Discriminative and Consistent Representation Distillation

Supplementary Material

Contents

6. Algorithm

Algorithm 1 provides the pseudo-code of DCD.

Algorithm 1 Pseudocode of DCD in a PyTorch-like style.

```
# f_s, f_t: student and teacher networks
# t_s, t_t: temperature for student and teacher
# tau: learnable temperature parameter
# b: learnable bias parameter
# max_tau: maximum temperature value

for x in loader: # load a minibatch x with N samples
   s = f_s.forward(x) # student embeddings: NxC
   s = normalize(s, dim=1) # L2 normalization

   with torch.no_grad(): # no gradients for teacher
      t = f_t.forward(x) # teacher embeddings: NxC
      t = normalize(t, dim=1) # L2 normalization

   # bound temperature
   tau = tau.exp().clamp(0, max_tau)

   # contrastive loss
   logits = torch.mm(f_s, f_t.t()) * tau + b
   labels = torch.arange(N) # diagonal labels
   contrastive_loss = cross_entropy(logits, labels)

   # consistent loss
   p1 = log_softmax(logits, dim=1)
   p2 = softmax(logits, dim=0)
   consistent_loss = kl_div(p1, p2)

   # total loss
   loss = contrastive_loss + 0.5 * consistent_loss

   # SGD update: student network and parameters
   loss.backward()
   update(f_s.params)
   update(tau) # update temperature
   update(b) # update bias
```

`mm`: matrix multiplication; `softmax`: exponential normalization across dimensions; `log_softmax`: logarithm of softmax; `cross_entropy`: negative log likelihood loss; `kl_div`: Kullback-Leibler divergence between distributions

7. Implementation Details

We implement DCD in PyTorch following CRD [31][1].

7.1. Baseline Methods

We compare our approach to the following state-of-the-art methods from the literature: (1) Knowledge Distillation (KD) [12]; (2) FitNets: Hints for Thin Deep Nets [27]; (3) Attention Transfer (AT) [36]; (4) Similarity-Preserving Knowledge Distillation (SP) [32]; (5) Correlation Congruence (CC) [25]; (6) Variational Information Distillation for Knowledge Transfer (VID) [1]; (7) Relational Knowledge Distillation (RKD) [23]; (8) Learning Deep Representations with Probabilistic Knowledge Transfer (PKT) [24]; (9) Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons (AB) [11]; (10) Paraphrasing Complex Network: Network Compression via Factor Transfer (FT) [15]; (11) A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning (FSP) [35]; (12) Like What You Like: Knowledge Distill via Neuron Selectivity Transfer (NST) [14]; (13) Contrastive Representation Distillation (CRD) [31]; (14) A Comprehensive Overhaul of Feature Distillation (OFD); (15) Rethinking Soft Labels for Knowledge Distillation: A Bias-Variance Tradeoff Perspective (WSLD) [41]; (16) Respecting Transfer Gap in Knowledge Distillation (IPWD) [22]; (17) Knowledge Distillation via Softmax Regression Representation Learning (SRRL) [34]; (18) Cross-Layer Distillation with Semantic Calibration (SemCKD) [2]; (19) Distilling Knowledge via Knowledge Review (ReviewKD) [5]; (20) Knowledge Distillation with the Reused Teacher Classifier (SimKD) [3]; (21) Searching A Fast Knowledge Distillation Process via Meta Optimization (DistPro) [6]; (22) Knowledge Distillation via N-to-One Representation Matching (NORM) [19]; (23) Wasserstein Contrastive Representation Distillation (WCoRD) [4]; (24) Complementary Relation Contrastive Distillation (CRCD) [42]; (25) Feature Kernel Distillation (FKD) [8]; (26) Information Theoretic Representation (ITRD) [21]; (27) Test-Time Memory Distillation (TTM) [40]; (28) Weighted Test-Time Memory Distillation (WTTM) [40]; (29) Decoupled Knowledge Distillation (DKD) [39]; (30) Feature Correlation-based Feature Distillation (FCFD) [18]; (31) Category-aligned Knowledge Distillation (CAT-KD) [7]; (32) Distillation with Instance-wise Selective Teaching (DIST) [13]; (33) Cross-layer Transfer Knowledge Distillation (CTKD) [16].

