 26] to distinguish between positive and negative samples. The theoretical foundations [2, 25] show that such objectives maximize a lower bound on mutual information, crucial for meaningful representations. Recent advances using momentum encoders [28], stronger augmentations [13], and methods that eliminate negative samples [24] have further improved self-supervised learning. Additionally, recent strategies explore invariance regularizers [48], while others prevent model collapse through redundancy reduction [69] or regularization [3]. Some approaches achieve this by eliminating negative samples through asymmetric Siamese structures or normalization [6, 14, 24]. Our method combines instance-level discrimination [63, 64] with an consistency constraint, ensuring the student learns both discriminative features and preserves the teacher's structural knowledge. Furthermore, our approach does not rely on fixed negative samples or momentum encoders [45]; instead, it employs a dynamic method that adapts to the model's current state during training. Our method shares the theoretical underpinnings of mutual information maximization but extends this framework to include explicit structural preservation, providing a more comprehensive approach to knowledge transfer.

3. Methodology

This section presents our methodology to improve the efficiency and accuracy of KD. Our method, ***D***iscriminative and ***C***onsistent ***D***istillation (DCD), focuses on learning representations that are both discriminative through contrastive learning and structurally consistent with the teacher model through a consistency regularization. DCD ensures that the student model learns to differentiate between different instances while preserving the distributional relationships captured by the teacher model. Figure 2 shows an overview of the proposed method in the latent space.

3.1. Preliminaries

KD involves transferring knowledge from a high-capacity teacher neural network, denoted a f^T, to a more compact student neural network f^S. Consider x_i as the input to these networks, typically an image. We represent the outputs at the penultimate layer (just before the final classification layer, or logits) as $\mathbf{z}_i^T = f^T(x_i)$ and $\mathbf{z}_i^S = f^S(x_i)$ for the teacher and student models, respectively. The primary objective of KD is to enable the student model to approximate the performance of the teacher model. The overall distillation process can be mathematically expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(y_i, \mathbf{z}_i^S) + \lambda \cdot \mathcal{L}_{\text{distill}}(\mathbf{z}_i^T, \mathbf{z}_i^S) \qquad (1)$$

where y_i represents the true label for the input x_i and λ is a hyperparameter that balances the supervised loss and the distillation loss. The supervised loss $\mathcal{L}_{\text{sup}}$ is the task-specific alignment error between the network prediction and annotation. For image classification [15, 47, 53, 57], this is typically cross-entropy loss, while for object detection [9, 40], it includes bounding box regression. The distillation loss $\mathcal{L}_{\text{distill}}$ is the mimic error of the student network towards the teacher network, typically implemented as KL divergence between student and teacher outputs [31].

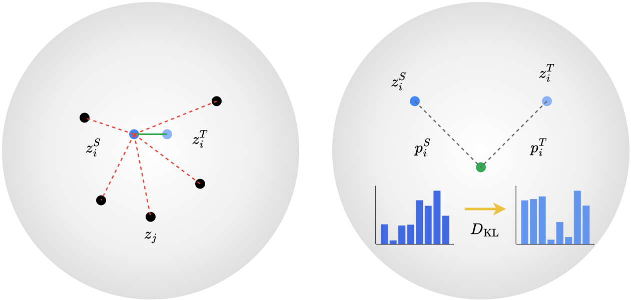


Figure 2. Overview of DCD. (a) Discriminative learning through contrastive distillation encourages student features (solid blue) to differentiate between instances by pulling them closer to their corresponding teacher features (transparent blue) while pushing away from other instances as negative samples (black dots). (b) Structural consistency through consistency regularization preserves the distributional relationship patterns captured by the teacher model by aligning the student and teacher feature similarities (represented by dotted lines) through KL divergence minimization.

