Frequency Attention for Knowledge Distillation
Supplemental Material

A.1. Implementation details in Section 4.1

CIFAR-100 dataset. For a fair comparison, we do experiments on standard teacher/student pairs following other distillation methods [1–3, 6]. This includes the distillation when teachers and students are in the same architecture and in different architectures. For training, we use the standard training procedure following [2, 6]. Specifically, we use SGD optimization and initialize the learning rate to 0.05 for ResNet and WRN, and 0.01 for MobileNet and ShuffleNet. All experiments are trained for 240 epochs with a batch size of 64, and we decay the learning rate by a factor of 10 at 150, 180, and 210 epochs. The weight decay is set to 5e-4. For the hyper-parameters α in distillation loss, we set α to 1 for ResNet56/ResNet20 and ResNet110/ResNet32. Meanwhile, we set α to 5 for WRN-40-2/WRN-16-2 and WRN-40-2/WRN-40-1, and set α to 20 for others.

ImageNet dataset. For ImagetNet dataset, we use ResNet34 and ResNet18 as the teacher and student pairs with the same architecture. Meanwhile, when teacher and student are in different architectures, we use ResNet50 as a teacher and MobileNet as a student. We train the model for a total of 100 epochs with batch size of 256. For optimization, we use SGD with a learning rate initialized to 0.1 and then divided by 10 for every 30 epochs. The hyper-parameter α for distillation loss is set to 1.

MS-COCO dataset. We follow the standard setting in [2, 4] for our experiments. Specifically, teacher models are pre-trained models provided by Detectron2 [5]. The training process comprises 180,000 iterations with a batch size of 8. We use SGD optimizer with the initial learning rate set to 0.01 and decaying by 10 at 120,000 and 160,000 iterations.

Teacher Student	ResNet110 ResNet32	ResNet32x4 ResNet8x4
ReviewKD [2]	73.89	75.63
FAM + ReviewKD	74.28	76.19

Table A.1. Impact of the FAM when integrating to ReviewKD [2]. The results are on the CIFAR-100 validation set.

A.2. More ablation studies

Effectiveness of the FAM module. We note that the differences between our enhanced review-based KD (refer to Figure 3 in our main paper) and the knowledge review [2] are that we use cross attention when fusing student's feature maps and place FAM after the cross attention while [2] uses ABF attention and places a convolutional layer after the attention. In Table A.1, we present the results when placing the FAM instead of a convolutional layer after the ABF in the knowledge review model [2]. The results show that the FAM module improves performance of the ReviewKD by 0.39% in which teacher/student pair is ResNet110/ResNet32, while the gain is 0.56% in which teacher/student pair is ResNet32x4/ResNet8x4. Those results confirm the effectiveness of the FAM module. In addition, by comparing the results of FAM+ReviewKD in Table A.1 (i.e., 74.28 and 76.19) with the corresponding results of FAM-KD (refer to Table 1 in main paper) (i.e., 74.45 and 76.84), we can see that in our enhanced review-based KD, cross attention is more effective than ABF attention [2] used in ReviewKD to fuse student's feature maps.

A.3. Variants of the high pass filter (HPF)

We conduct ablation studies to evaluate the performance of other HPF variants, including Gaussian HPF and Butterworth HPF. The results are shown in Table A.2, indicating that there is no significant difference between the different high-pass filter methods.

Teacher/Student	Butterworth HPF	Gaussian HPF	Ideal HPF
ResNet110/ResNet32	**74.50**	74.49	74.45
ResNet32x4/ResNet8x4	76.71	76.75	**76.84**

Table A.2. Comparisions of variant highpass filter for FAM module. The results (top-1 accuracy) are on CIFAR-100 validation set.

A.4. The values of γ_1 and γ_2 in the FAM module

γ_1 and γ_2 are two learnable parameters to control the contribution of the global and local branches in the FAM module, respectively. We report the values of these parameters after training on the ImageNet dataset with ResNet18 as a student model, and ResNet34 as a teacher model, and after training on the CIFAR-100 dataset with ResNet8x4 as a student model, and ResNet32x4 as a teacher model. In Table A.3 and Table A.4, it is clear that the global branch has much more contributions than the local branch to transformed student's feature maps.

	Stage 1	Stage 2	Stage 3	Stage 4
γ_1	1.18	3.33	3.78	3.44
γ_2	0.42	1.26	0.51	0.16

Table A.3. The values of γ_1 and γ_2 after training on the ImageNet dataset. The student model is ResNet18 and the teacher model is ResNet34. The FAM modules are applied at layer 5, layer 9, layer 13, and layer 17 of ResNet18 which are referred to as stage 1, stage 2, stage 3, and stage 4, respectively.

	Stage 1	Stage 2	Stage 3
γ_1	5.65	1.09	1.59
γ_2	0.37	0.20	0.23

Table A.4. The values of γ_1 and γ_2 after training on the CIFAR-100 dataset. The student model is ResNet8x4 and the teacher model is ResNet32x4. The FAM modules are applied at layer 3, layer 5, and layer 7 of ResNet8x4 which are referred to as stage 1, stage 2, and stage 3, respectively.

References

[1] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *CVPR*, 2021. 1

[2] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 1

[3] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 1

[4] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019. 1

[5] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[6] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 1