7.2. Network Architectures

We use the following network architectures as described in [31]: (1) Wide Residual Network (WRN) [37], where WRN-d-w represents a wide ResNet with depth d and width factor w; (2) ResNet [9], where resnet-d represents a CIFAR-style ResNet with 3 groups of basic blocks having 16, 32, and 64 channels, respectively, and resnet-8 $\times 4$ and resnet-32 $\times 4$ indicate a 4-times wider network with 64, 128, and 256 channels; (3) ResNet [9], where ResNet-d represents an ImageNet-style ResNet with Bottleneck blocks and more channels; (4) MobileNet-v2 [28], using a width multiplier of 0.5 in our experiments; (5) VGG [29], where the VGG network used is adapted from its original ImageNet counterpart; and (6) ShuffleNet-v1 [38] and ShuffleNet-v2 [20], which are adapted for efficient training with input sizes of 32×32.

7.3. Optimization

All methods evaluated in our experiments use SGD with 0.9 Nesterov momentum. For CIFAR-100, we initialize the learning rate as 0.05, and decay it by 0.1 every 30 epochs after the first 150 epochs until the last 240 epoch. For MobileNet-v2, ShuffleNet-v1, and ShuffleNet-v2, we use a learning rate of 0.01 as this learning rate is optimal for these models in a grid search, while 0.05 is optimal for other models. The batch size is set to 64 for CIFAR-100, and the weight decay is set to 5×10^{-4}. For ImageNet[2] , the initial learning rate is set to 0.1 and then divided by 10 at the 30th, 60th, and 90th epochs of the total 120 training epochs. The mini-batch size is set to 256, and the weight decay is set to 1×10^{-4}. All results are reported as means over three trials, except for the results on ImageNet, which are reported in a single trial. Our implementation for MS-COCO follows the settings in [39]. We use the two-stage method Faster R-CNN [26] with FPN [17] as the feature extractors. ResNet [9] models and MobileNet-V2 [28] are selected as teachers and students. All students are trained with the $1\times$ scheduler (schedulers and task-specific loss weights follow Detectron2 [33]).

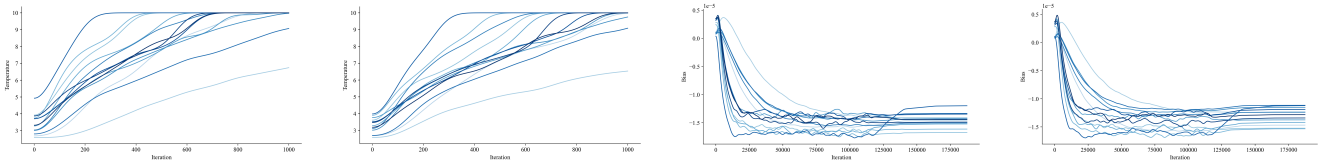
8. Additional Results

8.1. Evolution of Trainable Parameters

Figure 6 show how the trainable parameters change during training. The temperature parameter τ quickly reaches its optimal value of 10.0 in just 1,000 iterations for all teacher-student pairs, showing that the model rapidly finds the right scale for

[1]Available at: https://github.com/HobbitLong/RepDistiller.
[2]Available at: https://www.image-net.org/.

comparing features. The bias parameter β follows a different pattern, starting with a small upward spike before gradually decreasing and stabilizing between -1.2 to $-1.7 \cdot 10^{-5}$. The time it takes for the bias to stabilize depends on how similar the teacher and student architectures are.



(a) DCD τ (b) DCD+KD τ (c) DCD β (d) DCD+KD β

Figure 6. Evolution of trainable bias parameters τ and β over training iterations for both same-architecture pairs and different-architecture pairs. For parameter τ, we plot the first 1,000 training iterations.

8.2. Extended Comparison with State-of-the-Art

Table 5 and Table 6 provide a comprehensive overview of the top-1 accuracies of student networks trained with various state-of-the-art distillation techniques across a wide range of teacher-student architectural combinations. Our method benefits from its simplicity, as it has no trainable parameters, and the only hyperparameters involved are the loss coefficients.

8.3. Additional Inter-class Correlations

We present supplementary figures shown in Figure 7, which demonstrate the effectiveness of the DCD method across various student-teacher model architectures.

8.4. Additional t-SNE Visualizations

We present supplementary t-SNE visualizations shown in Figures 8 and 9, which demonstrate the effectiveness of the DCD method across various student-teacher model architectures.

9. Broader Impact

The research falls within knowledge distillation, aiming to reduce computational and energy demands for more sustainable AI deployment. However, this technique risks perpetuating teacher models' biases, particularly concerning in sensitive applications. While the algorithms' versatility enables broad application across vision tasks, this also presents dual-use concerns. Careful deployment is essential to ensure alignment with ethical guidelines and AI fairness. Future work should focus on developing robust bias detection and mitigation strategies within the distillation process. Regular auditing and transparency measures must be implemented to monitor the behavior of distilled models in real-world applications.

