A. Implementation Details

In this section, we describe the implementation details of the compared methods.

As mentioned in Sec. 4, we implemented FSKD for the pruned ResNet models according to its official codes[1], and we re-ran the official CD codes[2] to get its results. We only replaced the pruned models while keeping the hyperparameters unchanged. As for BP, we also used SGD as the optimizer, and the initial learning rate, weight decay, and momentum were 1e-3, 1e-4, and 0.9, respectively. For KD, we also used SGD and used the same learning rate, weight decay, and momentum as those in BP, in which we set the temperature $\tau = 2.0$ and the loss balancing factor was $\alpha = 0.7$.

B. Extra Results

In this part, we show results with fewer FLOPs pruned, results with connections pruned, and results on the CIFAR-10 dataset.

Pruning less FLOPs. As we reported in the main paper, we used 'Prune-C Normal' and 'Prune-C Residual' to prune the ResNet-34 model (*cf*. Tables 1 and 4, respectively). Here, we illustrated the ResNet-34 results of 'Prune-B Normal' (*cf*. Table 9) and 'Prune-B Residual' (*cf*. Table 10) on ILSVRC-2012. As shown in Table 9 and Table 10, MiR can outperform other methods regardless of pruning a large amount or a small amount of FLOPs, and the gap between MiR$_{after}$ and MiR$_{before}$ decreased when pruning less FLOPs.

Unstructured pruning results. We implemented our MiR to fine-tune the models which were pruned by an unstructured manner (*i.e*., connection pruning). We used the ℓ_1-norm weight pruning method to prune the less important weights and compared our MiR with the CD method under 90% weights pruned. Following the same settings in CD, we only pruned the weights in conv. layers, which means the weights in batch normalization layers and fully-connected layers were kept (but we can still update the weights in these layers). And 90% weights here means pruning 90% weights every conv. layer.

The representation ability was damaged because so much information was lost when pruning 90% weights with smaller ℓ_1-norm. So we tried a progressive way to prune and fine-tune weights. We pruned 20% weights and then fine-tuned with 400 iterations until we got a 90% pruned sparse model. At last, we fine-tuned the 90% pruned model with 4000 iterations. As for the results of the CD method, we re-ran the official codes and reported the mean and std. of top-1 accuracy.

CIFAR-10 results. We also conducted ResNet-56 on the CIFAR-10 dataset following the same pruning setting in CD [1]. ResNet-56 is a customized model for small datasets like CIFAR-10, and our pre-trained ResNet-56 model has 93.39%/99.87% of top-1/top-5 accuracy. We directly compared our MiR$_{before}$ with FSKD [18], FitNet [24], ThiNet [23], CP [14] and CD [1]. As for our MiR, no extra augmentation was used except for random horizontal flip, and training settings were the same as mentioned in Sec. 4.

In Table 12, we used 1/2/3/5/10/50 samples per class to fine-tune the pruned models and reported the mean and std. of top-1 accuracy under five independent trials. Results marked with * were copied from CD. As shown in Table 12, our MiR worked well in all cases.

[1] https://github.com/LTH14/FSKD

[2] https://github.com/haolibai/Cross-Distillation

Table 9. Mean and standard deviation of top-1/top-5 accuracy (%) on ILSVRC-2012. We used 'Prune-B Normal' to prune ResNet-34 and compared different methods with different training sizes. We used 50, 100, 500 random samples, and N-way-K-shot (N/K in the top row) settings. All the results were reported with five trials. **Bold** denotes the best results.

