Discriminative and Consistent Representation Distillation

Nikolaos Giakoumoglou
Imperial College London
London, UK, SW7 2AZ
nikos@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[1].*

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 [33]. This approach is particularly relevant today, as state-of-the-art vision models in tasks such as image classification [21, 48], object detection [44, 60], 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, 39], leading prac-
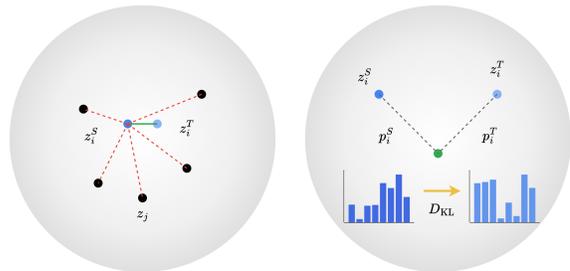

Figure 1. Overview of DCD. (a) Discriminative learning through contrastive distillation encourages student features (solid blue) to differentiate between instances by pulling them closer to their corresponding teacher features (transparent blue) while pushing away from other instances as negative samples (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.

titioners to seek more efficient alternatives through model compression techniques [5, 59].

The representation learning capabilities of neural networks play a crucial role in their performance [4, 41]. In the context of KD, while the original approach [33] focused on transferring knowledge through logit outputs, subsequent work has emphasized the importance of intermediate feature representations [56, 61, 67, 74]. These intermediate representations capture rich structural information and hierarchical features that are essential for robust model performance [38, 77]. Recent advances in KD have explored various ways to transfer this representational knowledge, including attention transfer [74], correlation congruence [58], and relational knowledge distillation [56].

Contrastive learning has recently revolutionized self-supervised representation learning [13, 30], demonstrating its effectiveness in learning discriminative features without labels. This success has inspired its adoption in KD frameworks [22, 66]. However, existing contrastive distillation approaches face several limitations: they often require large

---

[1]Code is available at: https://github.com/giakoumoglou/distillers.

memory banks to store negative samples [66], rely on fixed hyperparameters that limit their adaptability [13], and may not fully preserve the structural relationships captured by the teacher model [67]. 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 [58, 66].

To address these limitations, we propose *Discriminative and Consistent Distillation* (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 1). 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 [38].

Our **contributions** are twofold:

1. We propose 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 our approach through extensive experiments on standard benchmarks, showing significant improvements in both accuracy and robustness. DCD outperforms other methods, achieving a 20.31% relative improvement[2] 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 knowledge distillation and contrastive learning. Section 3 details our proposed methodology. Section 4 presents our experimental setup and results, and Section 5 concludes the paper.

2. Related Work

Our method combines contrastive learning with knowledge distillation by adding consistency regularization.

---

[2]Average relative improvement is calculated as: $\frac{1}{N} \sum_{i=1}^{N} \frac{\text{Acc}_{\text{DCD}}^i - \text{Acc}_{\text{KD}}^i}{\text{Acc}_{\text{KD}}^i - \text{Acc}_{\text{van}}^i}$, where $\text{Acc}_{\text{DCD}}^i$, Acc_{KD}^i, and $\text{Acc}_{\text{van}}^i$ represent the accuracies of DCD, KD, and vanilla training of the i-th student model, respectively [66].

2.1. Knowledge Distillation

The original knowledge distillation work by [33] introduced transferring knowledge through softened logit outputs using temperature scaling in the softmax. Similar to [33], our method uses temperature scaling but learns the optimal temperature during training for better knowledge transfer.

Logit-based methods. Several methods have improved logit-based distillation through techniques like label decoupling [81], instance-specific label smoothing [73], probability reweighting [55], and normalizing logits before softmax and KL divergence [65]. Some works focus on dynamic temperature adjustment [42] and separate learning of response-based and feature-based distillation [79]. Additional improvements include transformations for better teacher-student alignment [80] and methods for knowledge transfer from stronger teachers [35]. Unlike existing methods such as [42] that use fixed temperature schedules and MLPs, our approach introduces truly learnable temperature parameters that adapt during training, making it more flexible and efficient.