3.2. Discriminative and Consistent Distillation

We develop an objective function that ensures both discriminative and structurally consistent representations between the teacher's output $\mathbf{z}_i^T$ and the student's output $\mathbf{z}_i^S$. This objective combines a contrastive loss, which discriminatively aligns representations, with a consistency regularization term that preserves structural relationships in the feature space. The objective function is defined as:

$$\mathcal{L}_{\text{kd}}(\mathbf{z}_i^T, \mathbf{z}_i^S) = \mathcal{L}_{\text{contrast}}(\mathbf{z}_i^T, \mathbf{z}_i^S) + \alpha \cdot \mathcal{L}_{\text{consist}}(\mathbf{z}_i^T, \mathbf{z}_i^S) \quad (2)$$

where α is a hyperparameter that balances the contrastive loss $\mathcal{L}_{\text{contrast}}$ for discriminative learning and the consistency regularization term $\mathcal{L}_{\text{consist}}$ for preserving structural relationships.

Discriminative distillation. In our approach, we employ contrastive learning to align teacher and student representations at the instance level. This process creates similarity between representations of the same input while pushing apart those from different inputs [62]. Through this discriminative mechanism, the student network learns to mirror the teacher's ability to distinguish between distinct data points.

Instance contrastive learning [63] extends class-wise supervision to its logical extreme by treating each individual instance as its own class. However, this creates a practical challenge: with the number of "classes" matching the number of training instances, implementing a traditional softmax layer becomes computationally intractable. We resolve this challenge by implementing Noise Contrastive Estimation (NCE) to approximate the softmax, enabling instance-level discrimination without explicit class boundaries:

$$\mathcal{L}_{\text{contrast}}(\mathbf{z}_i^T, \mathbf{z}_i^S) = -\log \frac{\exp(\phi(\mathbf{z}_i^S, \mathbf{z}_i^T)/\tau + b)}{\sum_{j=1}^{N} \exp(\phi(\mathbf{z}_i^S, \mathbf{z}_j^T)/\tau + b)} \tag{3}$$

where $\phi(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/|\mathbf{u}||\mathbf{v}|$ represents the cosine similarity function, with τ serving as the temperature parameter, b as a bias parameter, and N as the total number of negatives. This formulation effectively transforms into a cross-entropy loss, where each student representation $\mathbf{z}_i^S$ must identify its corresponding teacher representation $\mathbf{z}_i^T$ among all other teacher representations in the batch. The objective essentially becomes a classification task: student embeddings must "classify" their matching teacher embeddings correctly, with the normalized similarities acting as logits and positive pair indices as class labels. The parameters τ and b provide fine-grained control over this classification process, determining its sharpness and scale respectively.

Consistent distillation. The consistency loss regularizes the student model to maintain the structural relationships in the teacher model's representations. Unlike the contrastive term, which operates at the instance level, the consistency loss considers the distributional patterns. The student's distribution is defined as the similarity between student's instance i and all other instances j in the batch, processed through a softmax layer:

$$p_i^S(j) = \frac{\exp(\phi(\mathbf{z}_i^S, \mathbf{z}_j^T)/\tau + b)}{\sum_{k=1}^{N} \exp(\phi(\mathbf{z}_i^S, \mathbf{z}_k^T)/\tau + b)} \tag{4}$$

Similarly for the teacher:

$$p_i^T(j) = \frac{\exp(\phi(\mathbf{z}_i^T, \mathbf{z}_j^S)/\tau + b)}{\sum_{k=1}^{N} \exp(\phi(\mathbf{z}_j^T, \mathbf{z}_k^S)/\tau + b)} \tag{5}$$

To achieve this, we adopt a relational consistency approach, which preserves the distributional patterns captured by the teacher model across instances. By aligning the pairwise relationships between instances in the student and teacher embeddings, the model maintains the structural integrity of the teacher's learned representations. Through KL divergence between the student and teacher similarity distributions, this approach matches not only individual representations but also the spatial configuration of all instances in the embedding space, ensuring robustness and transferability. The consistency regularization term ensures that the student model learns to preserve the structural relationships present in the teacher's representations by minimizing the KL divergence between these distributions:

$$\mathcal{L}_{\text{consist}}(\mathbf{z}_i^T, \mathbf{z}_i^S) = D_{\text{KL}}(\mathbf{p}_i^S \| \mathbf{p}_i^T) = \sum_{j=1}^{N} p_i^S(j) \log \frac{p_i^S(j)}{p_i^T(j)} \tag{6}$$

where D_{KL} denotes the KL divergence between the distributions $\mathbf{p}_i^T$ and $\mathbf{p}_i^S$, ensuring that the student model maintains similar relational patterns as the teacher model across different inputs.