Methods	50	100	500	1000/1	1000/2	1000/3
BP	$48.3_{\pm 1.55}/76.4_{\pm 0.94}$	$49.3_{\pm 0.44}/77.4_{\pm 0.41}$	$57.9_{\pm 0.19}/82.5_{\pm 0.09}$	$62.0_{\pm 0.27}/84.5_{\pm 0.20}$	$63.7_{\pm 0.23}/85.5_{\pm 0.13}$	$64.6_{\pm 0.14}/86.0_{\pm 0.10}$
KD	$52.7_{\pm 1.43}/78.8_{\pm 0.99}$	$54.0_{\pm 0.53}/80.1_{\pm 0.47}$	$60.3_{\pm 0.10}/83.8_{\pm 0.09}$	$62.5_{\pm 0.08}/84.9_{\pm 0.08}$	$63.4_{\pm 0.19}/85.4_{\pm 0.08}$	$63.7_{\pm 0.10}/85.6_{\pm 0.07}$
FSKD	$55.8_{\pm 0.38}/80.2_{\pm 0.26}$	$59.6_{\pm 0.35}/83.1_{\pm 0.17}$	$63.7_{\pm 0.13}/85.8_{\pm 0.04}$	$64.8_{\pm 0.07}/86.4_{\pm 0.08}$	$65.3_{\pm 0.06}/86.7_{\pm 0.07}$	$65.5_{\pm 0.11}/86.8_{\pm 0.05}$
CD	$62.7_{\pm 0.28}/85.1_{\pm 0.19}$	$62.8_{\pm 0.25}/85.2_{\pm 0.15}$	$67.1_{\pm 0.06}/88.0_{\pm 0.05}$	$67.5_{\pm 0.10}/88.2_{\pm 0.04}$	$67.8_{\pm 0.10}/88.4_{\pm 0.06}$	$68.1_{\pm 0.10}/88.5_{\pm 0.05}$
MiR$_{after}$	$65.7_{\pm 0.09}/87.3_{\pm 0.07}$	$66.6_{\pm 0.07}/87.8_{\pm 0.11}$	$68.3_{\pm 0.07}/88.7_{\pm 0.08}$	$68.9_{\pm 0.03}/89.1_{\pm 0.04}$	$69.3_{\pm 0.09}/89.2_{\pm 0.06}$	$69.5_{\pm 0.05}/89.4_{\pm 0.05}$
MiR$_{before}$	**67.5**$_{\pm 0.13}$/**88.3**$_{\pm 0.06}$	**68.1**$_{\pm 0.13}$/**88.7**$_{\pm 0.07}$	**69.2**$_{\pm 0.05}$/**89.3**$_{\pm 0.08}$	**69.7**$_{\pm 0.06}$/**89.5**$_{\pm 0.03}$	**69.9**$_{\pm 0.04}$/**89.7**$_{\pm 0.07}$	**70.0**$_{\pm 0.07}$/**89.8**$_{\pm 0.03}$

Table 10. Mean and standard deviation of top-1/top-5 accuracy (%) on ILSVRC-2012. We pruned ResNet-34 using 'Prune-B Residual' (*cf*. Table 2). **Bold** denotes the best results.