Feature-based methods. The work by [61] used intermediate feature hints to guide student learning, while [74] aligned attention maps between teacher and student. Several methods have explored structural relationships: [58] preserved feature space relationships, [45] ensured functional consistency, and [25] transferred class-level attention information. The method by [56] transferred mutual relations between data examples. Recent works have introduced cross-stage connection paths [12], direct reuse of teacher's classifier [8], and many-to-one representation matching [47]. Contrastive Representation Distillation (CRD) [66] used contrastive learning to maximize mutual information between representations but needed large memory buffers. Compared to [66], our approach uses efficient in-batch sampling and adds structural consistency, making it more practical. Our method also differs from previous approaches like [35, 57, 58, 67] in how we define and preserve structural relationships.

Architecture-aware methods. Recent studies have examined how network architecture affects distillation success. Some works developed meta-learning for architecture search [19], training-free student architecture selection [20], and graph-based architecture adaptation [50]. Methods by [28] and [52] addressed distillation between different architectures through unified feature spaces and intermediate networks respectively. These approaches aim to optimize not just the distillation 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 [34, 68]. These methods transform unsupervised learning into a classification problem, building on foundational work in metric learning [16, 27] to distinguish between positive and negative samples. The theoretical foundations [2, 26] show that such objectives maximize a lower bound on mutual information, crucial for meaningful representations. Recent advances using momentum encoders [30], stronger augmentations [13], and methods that eliminate negative samples [24] have further improved self-supervised learning. Additionally, recent strategies explore invariance regularizers [54], while others prevent model collapse through redundancy reduction [76] 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 [70, 71] 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 [51]; 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, *Discriminative and Consistent Distillation* (DCD), focuses on learning representations that are both discriminative through contrastive learning and structurally consistent with the teacher model through a consistency regularization. Our method ensures that the student model learns to differentiate between different instances while preserving the distributional relationships captured by the teacher model. Figure 1 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:

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

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, 53, 59, 63], this is typically cross-entropy loss, while for object detection [9, 46], 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 [33].

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 [68]. Through this discriminative mechanism, the student network learns to mirror the teacher's ability to distinguish between distinct data points.

Instance contrastive learning [70] 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:

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

where $\phi(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}/|\mathbf{u}||\mathbf{v}|$ 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:

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

Similarly for the teacher:

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

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:

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

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:

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

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 [66], we use the negative samples that naturally co-exist within the batch. This approach significantly reduces memory requirements while maintaining effective contrastive learning. Our in-batch negative sampling eliminates the need to maintain and update large memory banks that store thousands of feature vectors, which can consume significant GPU memory It also simplifies the implementation by removing the complexity of memory bank management, including challenges related to feature staleness and queue maintenance [30]. Unlike memory banks that may contain stale features from earlier training iterations, our approach always uses the most recent representations within the current batch.

Learnable temperature. 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:

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

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 $[0, \tau_{max}]$ 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 [66]. Beyond classification, we also validate our method's effectiveness on the more challenging object detection task.

4.1. Experimental Setup

We implement DCD in PyTorch following the implementation of [66].

Datasets. We conduct experiments on four popular datasets for model compression: (1) CIFAR-100 [40] 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. (5) MS-COCO [43] contains 118,000 training images and 5,000 validation images with annotations for object detection.

Setup. We experiment with student-teacher combinations of different capacities, such as ResNet [29] or Wide ResNet (WRN) [75], VGG [64], MobileNet [62], and ShuffleNet [49, 78]. We set $\alpha = 0.5$ and $\beta = 1$. The hyperparameter λ is set to 1.0 for the KL divergence loss to maintain consistency with [7, 8, 66]. 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. We empirically set $\tau_{max} = 10.0$.