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. 2, 5, 6, 7, 8

[2] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration, 2021. 2, 5, 6

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier, 2022. 2, 5, 6

[4] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation, 2021. 2, 5, 6

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review, 2021. 2, 5, 6

[6] Xueqing Deng, Dawei Sun, Shawn Newsam, and Peng Wang. Distpro: Searching a fast knowledge distillation process via meta optimization, 2022. 2, 5, 6

[7] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation, 2023. 2, 5, 6

[8] Bobby He and Mete Ozay. Feature kernel distillation. In *International Conference on Learning Representations*, 2022. 2, 5, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

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

[11] 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. 2, 5, 6, 7, 8

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2, 5, 6, 7, 8

[13] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher, 2022. 2, 5, 6

[14] 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. 2, 5, 6, 7, 8

[15] 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. 2, 5, 6, 7, 8

[16] 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

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. 2

[18] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation, 2023. 2, 5, 6

[19] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching, 2023. 2, 5, 6

[20] 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. 2

[21] Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information theoretic representation distillation, 2022. 2, 5, 6

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

[23] 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. 2, 5, 6, 7, 8

[24] 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, 7, 8

[25] 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. 2, 5, 6

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 2

[27] 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. 2, 5, 6, 7, 8

[28] 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. 2

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[30] Jianhua Sun, Hao-Shu Fang, Xianghui Zhu, Jiefeng Li, and Cewu Lu. Correlation field for boosting 3d object detection in structured scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2298–2306, 2022. 7, 8

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

[32] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 2, 5, 6, 7, 8

[33] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 2

[34] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021. 2, 5, 6

[35] 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. 2, 5

[36] 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. 2, 5, 6, 7, 8

[37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. 2

[38] 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. 2

[39] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation, 2022. 2, 5, 6

[40] Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching, 2024. 2, 5, 6

[41] 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

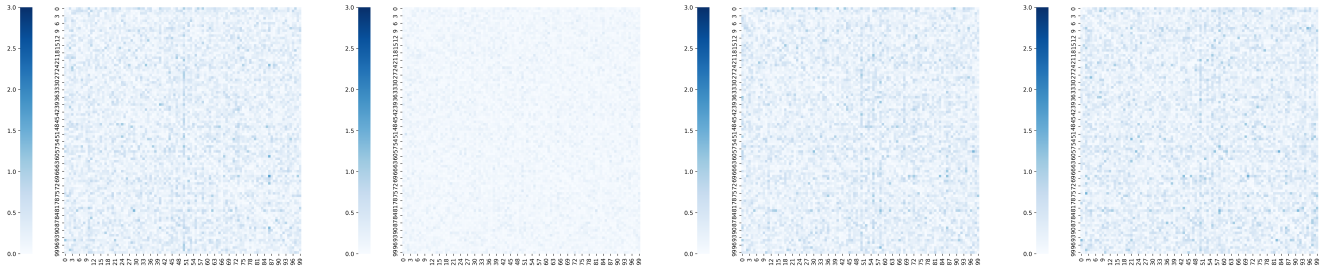
[42] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Aijun Yang, Mingzhe Rong, and Xiaohua Wang. Complementary relation contrastive distillation, 2021. 2, 5, 6

Table 5. Test top-1 accuracy (%) of student networks on CIFAR-100, comparing students and teachers of the same architecture using various distillation methods. ↑ denotes outperformance over KD and ↓ denotes underperformance.