Methods	50	100	500	1000/1	1000/2	1000/3
BP	$36.9_{\pm 0.94}/66.3_{\pm 1.04}$	$39.7_{\pm 0.36}/69.0_{\pm 0.25}$	$51.3_{\pm 0.21}/77.5_{\pm 0.18}$	$57.2_{\pm 0.22}/81.0_{\pm 0.13}$	$59.6_{\pm 0.19}/82.8_{\pm 0.14}$	$60.8_{\pm 0.15}/83.6_{\pm 0.06}$
KD	$42.4_{\pm 0.48}/70.1_{\pm 0.68}$	$44.9_{\pm 0.40}/72.2_{\pm 0.35}$	$54.0_{\pm 0.18}/79.0_{\pm 0.16}$	$57.2_{\pm 0.15}/81.0_{\pm 0.11}$	$58.6_{\pm 0.10}/82.0_{\pm 0.09}$	$58.9_{\pm 0.10}/82.2_{\pm 0.04}$
FSKD	$46.0_{\pm 0.49}/72.2_{\pm 0.46}$	$50.6_{\pm 0.19}/76.3_{\pm 0.14}$	$55.9_{\pm 0.25}/80.2_{\pm 0.15}$	$57.2_{\pm 0.09}/81.2_{\pm 0.14}$	$57.8_{\pm 0.11}/81.6_{\pm 0.13}$	$58.0_{\pm 0.05}/81.7_{\pm 0.08}$
MiR$_{after}$	$61.1_{\pm 0.20}/84.2_{\pm 0.24}$	$62.9_{\pm 0.18}/85.5_{\pm 0.16}$	$66.2_{\pm 0.12}/87.5_{\pm 0.07}$	$67.3_{\pm 0.06}/88.2_{\pm 0.07}$	$68.0_{\pm 0.07}/88.6_{\pm 0.05}$	$68.3_{\pm 0.03}/88.7_{\pm 0.06}$
MiR$_{before}$	**64.9**$_{\pm 0.25}$/**86.6**$_{\pm 0.21}$	**66.2**$_{\pm 0.10}$/**87.5**$_{\pm 0.12}$	**68.2**$_{\pm 0.12}$/**88.7**$_{\pm 0.05}$	**68.8**$_{\pm 0.05}$/**89.1**$_{\pm 0.06}$	**69.3**$_{\pm 0.06}$/**89.3**$_{\pm 0.03}$	**69.5**$_{\pm 0.06}$/**89.5**$_{\pm 0.03}$

Table 11. The top-1 accuracy of non-structured pruning. We pruned every layer except the first conv., and the sparsity is 0.9. **Bold** denotes the best results.

	Methods	50	100	500	1000/1	1000/2	1000/3
0.9	CD	$48.7_{\pm 0.48}$	$53.9_{\pm 0.07}$	$59.1_{\pm 0.19}$	$60.3_{\pm 0.11}$	$61.3_{\pm 0.03}$	$61.7_{\pm 0.08}$
	MiR$_{before}$	**52.7**$_{\pm 0.53}$	**55.5**$_{\pm 0.56}$	**60.9**$_{\pm 0.29}$	**63.2**$_{\pm 0.11}$	**64.3**$_{\pm 0.13}$	**64.7**$_{\pm 0.02}$

Table 12. Mean and standard deviation of top-1 accuracy (%) on CIFAR-10. We pruned ResNet-56 with 'Res-50%' used in CD [1] and used 1/2/3/5/10/50 samples per class for tuning. Results were reported with five trials. Methods marked with * were copied from CD. **Bold** denotes the best results.

Methods	1	2	3	5	10	50
FSKD*	$84.26_{\pm 1.42}$	$85.79_{\pm 1.31}$	$85.99_{\pm 1.29}$	$87.53_{\pm 1.06}$	$88.15_{\pm 0.71}$	$88.70_{\pm 0.55}$
FitNet*	$86.85_{\pm 1.91}$	$87.95_{\pm 2.13}$	$88.94_{\pm 1.85}$	$89.43_{\pm 1.60}$	$91.03_{\pm 1.14}$	$91.89_{\pm 0.87}$
ThiNet*	$88.40_{\pm 1.26}$	$88.76_{\pm 1.18}$	$88.95_{\pm 1.19}$	$89.54_{\pm 0.84}$	$90.36_{\pm 0.76}$	$90.89_{\pm 0.49}$
CP*	$88.53_{\pm 1.37}$	$88.69_{\pm 1.09}$	$88.79_{\pm 0.94}$	$89.39_{\pm 0.80}$	$89.91_{\pm 0.69}$	$90.45_{\pm 0.43}$
CD*	$89.00_{\pm 1.59}$	$89.45_{\pm 1.43}$	$89.56_{\pm 1.32}$	$90.14_{\pm 1.19}$	$90.82_{\pm 0.79}$	$91.24_{\pm 0.33}$
MiR$_{before}$	**89.27**$_{\pm 0.21}$	**90.43**$_{\pm 0.12}$	**90.70**$_{\pm 0.15}$	**91.14**$_{\pm 0.23}$	**91.57**$_{\pm 0.14}$	**92.16**$_{\pm 0.11}$