Comparison. We compare our approach to the following state-of-the-art methods: (1) KD [33]; (2) FitNets [61]; (3) AT [74]; (4) SP [67]; (5) CC [58]; (6) VID [1]; (7) RKD [56]; (8) PKT [57]; (9) AB [32]; (10) FT [37]; (11) FSP [72]; (12) NST [36]; (13) CRD [66]; (14) OFD [31]; (15) WSLD [81]; (16) IPWD [55]; (17) CTKD [42].

Computational cost. While CRD's memory bank requires approximately 8MB per class (16k features $\times$128 dimensions $\times$4 bytes), DCD's in-batch sampling needs only 0.13MB total ($256 \times 128 \times 4$ bytes). This efficiency extends

to training time - on a 4-GPU machine, DCD completes ImageNet training in approximately 72 hours compared to CRD's 88 hours, representing an 18% reduction in training time.

4.2. Main Results

We benchmark our method on image classification and object detection tasks.

Results on CIFAR-100. Table 1 compares top-1 accuracies of various KD methods on CIFAR-100. Our DCD+KD achieves superior performance, surpassing the teacher network by **+0.45%** in same-architecture (WRN-40-2 to WRN-16-2) and **+0.90%** in cross-architecture (WRN-40-2 to ShuffleNet-v1) scenarios. The method shows significant improvements over baseline students: **+2.82%** for same-architecture and **+5.25%** for cross-architecture pairs, outperforming memory bank-based CRD. DCD alone performs slightly below CRD, which uses a large 16k-feature memory bank. However, DCD with KD achieves better results by combining complementary objectives: KD's soft targets provide direct class-level supervision through logit-space KL divergence, while DCD ensures feature-space consistency.

Transferability of representations. Table 2 compares the top-1 test accuracy of WRN-16-2 (student) distilled from WRN-40-2 (teacher) and evaluated on STL-10 and Tiny ImageNet. The student is trained on CIFAR-100, either directly or via distillation, and serves as a frozen feature extractor with a linear classifier. We assess how well different distillation methods enhance feature transferability. Our results show that DCD, both standalone and combined with KD, significantly improves transferability.

Results on ImageNet. Table 3 presents the top-1 accuracy of student networks trained with various distillation methods on ImageNet. The results highlight the effectiveness of our approach in large-scale settings, demonstrating its ability to distill knowledge from complex models and enhance student performance. Our method consistently surpasses KD and achieves competitive results across different architectures.

Results on COCO. Table 4 shows our performance on the MS-COCO object detection task. Following [79], we adopt Faster R-CNN [60] with Feature Pyramid Network (FPN) [44] as our detection framework. We evaluate two teacher-student scenarios: ResNet-101 to ResNet-50 and ResNet-50 to MobileNet-V2. For ResNet-101 to ResNet-50 distillation, DCD+KD achieves 39.01 AP and 60.07 AP_{50}, surpassing the baseline by +1.08 AP. When distilling from ResNet-50 to MobileNet-v2, we obtain substantial improvements of +3.81 AP over the baseline, demonstrating effective knowledge transfer even across different architectures.

Table 1. Test top-1 accuracy (%) on CIFAR-100 of student networks trained with various distillation methods across different teacher-student architectures. Architecture abbreviations: W: WideResNet, R: ResNet, MN: MobileNet, SN: ShuffleNet. Results adapted from [66]. Results for our method are averaged over *five* runs.