The combination of contrastive loss and consistency regularization ensures that the learned representations are both discriminative and structurally consistent with the teacher model. This is formalized by the following theorem:

Final objective. The final objective function, which includes the supervised loss and standard KL divergence, is given by:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}(y_i, \mathbf{z}_i^S) + \lambda \cdot \mathcal{L}_{\text{distill}}(\mathbf{z}_i^T, \mathbf{z}_i^S) + \beta \cdot \mathcal{L}_{\text{kd}}(\mathbf{z}_i^T, \mathbf{z}_i^S) \tag{7}$$

where β is a hyperparameter that balances $\mathcal{L}_{\text{kd}}$.

3.3. Implementation Details

We implement the objective using mini-batch stochastic gradient descent. The representations $\mathbf{z}_i^T$ and $\mathbf{z}_i^S$ are obtained from the last layer of the teacher and student models, respectively. We further encode $\mathbf{z}_i^T$ and $\mathbf{z}_i^S$ using a projection head to match the dimensions. The projection head is trained using stochastic gradient descent as well. This ensures that the representations from both models are compatible for comparison and alignment. Additionally, we normalize the outputs $\mathbf{z}_i^T$ and $\mathbf{z}_i^S$ before computing the loss, ensuring that the representations lie on a unit hypersphere. This ensures that the representations from both models are compatible for comparison and alignment.

Memory-efficient sampling. Instead of using a large memory buffer for contrasting representations as in CRD [60], we use the negative samples that naturally co-exist within the batch. This approach significantly reduces memory requirements while maintaining effective contrastive learning.

Adaptive temperature scaling. Contrary to contrastive learning objectives that use a constant temperature parameter, we parameterize the temperature using $\exp(\tau)$ where τ is a learnable parameter, along with a learnable bias b. For a batch of normalized embeddings $\mathbf{z}_i^S$ and $\mathbf{z}_i^T$, we compute the similarity matrix through:

$$\ell_{ij} = \phi(\mathbf{z}_i^S, \mathbf{z}_j^T) \cdot \exp(\tau) + b \tag{8}$$

where $\phi(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/|\mathbf{u}||\mathbf{v}|$ is implemented efficiently as a normalized matrix multiplication. The exponential parameterization ensures the temperature remains positive while allowing unconstrained optimization of τ, which is clamped to $[0, \tau_{\text{max}}]$ for numerical stability. The learnable bias b provides an additive degree of freedom that helps adjust the logit scale. This adaptive approach allows the model to automatically tune the contrast level and logit scaling during training, leading to more robust knowledge transfer compared to fixed hyperparameter approaches.

4. Experiments

We evaluate our DCD framework in the KD task for model compression of a large network to a smaller one, similar to [60]. This method aligns with common practices in the field, ensuring a fair comparison between different techniques.

4.1. Experimental Setup

We implement DCD in PyTorch following the implementation of CRD [60]. The detailed algorithm of DCD is provided in the supplementary material.

Datasets. We conduct experiments on four popular datasets for model compression: (1) CIFAR-100 [37] contains 50,000 training images with 500 images per class and 10,000 test images. (2) ImageNet ILSVRC-2012 [18] includes 1.2 million images from 1,000 classes for training and 50,000 for validation. (3) STL-10 [17] consists of a training set of 5,000 labeled images from 10 classes and 100,000 unlabeled images, and a test set of 8,000 images. (4) Tiny ImageNet [18] comprises 200 classes, each with 500 training images and 50 validation images.

Setup. We experiment with student-teacher combinations of different capacities, such as ResNet [27] or Wide ResNet (WRN) [68], VGG [58], MobileNet [56], and ShuffleNet [43, 71] (more details are described in the supplementary material). We set $\alpha = 0.5$ and $\beta = 1$ and ablate both α and β in the supplementary material. The hyperparameter λ is set to 1.0 for the KL divergence loss to maintain consistency with [7, 8, 60] (we also ablate λ in the supplementary material). Both the student and teacher outputs are projected to a 128-dimensional space using a projection head consisting of a single linear layer, followed by ℓ_2 normalization. The projection layer, beyond matching dimensions, plays a crucial role in KD by implicitly encoding relational information from previous samples [46]. We empirically set $\tau_{\text{max}} = 10.0$. More details on the training procedures can be found in the supplementary material.