Teacher	WRN-40-2	WRN-40-2	resnet-56	resnet-110	resnet-110	resnet-32x4	VGG-13
Student	WRN-16-2	WRN-40-1	resnet-20	resnet-20	resnet-32	resnet-8x4	VGG-8
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
KD [12]	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet [27]	73.58 (↓)	72.24 (↓)	69.21 (↓)	68.99 (↓)	71.06 (↓)	73.50 (↑)	71.02 (↓)
AT [36]	74.08 (↓)	72.77 (↓)	70.55 (↓)	70.22 (↓)	72.31 (↓)	73.44 (↑)	71.43 (↓)
SP [32]	73.83 (↓)	72.43 (↓)	69.67 (↓)	70.04 (↓)	72.69 (↓)	72.94 (↓)	72.68 (↓)
CC [25]	73.56 (↓)	72.21 (↓)	69.63 (↓)	69.48 (↓)	71.48 (↓)	72.97 (↓)	70.81 (↓)
VID [1]	74.11 (↓)	73.30 (↓)	70.38 (↓)	70.16 (↓)	72.61 (↓)	73.09 (↓)	71.23 (↓)
RKD [23]	73.35 (↓)	72.22 (↓)	69.61 (↓)	69.25 (↓)	71.82 (↓)	71.90 (↓)	71.48 (↓)
PKT [24]	74.54 (↓)	73.45 (↓)	70.34 (↓)	70.25 (↓)	72.61 (↓)	73.64 (↑)	72.88 (↓)
AB [11]	72.50 (↓)	72.38 (↓)	69.47 (↓)	69.53 (↓)	70.98 (↓)	73.17 (↓)	70.94 (↓)
FT [15]	73.25 (↓)	71.59 (↓)	69.84 (↓)	70.22 (↓)	72.37 (↓)	72.86 (↓)	70.58 (↓)
FSP [35]	72.91 (↓)	n/a	69.95 (↓)	70.11 (↓)	71.89 (↓)	72.62 (↓)	70.33 (↓)
NST [14]	73.68 (↓)	72.24 (↓)	69.60 (↓)	69.53 (↓)	71.96 (↓)	73.30 (↓)	71.53 (↓)
CRD [31]	75.48 (↑)	74.14 (↑)	71.16 (↑)	71.46 (↑)	73.48 (↑)	75.51 (↑)	73.94 (↑)
CRD+KD [31]	75.64 (↑)	74.38 (↑)	71.63 (↑)	71.56 (↑)	73.75 (↑)	75.46 (↑)	74.29 (↑)
OFD [10]	75.24 (↑)	74.33 (↑)	70.38 (↓)	n/a	73.23 (↑)	74.95 (↑)	73.95 (↑)
WSLD [41]	n/a	73.74 (↑)	71.53 (↑)	n/a	73.36 (↑)	74.79 (↑)	n/a
IPWD [22]	n/a	74.64 (↑)	71.32 (↑)	n/a	73.91 (↑)	76.03 (↑)	n/a
SRRL [34]	n/a	74.64 (↑)	n/a	n/a	n/a	75.39 (↑)	n/a
SemCKD [2]	n/a	74.41 (↑)	n/a	n/a	n/a	76.23 (↑)	n/a
ReviewKD [5]	76.12 (↑)	75.09 (↑)	71.89 (↑)	n/a	73.89 (↑)	75.63 (↑)	74.84 (↑)