	Same architecture							Different architecture					
Teacher	W-40-2	W-40-2	R-56	R-110	R-110	R-32x4	VGG-13	VGG-13	R-50	R-50	R-32x4	R-32x4	W-40-2
Student	W-16-2	W-40-1	R-20	R-20	R-32	R-8x4	VGG-8	MN-v2	MN-v2	VGG-8	SN-v1	SN-v2	SN-v1
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64	74.64	79.34	79.34	79.42	79.42	75.61
Student	73.26	71.98	69.06	69.06	71.14	72.50	70.36	64.60	64.60	70.36	70.50	71.82	70.50
KD [33]	74.92	73.54	70.66	70.67	73.08	73.33	72.98	67.37	67.35	73.81	74.07	74.45	74.83
FitNet [61]	73.58	72.24	69.21	68.99	71.06	73.50	71.02	64.14	63.16	70.69	73.59	73.54	73.73
AT [74]	74.08	72.77	70.55	70.22	72.31	73.44	71.43	59.40	58.58	71.84	71.73	72.73	73.32
SP [67]	73.83	72.43	69.67	70.04	72.69	72.94	72.68	66.30	68.08	73.34	73.48	74.56	74.52
CC [58]	73.56	72.21	69.63	69.48	71.48	72.97	70.81	64.86	65.43	70.25	71.14	71.29	71.38
VID [1]	74.11	73.30	70.38	70.16	72.61	73.09	71.23	65.56	67.57	70.30	73.38	73.40	73.61
RKD [56]	73.35	72.22	69.61	69.25	71.82	71.90	71.48	64.52	64.43	71.50	72.28	73.21	72.21
PKT [57]	74.54	73.45	70.34	70.25	72.61	73.64	72.88	67.13	66.52	73.01	74.10	74.69	73.89
AB [32]	72.50	72.38	69.47	69.53	70.98	73.17	70.94	66.06	67.20	70.65	73.55	74.31	73.34
FT [37]	73.25	71.59	69.84	70.22	72.37	72.86	70.58	61.78	60.99	70.29	71.75	72.50	72.03
FSP [72]	72.91	n/a	69.95	70.11	71.89	72.62	70.33	58.16	64.96	71.28	74.12	74.68	76.09
CRD [66]	75.48	74.14	71.16	71.46	73.48	75.51	73.94	69.73	69.11	74.30	75.11	75.65	76.05
CRD+KD [66]	75.64	74.38	71.63	71.56	73.75	75.46	**74.29**	**69.94**	69.54	74.58	75.12	76.05	76.27
OFD [31]	75.24	74.33	70.38	n/a	73.23	74.95	73.95	69.48	69.04	n/a	75.98	76.82	75.8
WSLD [81]	n/a	73.74	71.53	n/a	73.36	74.79	n/a	n/a	68.79	73.80	75.09	n/a	75.23
IPWD [55]	n/a	74.64	71.32	n/a	**73.91**	76.03	n/a	n/a	**70.25**	74.97	76.03	n/a	76.44
CTKD [42]	75.45	73.93	71.19	70.99	73.52	n/a	73.52	68.46	68.47	n/a	74.78	75.31	75.78
DCD (ours)	74.99	73.69	71.18	71.00	73.12	74.23	73.22	68.35	67.39	73.85	74.26	75.26	74.98
DCD+KD (ours)	**76.06**	**74.76**	**71.81**	**72.03**	73.62	75.09	73.95	69.77	70.03	74.08	76.01	**76.95**	**76.51**

Table 2. 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.

	Teacher	Student	KD [33]	AT [74]	FitNet [61]	CRD [66]	CRD+KD [66]	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**

4.3. Visualization

We present comprehensive visualizations that analyze the learned representations and knowledge transfer patterns across different distillation approaches, providing qualitative insights.

Inter-class correlations. Figure 2 evaluates the effectiveness of distillation methods on the CIFAR-100 KD task using WRN-40-2 (teacher) and WRN-40-1 (student). We compare models trained without distillation, with KL divergence [33], with CRD [66], and our proposed DCD method. DCD achieves strong correlation alignment between teacher and student logits, reducing discrepancies in their correlation matrices.

t-SNE visualization. Figure 3 presents t-SNE [69] visualizations of embeddings from WRN-40-2 (teacher) and WRN-40-1 (student) on CIFAR-100. Compared to standard training and [74], DCD better aligns student embeddings with the teacher, preserving structural relationships in the feature space. The preservation of semantic relationships between class clusters demonstrates our method's effectiveness in transferring the teacher's knowledge organization.