Comparison. We compare our approach to the following state-of-the-art methods: (1) KD [31]; (2) FitNets [55]; (3) AT [67]; (4) SP [61]; (5) CC [52]; (6) VID [1]; (7) RKD [50]; (8) PKT [51]; (9) AB [30]; (10) FT [34]; (11) FSP [65]; (12) NST [33]; (13) CRD [60]; (14) OFD [29]; (15) WSLD [72]; (16) IPWD [49]. In the supplementary material, we include additional methods.

4.2. Results on CIFAR-100

Table 1 and Table 2 provide a comprehensive comparison of top-1 accuracies across various KD methods for both identical and differing architectures between student and teacher models on the CIFAR-100 dataset (extended comparison with more state-of-the-art methods is included in the supplementary material). Specifically, Table 1 focuses on scenarios where the student and teacher share the same architecture, while Table 2 explores settings with differing architectures. Our proposed method, DCD, and its combination with KD, consistently achieve superior performance compared to other distillation objectives, including the original KD. Our method surpasses the teacher network's performance in both same-architecture (WRN-40-2 to WRN-16-2) and cross-architecture (WRN-40-2 to ShuffleNet-v1) scenarios, achieving accuracy gains of 0.45% and 0.90%, respectively (see 1). As shown in the rightmost column of both tables, DCD, when combined with KD, achieves the highest average improvement over the baseline student model, with +2.82% for same-architecture scenarios and +5.25% for cross-architecture scenarios. The strong performance of both CRD and DCD validates the effectiveness of contrastive learning objectives in knowledge distillation, with our adaptive approach showing consistent improvements over the memory bank-based CRD method.

4.3. Results on ImageNet

Table 3 showcases the top-1 accuracy of student networks that were trained using various distillation methods on ImageNet. The results demonstrate the effectiveness of our method on large-scale datasets, highlighting its ability to distill knowledge from complex models and enhance the performance of student networks. Our approach achieves competitive results, surpassing KD and attaining state-of-the-art performance in ResNet-50 to ResNet-18 distillation. Our approach also shows improvement across different architectures, demonstrating its effectiveness in various distillation scenarios. DCD with KD consistently outperforms the standard KD method across all teacher-student combinations.

Table 1. Test top-1 accuracy (%) of student networks on CIFAR-100, comparing various distillation methods. The values in bold indicate the maximum of each column while underlined values mark the second best. Δ represents the mean accuracy improvement (%) over the baseline student model.