SimKD [3]	n/a	75.56 (↑)	n/a	n/a	n/a	78.08 (↑)	n/a
DistPro [6]	76.36 (↑)	n/a	72.03 (↑)	n/a	73.74 (↑)	n/a	n/a
NORM [19]	75.65 (↑)	74.82 (↑)	71.35 (↑)	71.55 (↑)	73.67 (↑)	76.49 (↑)	73.95 (↑)
NORM+KD [19]	76.26 (↑)	75.42 (↑)	71.61 (↑)	72.00 (↑)	74.95 (↑)	76.98 (↑)	74.46 (↑)
NORM+CRD [19]	76.02 (↑)	75.37 (↑)	71.51 (↑)	71.90 (↑)	73.81 (↑)	76.49 (↑)	73.58 (↑)
WCoRD [4]	75.88 (↑)	74.73 (↑)	71.56 (↑)	71.57 (↑)	73.81 (↑)	75.95 (↑)	74.55 (↑)
WCoRD+KD [4]	76.11 (↑)	74.72 (↑)	71.92 (↑)	71.88 (↑)	74.20 (↑)	76.15 (↑)	74.72 (↑)
CRCD [42]	76.67 (↑)	75.95 (↑)	73.21 (↑)	72.33 (↑)	74.98 (↑)	76.42 (↑)	74.97 (↑)
FKD [8]	n/a	n/a	n/a	n/a	n/a	75.57 (↑)	73.78 (↑)
ITRD (corr) [21]	75.85 (↑)	74.90 (↑)	71.45 (↑)	71.77 (↑)	74.02 (↑)	75.63 (↑)	74.70 (↑)
ITRD (corr+mi) [21]	76.12 (↑)	75.18 (↑)	71.47 (↑)	71.99 (↑)	74.26 (↑)	76.19 (↑)	74.93 (↑)
TTM [40]	76.23 (↑)	74.32 (↑)	71.83 (↑)	71.46 (↑)	73.97 (↑)	76.17 (↑)	74.33
WTTM [40]	76.37 (↑)	74.58 (↑)	71.92 (↑)	71.67 (↑)	74.13 (↑)	76.06 (↑)	74.44
WTTM+CRD [40]	76.61 (↑)	74.94 (↑)	72.20 (↑)	72.13 (↑)	74.52 (↑)	76.65 (↑)	74.71
WTTM+ITRD [40]	76.65 (↑)	75.34 (↑)	72.16 (↑)	72.20 (↑)	74.36 (↑)	77.36 (↑)	75.13
DKD [39]	76.24 (↑)	74.81 (↑)	71.97 (↑)	n/a	74.11 (↑)	76.32 (↑)	74.68
FCFD [18]	76.34 (↑)	75.43 (↑)	71.68 (↑)	n/a	n/a	76.80 (↑)	74.86
FCFD+KD [18]	76.43 (↑)	75.46 (↑)	71.96 (↑)	n/a	n/a	76.62 (↑)	75.22
CAT-KD [7]	75.60 (↑)	74.82 (↑)	71.62 (↑)	n/a	73.62 (↑)	76.91 (↑)	74.65
DIST [13]	n/a	74.73 (↑)	71.75 (↑)	n/a	n/a	76.31 (↑)	n/a
CTKD [16]	75.45 (↑)	73.93 (↑)	71.19 (↑)	70.99 (↑)	73.52 (↑)	n/a	73.52
DCD (ours)	74.99 (↑)	73.69 (↑)	71.18 (↑)	71.00 (↑)	73.12 (↑)	74.23 (↑)	73.22 (↑)
DCD+KD (ours)	76.06 (↑)	74.76 (↑)	71.81 (↑)	72.03 (↑)	73.62 (↑)	75.09 (↑)	73.95 (↑)