4.4. Ablation Study

We analyze our method through ablation experiments presented in Figures 4 and 5. We first study the effects of consistency regularization and temperature scaling, then investigate the sensitivity to hyperparameters α, β, and λ.

Table 3. Test top-1 accuracy (%) on ImageNet validation set for student networks trained with various distillation methods across different teacher-student architectures. Results for our method are based on a *single* run.

	Teacher	*Student*	KD [33]	AT [74]	SP [67]	CC [58]	RKD [56]	CRD [66]	DCD	DCD+KD
ResNet-34→ResNet-18	73.31	69.75	70.67	71.03	70.62	69.96	70.40	71.17	71.10	**71.71**
ResNet-50→ResNet-18	76.16	69.75	71.29	71.18	71.08	n/a	n/a	71.25	71.38	**71.65**
ResNet-50→MobileNet-v2	76.16	69.63	70.49	70.18	n/a	n/a	68.50	69.07	70.51	**71.55**

Table 4. Object detection performance on MS-COCO `val2017` using Faster R-CNN with FPN backbone. We use mean Average Precision (AP) and AP at IoU thresholds of 0.5 and 0.75 (AP_{50}, AP_{75}). Results for our method are based on a *single* run.

Method	ResNet-101 → ResNet-18			ResNet-101 → ResNet-50			ResNet-50 → MobileNet-v2		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Teacher	42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
Student	33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
KD [33]	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
FitNet [74]	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
DCD (ours)	**36.98**	**57.44**	**39.79**	**39.01**	**60.07**	**42.88**	**33.28**	**52.97**	**35.15**

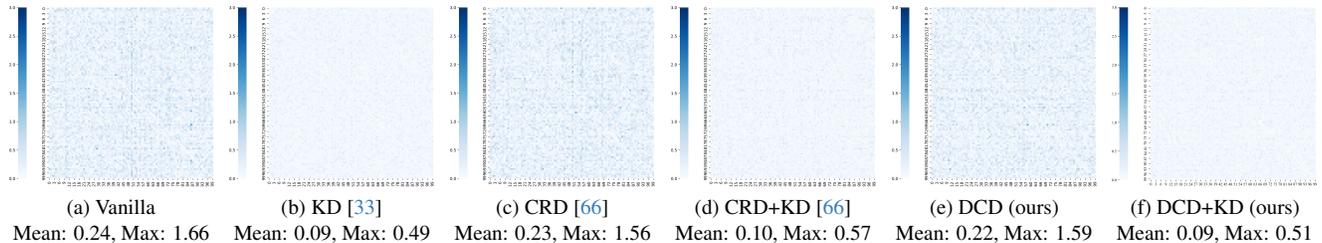| (a) Vanilla	(b) KD [33]	(c) CRD [66]	(d) CRD+KD [66]	(e) DCD (ours)	(f) DCD+KD (ours)
Mean: 0.24, Max: 1.66	Mean: 0.09, Max: 0.49	Mean: 0.23, Max: 1.56	Mean: 0.10, Max: 0.57	Mean: 0.22, Max: 1.59	Mean: 0.09, Max: 0.51

Figure 2. Correlation matrix of the average logit difference between teacher and student logits on CIFAR-100. We use WRN-40-2 as the teacher and WRN-40-1 as the student. Methods have been re-implemented according to [66].

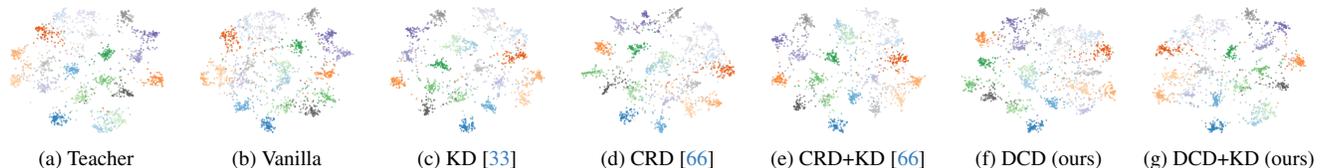| (a) Teacher	(b) Vanilla	(c) KD [33]	(d) CRD [66]	(e) CRD+KD [66]	(f) DCD (ours)	(g) DCD+KD (ours)

Figure 3. t-SNE visualizations of embeddings from teacher and student networks on CIFAR-100 (first 20 classes). We use WRN-40-2 as the teacher and WRN-40-1 as the student. Methods have been re-implemented according to [66].