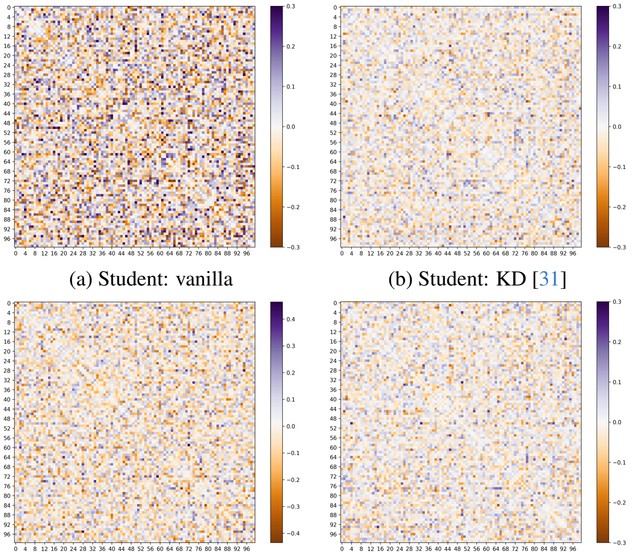
Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	resnet-56 resnet-20	resnet-110 resnet-20	resnet-110 resnet-32	resnet-32x4 resnet-8x4	VGG-13 VGG-8	Δ
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64	-
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36	0.00
KD [31]	74.92	73.54	70.66	70.67	73.08	73.33	72.98	+1.69
FitNet [55]	73.58	72.24	69.21	68.99	71.06	73.50	71.02	+0.32
AT [67]	74.08	72.77	70.55	70.22	72.31	73.44	71.43	+1.06
SP [61]	73.83	72.43	69.67	70.04	72.69	72.94	72.68	+0.99
CC [52]	73.56	72.21	69.63	69.48	71.48	72.97	70.81	+0.40
VID [1]	74.11	73.30	70.38	70.16	72.61	73.09	71.23	+1.07
RKD [50]	73.35	72.22	69.61	69.25	71.82	71.90	71.48	+0.18
PKT [51]	74.54	73.45	70.34	70.25	72.61	73.64	72.88	+1.48
AB [30]	72.50	72.38	69.47	69.53	70.98	73.17	70.94	+0.20
FT [34]	73.25	71.59	69.84	70.22	72.37	72.86	70.58	+0.45
FSP [65]	72.91	n/a	69.95	70.11	71.89	72.62	70.33	+0.41
NST [33]	73.68	72.24	69.60	69.53	71.96	73.30	71.53	+0.67
CRD [60]	75.48	74.14	71.16	71.46	73.48	75.51	73.94	+2.54
CRD+KD [60]	75.64	74.38	71.63	71.56	73.75	75.46	**74.29**	+2.48
OFD [29]	75.24	74.33	70.38	n/a	73.23	74.95	73.95	+2.30
WSLD [72]	n/a	73.74	71.53	n/a	73.36	74.79	n/a	+2.19
IPWD [49]	n/a	74.64	71.32	n/a	**73.91**	**76.03**	n/a	+2.81
DCD (ours)	74.99	73.69	71.18	71.00	73.12	74.23	73.22	+1.72
DCD+KD (ours)	**76.06**	**74.76**	**71.81**	**72.03**	73.62	75.09	73.95	**+2.82**

Table 2. Test top-1 accuracy (%) of student networks on CIFAR-100 involving students and teachers from different architectures, using various distillation methods. The values in bold indicate the maximum of each column while underlined values mark the second best. Δ represents the mean accuracy improvement (%) over the baseline student model.

Teacher Student	VGG-13 MobileNet-v2	ResNet-50 MobileNet-v2	ResNet-50 VGG-8	ResNet-32x4 ShuffleNet-v1	ResNet-32x4 ShuffleNet-v2	WRN-40-2 ShuffleNet-v1	Δ
Teacher	74.64	79.34	79.34	79.42	79.42	75.61	-
Student	64.60	64.60	70.36	70.50	71.82	70.50	0.00
KD [31]	67.37	67.35	73.81	74.07	74.45	74.83	+3.25
FitNet [55]	64.14	63.16	70.69	73.59	73.54	73.73	+1.08
AT [67]	59.40	58.58	71.84	71.73	72.73	73.32	-0.80
SP [61]	66.30	68.08	73.34	73.48	74.56	74.52	+2.98
CC [52]	64.86	65.43	70.25	71.14	71.29	71.38	+0.50
VID [1]	65.56	67.57	70.30	73.38	73.40	73.61	+1.91
RKD [50]	64.52	64.43	71.50	72.28	73.21	72.21	+0.96
PKT [51]	67.13	66.52	73.01	74.10	74.69	73.89	+2.83
AB [30]	66.06	67.20	70.65	73.55	74.31	73.34	+2.12
FT [34]	61.78	60.99	70.29	71.75	72.50	72.03	-0.51
NST [33]	58.16	64.96	71.28	74.12	74.68	76.09	+1.15
CRD [60]	69.73	69.11	74.30	75.11	75.65	76.05	+4.43
CRD+KD [60]	**69.94**	69.54	74.58	75.12	76.05	76.27	+4.69
OFD [29]	69.48	69.04	n/a	75.98	76.82	75.85	+4.83
WSLD [72]	n/a	68.79	73.80	75.09	n/a	75.23	+4.24
IPWD [49]	n/a	**70.25**	**74.97**	**76.03**	n/a	76.44	**+5.43**
DCD (ours)	68.35	67.39	73.85	74.26	75.26	74.98	+3.62
DCD+KD (ours)	69.77	70.03	74.08	76.01	**76.95**	**76.51**	+5.25

Table 3. Test top-1 (%) on ImageNet validation test using various distillation methods. The table compares students and teachers of the same and different architecture. The values in bold indicate the maximum of each column while underlined values mark the second best.

