A Good Teacher Adapts Their Knowledge for Distillation

Supplementary Material

Teacher **Student**	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32×4 ResNet8×4	VGG13 VGG8
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
			Feature-based				
FitNet *ICLR2015*	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT *ICLR2017*	74.08	72.77	70.55	70.65	72.31	73.44	71.43
RKD *CVPR2019*	73.35	72.22	69.61	69.25	71.82	71.90	71.48
CCKD *ICCV2019*	73.56	72.21	69.63	69.48	71.48	72.97	70.71
CRD *ICLR2020*	75.48	74.14	71.16	71.46	73.48	75.51	73.94
WCoRD *CVPR2021*	75.88	74.73	71.56	71.57	73.81	75.95	74.55
ReviewKD *CVPR2021*	76.12	75.09	71.89	71.62	73.89	75.63	74.84
TaT *CVPR2022*	76.06	74.97	71.59	71.70	74.05	75.89	74.39
DPK *ICLR2023*	76.42	75.27	72.37	72.44	**74.89**	-	74.96
			Logit-based				
KD *NIPS2014*	74.92	73.54	70.66	70.67	73.08	73.33	72.98
DKD *CVPR2022*	76.24	74.81	71.97	72.01	74.11	76.32	74.68
MLKD *CVPR2023*	76.63	75.35	72.19	71.89	74.11	77.08	**75.18**
NormKD *arXiv2023*	76.40	74.84	71.40	-	73.91	76.57	74.45
STD *CVPR2024*	76.11	74.37	71.43	71.48	74.17	76.62	74.36
Ours	**76.66**	**75.62**	**72.50**	**72.68**	74.53	**77.73**	74.93

Table 10. Top-1 accuracy(%) on CIFAR-100. The best results are bolded. The best results of logit-based methods are underlined.

	fine-tune	KD
Epoch	10 - 30	240
LR	0.005	0.05
WD	0	0.0005
DR	-	0.1
Optim	sgd	sgd
τ	4	4

Table 11. Experiment Details for training and finetuning.

6. More Experiments

In Table 3 and Table 10, we compare our proposed method with existing SOTA approaches across various common teacher-student pairings on the CIFAR-100 dataset. The results demonstrate that our method consistently outperforms existing methods. Table 10 presents results where the teachers and students belong to the same network families, whereas Table 3 shows results for teachers and students from different network families.

Notably, our method employs the vanilla KD loss, a logit-based approach, unlike feature-based approaches, which require training additional networks to match the features of teachers and students. For the pairing of ResNet32×4 and WRN-40-2, the student model WRN-40-2 is relatively large, with 2.25 million parameters compared to smaller models like ResNet20, which has only 0.27 million parameters. Despite the larger size of the student model, our method still improves its perfor-

mance. These results indicate that our method can be considered a new baseline for KD, effectively alleviating the capacity gap problem.

7. More Experiment Details

In Table 11, we present detailed experimental parameters for our experiments for CIFAT-100. We fine-tune the teacher models before training the students, requiring 10 - 30 epochs for finetuning. The finetuning parameters are labeled as 'fine-tune', while the parameters for student training are denoted as 'KD'. Table 11 includes information on the number of epochs (Epoch), learning rate (LR), weight decay (WD), decay rate (DR), optimizer (Optim), and temperature (τ). During student training, the learning rate is reduced by the decay rate at epochs 150, 180, and 210. For KD, the learning rate is set to 0.01 for MobileNetV2 and ShuffleNetV2. The pseudo code for our method is shown in Algorithm 1.