Ablation on DCD. To examine the effectiveness of consistency regularization and temperature scaling, we conduct comprehensive ablation studies on CIFAR-100, as shown in Figure 4. Starting with a pure discriminative variant ($\alpha = 0$), we observe that adding consistency regularization with fixed temperature ($\alpha = 0.5$, $\tau = 0.07$, $b = 0$) improves performance across all architectures. Our proposed method with trainable temperature parameters further enhances the results, achieving improvements of up to **+1.69%** without KD and **+2.21%** with KD over the baseline. These results demonstrate that both consistency regularization and

adaptive temperature scaling contribute significantly to the method's performance.

Ablation on α. We then tested different values for the loss coefficient α of Equation (2): $\alpha = \{0.01, 0.1, 0.3, 0.5, 0.7, 1, 2, 5\}$ while setting $\beta = 1$ and $\lambda = 0$. For all the following ablations we use WRN-40-2 as teacher and WRN-16-2 as student on CIFAR-100. As shown in Figure 5a, our method remains robust across changes in α, with no significant difference in performance at low or high values. This robustness can be attributed to the adaptive
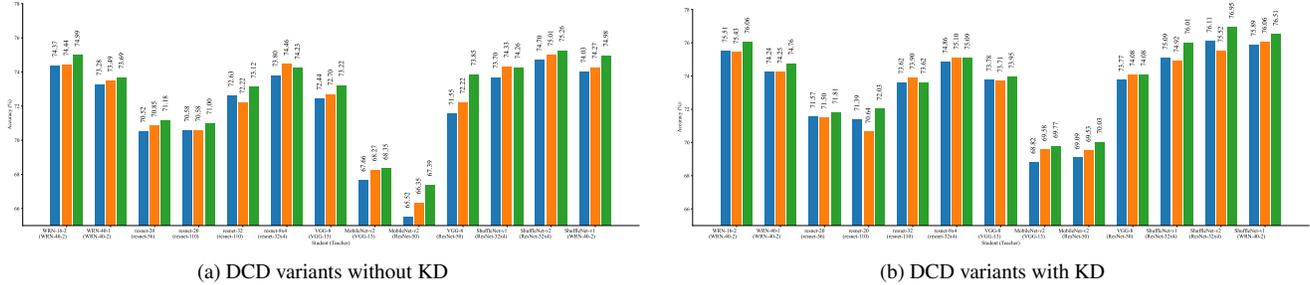
(a) DCD variants without KD

(b) DCD variants with KD

Figure 4. Ablation study results on CIFAR-100. We show results for discriminative training ($\alpha = 0$), discriminative and consistent training ($\alpha = 0.5$, $\tau = 0.07$, $b = 0$), and our proposed DCD approach ($\alpha = 0.5$, trainable τ and b). The colors correspond to each respective variant. (a) compares DCD variants without knowledge distillation, while (b) shows improvements when combined with KD. Results are based on a *single* run.