Teacher Student	ResNet-34 ResNet-18	ResNet-50 ResNet-18	ResNet-50 MobileNet
Teacher	73.31	76.16	76.16
Student	69.75	69.75	69.63
KD [31]	70.67	71.29	70.49
AT [67]	71.03	71.18	70.18
SP [61]	70.62	71.08	n/a
CC [52]	69.96	n/a	n/a
VID [1]	n/a	71.11	n/a
RKD [50]	70.40	n/a	68.50
AB [30]	n/a	n/a	68.89
FT [34]	n/a	n/a	69.88
FSP [65]	70.58	n/a	n/a
NST [33]	70.29	n/a	n/a
CRD [60]	71.17	71.25	69.07
OFD [29]	71.03	n/a	71.33
WSLD [72]	**72.04**	n/a	71.52
IPWD [49]	71.88	n/a	**72.65**
DCD (ours)	71.10	71.38	70.51
DCD+KD (ours)	71.71	**71.65**	71.55



(a) Student: vanilla

(b) Student: KD [31]

(c) Student: AT [67]

(d) Student: DCD (ours)

Figure 3. Comparison of correlation matrix differences between teacher and student logits across various distillation methods on the CIFAR-100 task. Subfigures show results for (a) students trained without distillation, (b) with KL divergence [31], (c) with attention transfer (AT) [67], and (d) with our DCD method, highlighting better matching between student's and teacher's correlations. Results have been re-implemented according to [60].

4.4. Capturing Inter-class Correlations

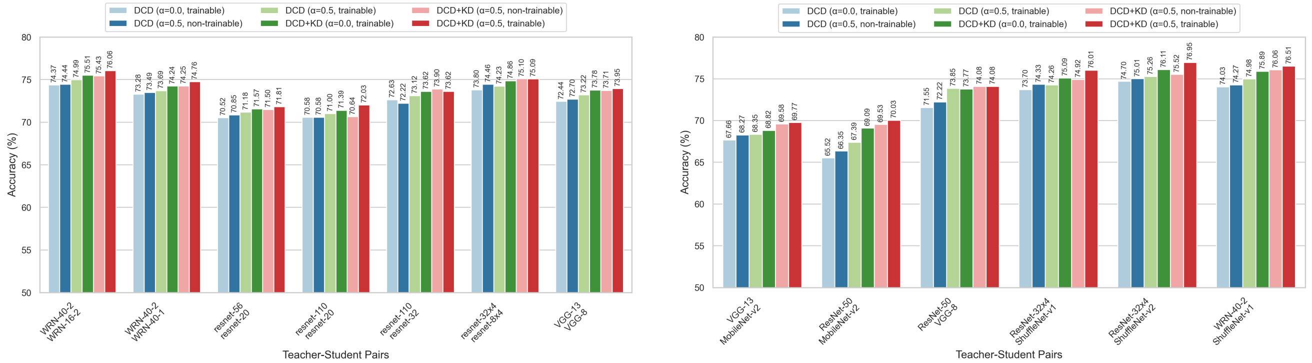
Cross-entropy loss ignores the correlations among class logits in a teacher network, often leading to suboptimal knowledge transfer. By employing ”soft targets”, distillation methods such as those described by [31] have successfully captured these correlations, enhancing student learning. Figure 3 evaluates the effectiveness of various distillation approaches on the CIFAR-100 KD task using WRN-40-2 as the teacher and WRN-40-1 as the student. Specifically, we compare students trained without distillation, with attention transfer [67], with KL divergence [31], and with our proposed DCD method. Our results demonstrate that DCD achieves close correlation alignment between teacher and student logits, as evidenced by the minimal differences in their correlation matrices. However, compared to CRD [60], our method achieves less optimal matching, though it still significantly enhances learning efficiency and lowers error rates. The smaller discrepancies between teacher and student logits clearly indicate that the DCD objective captures a substantial amount of correlation structure in the logit, resulting in lower error rates, although it is surpassed slightly by CRD in achieving the closest match. Moreover, our method also enhances representation learning as it employs a contrastive objective.