Table 6. Test top-1 accuracy (%) of student networks on CIFAR-100 involving students and teachers from different architectures, using various distillation methods. ↑ denotes outperformance over KD and ↓ denotes underperformance.

Teacher	VGG-13	ResNet-50	ResNet-50	ResNet-32x4	ResNet-32x4	WRN-40-2
Student	MobileNet-v2	MobileNet-v2	VGG-8	ShuffleNet-v1	ShuffleNet-v2	ShuffleNet-v1
Teacher	74.64	79.34	79.34	79.42	79.42	75.61
Student	64.60	64.60	70.36	70.5	71.82	70.5
KD [12]	67.37	67.35	73.81	74.07	74.45	74.83
FitNet [27]	64.14 (↓)	63.16 (↓)	70.69 (↓)	73.59 (↓)	73.54 (↓)	73.73 (↓)
AT [36]	59.40 (↓)	58.58 (↓)	71.84 (↓)	71.73 (↓)	72.73 (↓)	73.32 (↓)
SP [32]	66.30 (↓)	68.08 (↑)	73.34 (↓)	73.48 (↓)	74.56 (↑)	74.52 (↓)
CC [25]	64.86 (↓)	65.43 (↓)	70.25 (↓)	71.14 (↓)	71.29 (↓)	71.38 (↓)
VID [1]	65.56 (↓)	67.57 (↑)	70.30 (↓)	73.38 (↓)	73.40 (↓)	73.61 (↓)
RKD [23]	64.52 (↓)	64.43 (↓)	71.50 (↓)	72.28 (↓)	73.21 (↓)	72.21 (↓)
PKT [24]	67.13 (↓)	66.52 (↓)	73.01 (↓)	74.10 (↑)	74.69 (↑)	73.89 (↓)
AB [11]	66.06 (↓)	67.20 (↓)	70.65 (↓)	73.55 (↓)	74.31 (↓)	73.34 (↓)
FT [15]	61.78 (↓)	60.99 (↓)	70.29 (↓)	71.75 (↓)	72.50 (↓)	72.03 (↓)
NST [14]	58.16 (↓)	64.96 (↓)	71.28 (↓)	74.12 (↑)	74.68 (↑)	76.09 (↑)
CRD [31]	69.73 (↑)	69.11 (↑)	74.3 (↑)	75.11 (↑)	75.65 (↑)	76.05 (↑)
CRD+KD [31]	69.94 (↑)	69.54 (↑)	74.58 (↑)	75.12 (↑)	76.05 (↑)	76.27 (↑)
OFD [10]	69.48 (↑)	69.04 (↑)	n/a	75.98 (↑)	76.82 (↑)	75.85 (↑)
WSLD [41]	n/a	68.79 (↑)	73.80 (↓)	75.09 (↑)	n/a	75.23 (↑)
IPWD [22]	n/a	70.25 (↑)	74.97 (↑)	76.03 (↑)	n/a	76.44 (↑)
SRRL [34]	n/a	n/a	n/a	75.18 (↑)	n/a	n/a
SemCKD [2]	n/a	n/a	n/a	n/a	77.62 (↑)	n/a
ReviewKD [5]	70.37 (↑)	69.89 (↑)	n/a	77.45 (↑)	77.78 (↑)	77.14 (↑)
SimKD [3]	n/a	n/a	n/a	77.18 (↑)	n/a	n/a
DistPro [6]	n/a	n/a	n/a	77.18 (↑)	77.54 (↑)	77.24 (↑)
NORM [19]	68.94 (↑)	70.56 (↑)	75.17 (↑)	77.42 (↑)	78.07 (↑)	77.06 (↑)
NORM+KD [19]	69.38 (↑)	71.17 (↑)	75.67 (↑)	77.79 (↑)	78.32 (↑)	77.63 (↑)
NORM+CRD [19]	69.17 (↑)	71.08 (↑)	75.51 (↑)	77.50 (↑)	77.96 (↑)	77.09 (↑)
WCoRD [4]	69.47 (↑)	70.45 (↑)	74.86 (↑)	75.40 (↑)	75.96 (↑)	76.32 (↑)
WCoRD+KD [4]	70.02 (↑)	70.12 (↑)	74.68 (↑)	75.77 (↑)	76.48 (↑)	76.68 (↑)
CRCD [42]	n/a	n/a	n/a	n/a	n/a	n/a
FKD [8]	n/a	n/a	74.61 (↑)	75 (↑)	n/a	n/a
ITRD (corr) [21]	69.97 (↑)	71.41 (↑)	75.71 (↑)	76.8 (↑)	77.27 (↑)	77.35 (↑)
ITRD (corr+mi) [21]	70.39 (↑)	71.34 (↑)	75.49 (↑)	76.91 (↑)	77.40 (↑)	77.09 (↑)
TTM [40]	68.98 (↑)	69.24 (↑)	74.87 (↑)	74.18 (↑)	76.57 (↑)	75.39
WTTM [40]	69.16 (↑)	69.59 (↑)	74.82 (↑)	74.37 (↑)	76.55 (↑)	75.42
WTTM+CRD [40]	70.30 (↑)	70.84 (↑)	75.30 (↑)	75.82 (↑)	77.04 (↑)	76.86
WTTM+ITRD [40]	70.70 (↑)	71.56 (↑)	76.00 (↑)	77.03 (↑)	77.68 (↑)	77.44
DKD [39]	69.71 (↑)	70.35 (↑)	n/a	76.45 (↑)	77.07 (↑)	76.70
FCFD [18]	70.67 (↑)	71.07 (↑)	n/a	78.12 (↑)	78.20 (↑)	77.81
FCFD+KD [18]	70.65 (↑)	71.00 (↑)	n/a	78.12 (↑)	78.18 (↑)	77.99
CAT-KD [7]	69.13 (↑)	71.36 (↑)	n/a	78.26 (↑)	78.41 (↑)	77.35
DIST [13]	n/a	68.66 (↑)	n/a	76.34 (↑)	77.35 (↑)	n/a
CTKD [16]	68.46 (↑)	68.47 (↑)	n/a	74.78 (↑)	75.31 (↑)	75.78
DCD (ours)	68.35 (↑)	67.39 (↑)	73.85 (↑)	74.26 (↑)	75.26 (↑)	74.98 (↑)
DCD+KD (ours)	69.77 (↑)	70.03 (↑)	74.08 (↑)	76.01 (↑)	76.95 (↑)	76.51 (↑)