(a) Effect of varying α

(b) Effect of varying β

(c) Effect of varying λ
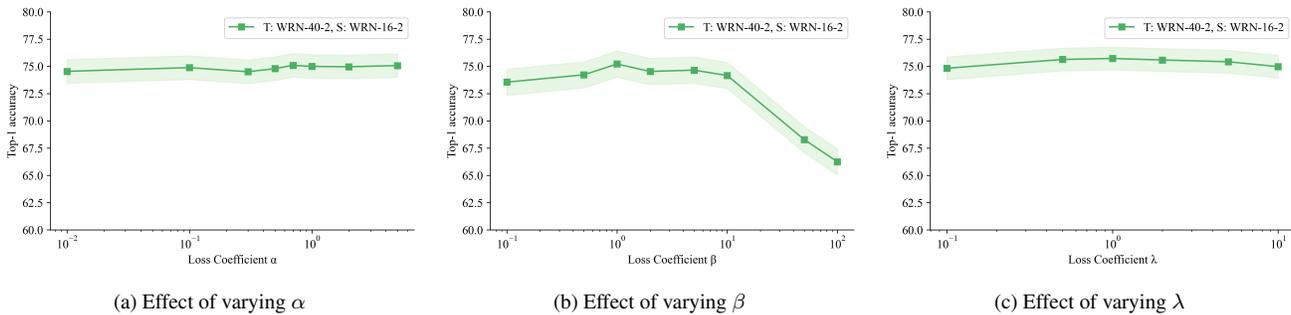
Figure 5. Ablation study results on CIFAR-100 using WRN-40-2 as the teacher and WRN-16-2 as the student. (a) Effect of the internal DCD coefficient α in Equation (2). (b) Effect of DCD loss coefficient β in Equation (7). (c) Effect of loss coefficient λ in Equation (7). Results are averaged over *five* runs.

temperature scaling mechanism, which enables automatic tuning of contrast level and logit scaling during training, providing more stable knowledge transfer.

Ablation on β. We varied β of Equation (7) from $\beta = 0.1$ to $\beta = 100$, considering values of $\beta = \{0.1, 0.5, 1, 2, 5, 10, 50, 100\}$ while fixing $\alpha = 0.5$ and $\lambda = 0$. As illustrated in Figure 5b, extremely high β values cause significant degradation in performance due to the overwhelming contribution of the DCD loss relative to other loss terms. Very low values of β also lead to a slight decrease in performance. The optimal range for β is between $\beta = 0.5$ and $\beta = 10$, suggesting that the DCD loss should be weighted similarly to other loss terms for the best results.

Ablation on λ. While λ of Equation (7) is typically set to $\lambda = 1.0$ [7, 8, 66], we tested values from $\lambda = 0.1$ to $\lambda = 10$, considering values of $\lambda = \{0.1, 0.5, 1, 2, 5, 10\}$. For these experiments we fixed $\alpha = 0.5$ and $\beta = 1$. As shown in Figure 5c, performance remains stable across all tested values, with the best results achieved at $\lambda = 1.0$ (we also found that higher values of $\lambda = 50$ and $\lambda = 100$ lead to training collapse, not shown in figure). This confirms the

common practice of setting $\lambda = 1.0$ in prior work is optimal.

5. Conclusions

We have presented DCD, 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, COCO, STL-10, and Tiny ImageNet datasets, we have demonstrated significant improvements over existing methods. Unlike previous methods such as CRD [66] that require large memory banks or WSLD [81] and IPWD [55] 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. We hope this work will inspire future research in knowledge distillation.

Acknowledgments

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

[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. 1

[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] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation, 2023. 2

[26] 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

[27] 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

[28] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation, 2023. 2

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

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 1, 3, 4

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

[32] 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

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

[34] 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

[35] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher, 2022. 2

[36] 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

[37] 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

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

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

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

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

[42] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation, 2022. 2, 5, 6

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 1, 5

[45] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation, 2023. 2

[46] 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

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

[48] 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

[49] 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

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

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

[52] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. 2

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

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

[55] Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. Respecting transfer gap in knowledge distillation, 2022. 2, 5, 6, 8

[56] 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

[57] 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. 2, 5, 6

[58] 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

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

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1, 5

[61] 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

[62] 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

[63] 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

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

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

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

[67] 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

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

[69] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 6

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

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

[72] 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

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

[74] 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

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

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

[77] 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

[78] 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

[79] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022. 2, 5

[80] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching, 2024. 2

[81] 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, 8