Supplemental Material

1. Datasets and Architectures

We conduct experiments on several benchmark datasets, including CIFAR100 [4], STL-10 [1], TinyImageNet and ImageNet [2]. Four architectures are used for the teacher and student networks, namely ResNet [3], VGG [8], ShuffleNet [10], MobileNet [7].

2. Implementation Details

The CIFAR100 dataset consists of 50,000 images of size 32×32 with 500 images per class and 10,000 test images. The TinyImageNet dataset is a subset of ImageNet, consisting of 100,000 images of size 64×64 from 200 classes. STL-10 consists of 5000 labeled training images from 10 classes and 100,000 unlabeled images, and a test set of 8,000 images. To keep our cross-modal transfer experiment's consistency, we down-sample each image to size 32×32. We normalized all images by channel means and standard deviations.

Following the same experimental settings of existing works [9, 5, 6], we use the SGD optimizer with momentum for all networks. For MobileNetV2 and ShuffleNet, we use a learning rate of 0.01. For the rest of the networks, the learning rate is initialized with 0.05. All the learning rates are decayed by 0.1 every 30 epochs after the first 150 epochs until the last 240 epoch. We implement the networks and training procedures in Pytorch.

3. Additional Visualization Results

We further provide more visualization results on the CIFAR-100 datasets, which is illustrated in Fig 1 and Fig 2. We observe the similar results that the proposed HKD method have similar topolopy structure with the teahcer network, which demonstrates the effectiveness of the proposed HKD method.

4. Additional Parameter Analysis

We further provide the parameter analysis on the number of layers in Graph Neural Networks. More specifically, we test the one layer(L=1) and two layer(L=2) graph neural networks. We do not set higher number of layers to avoid the over-smoothing problem.

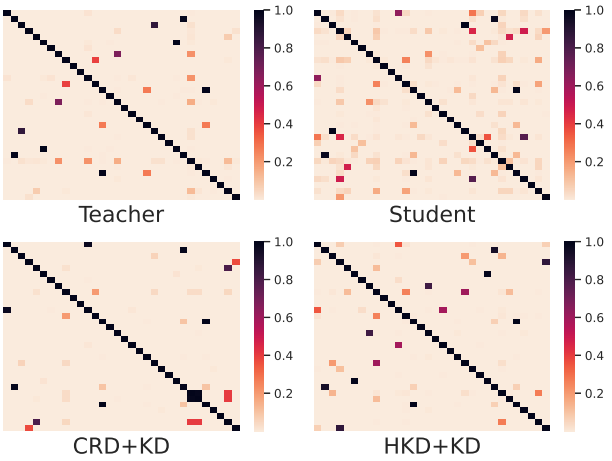

Figure 1. Visualization results on the CIFAR-100 datasets.

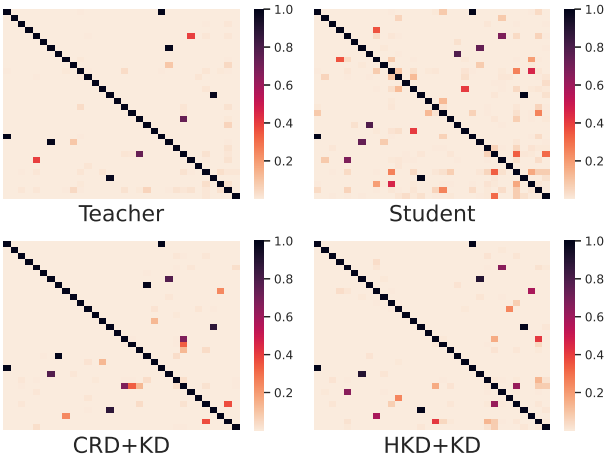

Figure 2. Visualization results on the CIFAR-100 datasets.

Tabel 1 illustrates the experimental results on the CIFAR100 dataset, from which we can observe that the HKD method is not sensitive to the number of layers.

Table 1. Parameter analysis on number of layers L

Teacher	ResNet32x4	VGG13	ResNet50
Student	ResNet8x4	MobileNetV2	VGG8
L=1	76.13 ± 0.05	70.48±0.25	74.85±0.26
L=2	76.05± 0.11	70.28±0.07	74.82±0.24

References

[1] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 215–223, 2011. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 1

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[5] 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

[6] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In Proceedings of the IEEE International Conference on Computer Vision, pages 5007–5016, 2019. 1

[7] 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. 1

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 1

[9] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. arXiv preprint arXiv:1910.10699, 2019. 1

[10] 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. 1