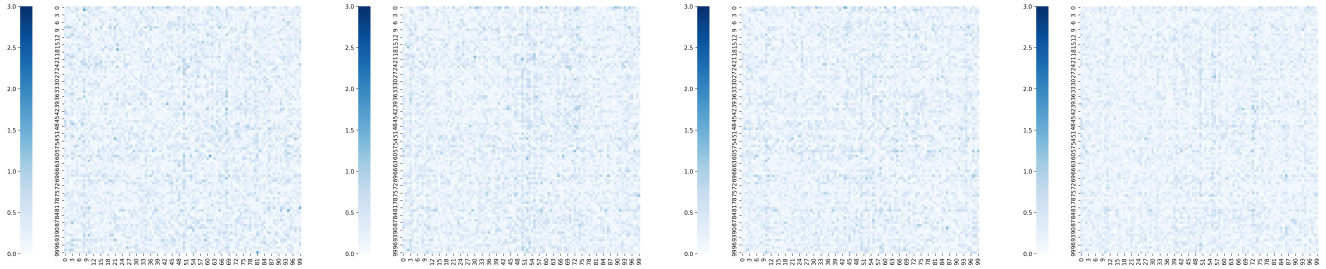


(a) Vanilla
Mean: 0.24, Max: 1.66

(b) KD [12]
Mean: 0.09, Max: 0.49

(c) FitNets [27]
Mean: 0.25, Max: 1.38

(d) AT [36]
Mean: 0.26, Max: 1.43

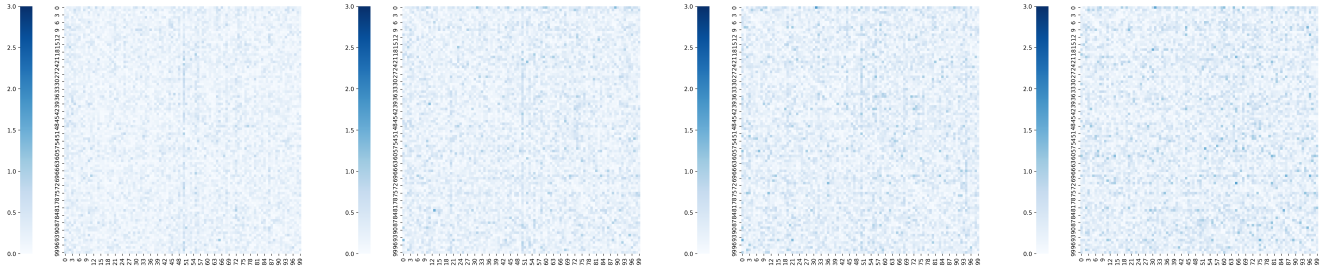


(e) SP [32]
Mean: 0.25, Max: 1.57

(f) CC [30]
Mean: 0.234, Max: 1.37

(g) VID [1]
Mean: 0.23, Max: 1.46

(h) RKD [23]
Mean: 0.21, Max: 1.24

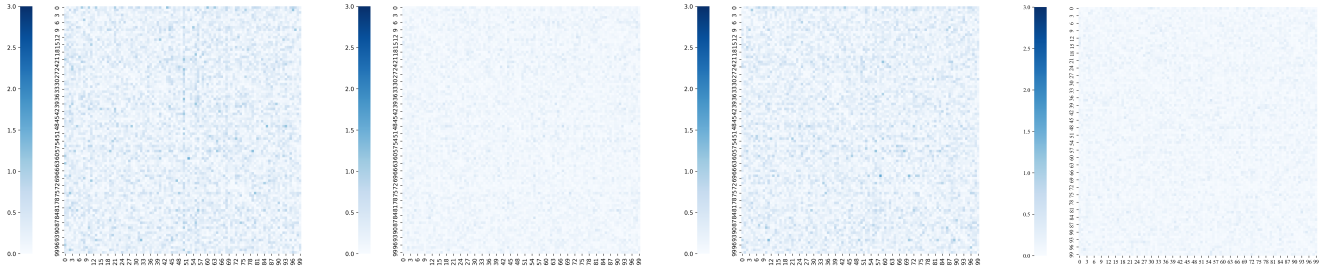


(i) PKT [24]
Mean: 0.19, Max: 1.14

(j) AB [11]
Mean: 0.24, Max: 1.49

(k) FT [15]
Mean: 0.26, Max: 1.76

(l) NST [14]
Mean: 0.28, Max: 1.65



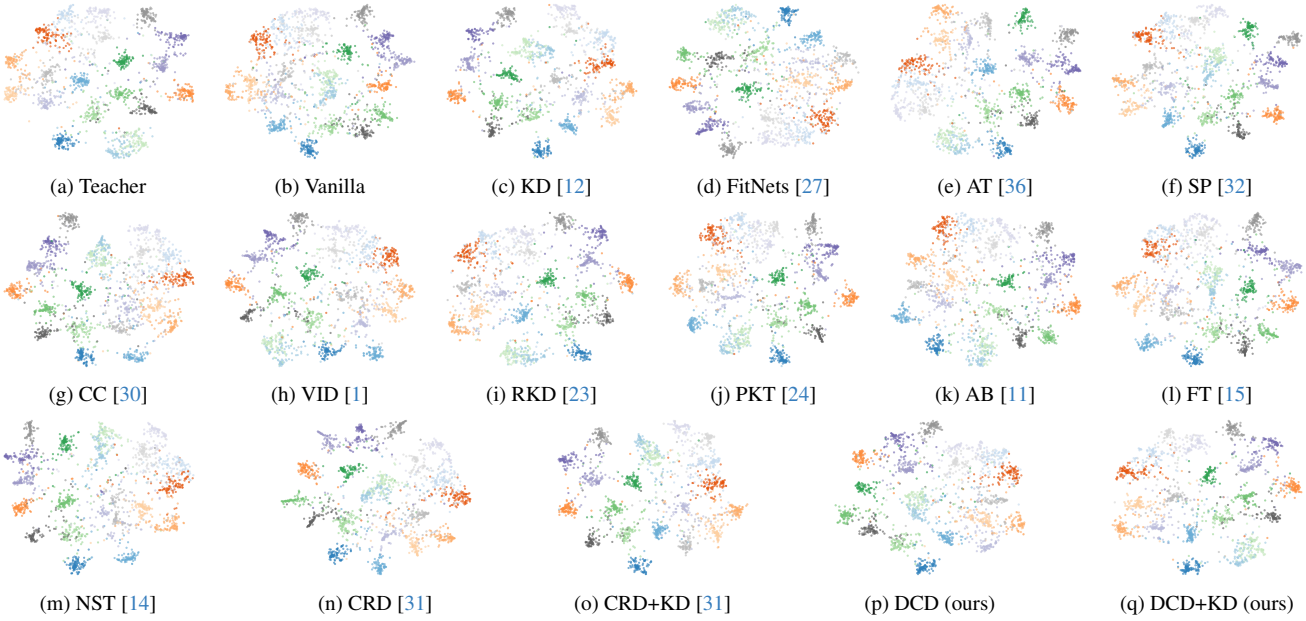
(m) CRD [31]
Mean: 0.23, Max: 1.56

(n) CRD+KD [31]
Mean: 0.10, Max: 0.57

(o) DCD (ours)
Mean: 0.21, Max: 1.19

(p) DCD+KD (ours)
Mean: 0.07, Max: 0.55