4.5. Transferability of Representations

Our research focuses on transferring knowledge from a teacher network to a student network while learning representations that embody general knowledge applicable across various tasks and datasets. To investigate this, we employ a distillation process where a WRN-40-2 teacher network transfers its learned representations to a WRN-16-2 student network, which can either be trained directly from the CIFAR-100 dataset or through distillation. In our experiments, the student network functions as a fixed feature extractor, processing images from STL-10 and Tiny ImageNet, both resized to 32×32. To assess the generalizability of these representations, we train a linear classifier on top of the last feature layer to perform 10-way classification for STL-10 and 200-way classification for Tiny ImageNet. The effectiveness of different distillation methods in enhancing the transferability of these representations is detailed in Table 4. Our results indicate that the DCD method, both standalone and in combination with KD, significantly outperforms all other distillation methods in improving the transferability of learned representations across different datasets. This superior transferability suggests that DCD encourages the student to learn more general and robust features, which are less overfitted to the specific training dataset and more applicable to diverse visual recognition tasks.

	Teacher	*Student*	KD	AT	FitNet	CRD	CRD+KD	DCD	DCD+KD
CIFAR-100→STL-10	68.6	69.7	70.9	70.7	70.3	71.6	72.2	71.2	**72.5**
CIFAR-100→Tiny ImageNet	31.5	33.7	33.9	34.2	33.5	35.6	35.5	35.0	**36.2**

Table 4. Test top-1 accuracy (%) of WRN-16-2 (student) distilled from WRN-40-2 (teacher). In this setup, the representations learned from the CIFAR-100 dataset are transferred to the STL-10 and Tiny ImageNet datasets. The network is frozen, and a linear classifier is trained on the last feature layer to perform classification with 10 classes (STL-10) or 200 classes (Tiny ImageNet). The values in bold indicate the maximum of each row.



(a) Same-architecture distillation

(b) Cross-architecture distillation

Figure 4. Ablation results on CIFAR-100 comparing DCD variants. (a) Results for same-architecture pairs, where teacher and student share network architecture. (b) Results for cross-architecture pairs, where teacher and student are from different architectures. Three configurations are analyzed: a discriminative-only loss ($\alpha = 0$), fixed temperature scaling with consistency regularization ($\alpha = 0.5$, $\tau = 0.07$, $b = 0$), and our proposed method with consistency regularization and trainable temperature parameters ($\alpha = 0.5$, trainable τ and b).

4.6. Ablation Study

To examine the effectiveness of consistency regularization and temperature scaling, we conduct ablation studies with two DCD variants in both Figure 4a and Figure 4b. First, we evaluate a pure discriminative variant ($\alpha = 0$) without consistency regularization. Then, we test adding consistency regularization ($\alpha = 0.5$) with fixed temperature parameters ($\tau = 0.07$, $b = 0$). The results demonstrate that incorporating both consistency regularization and trainable temperature scaling yields the best performance, with improvements of up to 0.62% and 0.69% over the discriminative-only variant in same-architecture and cross-architecture scenarios, respectively. Additionally, removing the trainable temperature and bias parameters leads to a decrease in performance. This suggests that adaptive temperature scaling plays a crucial role in maximizing the benefits of consistency regularization during knowledge distillation. Additional ablation studies on the loss coefficients α, β, and λ are provided in the supplementary material. Our method demonstrates robust performance across different values of α and β due to the adaptive temperature scaling mechanism with trainable parameters τ and b, which automatically tunes the contrast level and logit scaling during training. Following prior work [7, 8, 60], we set $\lambda = 1.0$ as the default value.