Figure 7. 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 [31].

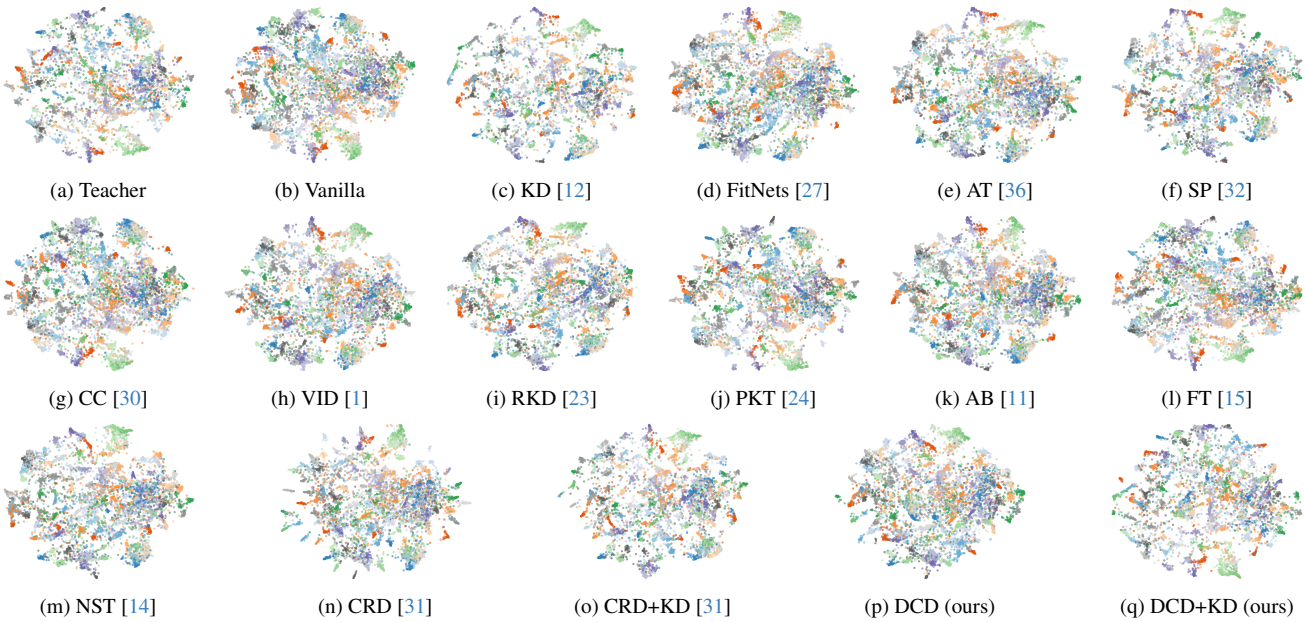


(a) Teacher (b) Vanilla (c) KD [12] (d) FitNets [27] (e) AT [36] (f) SP [32]

(g) CC [30] (h) VID [1] (i) RKD [23] (j) PKT [24] (k) AB [11] (l) FT [15]

(m) NST [14] (n) CRD [31] (o) CRD+KD [31] (p) DCD (ours) (q) DCD+KD (ours)

Figure 8. 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 [31].



(a) Teacher (b) Vanilla (c) KD [12] (d) FitNets [27] (e) AT [36] (f) SP [32]

(g) CC [30] (h) VID [1] (i) RKD [23] (j) PKT [24] (k) AB [11] (l) FT [15]

(m) NST [14] (n) CRD [31] (o) CRD+KD [31] (p) DCD (ours) (q) DCD+KD (ours)

Figure 9. t-SNE visualizations of embeddings from teacher and student networks 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 [31].