5. Conclusions

We have presented DCD (***D***iscriminative and ***C***onsistent ***D***istillation), a knowledge distillation method that combines contrastive learning with consistency regularization to improve the traditional KD process. Our method achieves state-of-the-art performance through memory-efficient in-batch negative sampling and adaptive temperature scaling, eliminating the need for large memory banks while automatically tuning contrast levels during training. Through extensive experimentation across CIFAR-100, ImageNet, STL-10, and Tiny ImageNet datasets, we have demonstrated significant improvements over existing methods, achieving 20.31% relative improvement over the original KD, and 73.87% when combined with KD. Unlike previous methods such as CRD that require large memory banks or WSLD and IPWD that focus solely on instance discrimination, our approach achieves superior performance while being more memory-efficient and capturing both local and global structural information. The effectiveness of DCD has been validated in both same-architecture and cross-architecture scenarios, with student models in several cases exceeding their teachers' performance, particularly in WRN-40-2 to WRN-16-2 and WRN-40-2 to ShuffleNet-v1 configurations. We hope this work will inspire future research in knowledge distillation.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019. 5, 6, 7

[2] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019. 3

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. 3

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013. 1

[5] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 1

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 3

[7] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration, 2021. 5, 8

[8] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022. 2, 5, 8

[9] Gongfan Chen, Yuting Wang, Jiajun Xu, Zhe Du, Qionghai Dai, Shiyang Geng, and Tao Mei. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017. 1

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 1

[12] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review, 2021. 2

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 2, 3

[14] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning, 2020. 3

[15] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4794–4802, 2019. 3

[16] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 539–546 vol. 1, 2005. 3

[17] Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 921–928, Madison, WI, USA, 2011. Omnipress. 5

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[19] Xueqing Deng, Dawei Sun, Shawn Newsam, and Peng Wang. Distpro: Searching a fast knowledge distillation process via meta optimization, 2022. 2

[20] Peijie Dong, Lujun Li, and Zimian Wei. Diswot: Student architecture search for distillation without training, 2023. 2

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[22] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation, 2021. 2

[23] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning, 2019. 1

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 3

[25] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 3

[26] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, page 1735–1742. IEEE, 2006. 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 1, 3

[29] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation, 2019. 5, 6, 7

[30] Byeongho Heo, Minsik Lee, Seong Joon Yun, Jin Young Choi, and In So Kweon. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2019. 5, 6, 7

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 2, 3, 5, 6, 7

[32] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019. 3

[33] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. In *Advances in Neural Information Processing Systems*, pages 185–195, 2017. 5, 6, 7

[34] Jangho Kim, Seongwon Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*, pages 2760–2769, 2018. 5, 6, 7

[35] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning, 2019. 1, 2

[36] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better?, 2019. 1

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. 5

[38] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 1

[40] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018. 3

[41] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching, 2023. 2

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1

[43] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 5

[44] Yuchen Ma, Yanbei Chen, and Zeynep Akata. Distilling knowledge from self-supervised teacher by embedding graph alignment, 2022. 2

[45] Nicolas Michel, Maorong Wang, Ling Xiao, and Toshihiko Yamasaki. Rethinking momentum knowledge distillation in online continual learning, 2024. 3

[46] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation, 2024. 5

[47] Arun Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *International Conference on Learning Representations*, 2017. 3

[48] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms, 2020. 3

[49] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation, 2022. 2, 5, 6, 7

[50] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 1, 2, 5, 6, 7

[51] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018. 5, 6

[52] Baoyun Peng, Xi Li, Yifan Wu, Yizhou Fan, Bo Wang, Qi Tian, and Jun Liang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5007–5016, 2019. 1, 2, 5, 6, 7

[53] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018. 1, 3

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[55] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the 4th International Conference on Learning Representations*, 2014. 1, 2, 5, 6

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[57] Li Shen and Marios Savvides. Amalgamating knowledge towards comprehensive classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1687–1696, 2020. 3

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 5

[59] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation, 2024. 2

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation, 2022. 1, 2, 4, 5, 6, 7, 8

[61] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 1, 2, 5, 6, 7

[62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 3, 4

[63] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 3, 4

[64] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature, 2019. 3

[65] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 5, 6, 7

[66] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization, 2021. 2

[67] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the 5th International Conference on Learning Representations*, 2016. 1, 2, 5, 6, 7

[68] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. 5

[69] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. 3

[70] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020. 1

[71] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 5

[72] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective, 2021. 2, 5, 6, 7