

# Supplementary Materials: MDR: Multi-stage Decoupled Relational Knowledge Distillation with Adaptive Stage Selection

Anonymous Authors

## 1 IMPLEMENTATIONS

**Training details.** On CIFAR-100, the batch size and initial learning rate are set to 64 and 0.05. We train the models for 240 epochs in total with SGD optimizer, and decay the learning rate by 0.1 at 150, 180, and 210 epochs. The weight decay and the momentum are set to  $5e-4$  and 0.9. We set  $\tau$  in  $\mathcal{L}_{kd}$  for  $\mathcal{P}$  to be 4,  $\bar{\mathcal{P}}$  to be 0.75,  $\tau$  in  $\mathcal{L}_{ang}$  and  $\mathcal{L}_{len}$  to be 0.5. We set  $\lambda_1 = 0.1$ ,  $\lambda_2 = 2.7$ ,  $\lambda_3 = 300$ ,  $\lambda_4 = 1.0$ . We conduct experiments on one Tesla T4 GPU. On ImageNet, we adopt the SGD optimizer to train the student networks for 100 epochs with a batch size of 512. The initial learning rate is 0.2 and decayed by 10 when the epoch is 30, 60 and 90. The optimizer with  $5e-4$  weight decay and 0.9 momentum is adopted. We set  $\tau$  in  $\mathcal{L}_{kd}$  for  $\mathcal{P}$  to be 1,  $\bar{\mathcal{P}}$  to be 1,  $\tau$  in  $\mathcal{L}_{ang}$  and  $\mathcal{L}_{len}$  to be 0.5. We set  $\lambda_1 = 1.0$ ,  $\lambda_2 = 2.0$ ,  $\lambda_3 = 300$ ,  $\lambda_4 = 1.0$  in Eqn. ???. Our implementation on Pascal VOC for object detection follows the same setting used in [2]. Following the consistent protocol, we use trainval07 + 12 for training and test07 for evaluation. We train the model by SGD optimizer with a momentum of 0.9 and a weight decay of  $1 \times 10^{-4}$ . The initial learning rate starts at 0.01 and is decayed by a factor of 10 at the third epoch within a total of four epochs. We conduct experiments on 4 Tesla V100 GPUs. On STL-10 and TinyImageNet, we train the linear classifiers by the SGD optimizer with a momentum of 0.9, a batch size of 64 and a weight decay of 0. The initial learning rate starts at 0.1 and is decayed by a factor of 10 at 30, 60 and 90 epochs within the total 100 epochs.

To ensure a fair comparison, we use the same data augmentation MixUp and classic KD loss for all methods. For the identical network, we maintain the same training hyperparameters, including learning rate, epoch and weight decay. In particular, we align the data augmentation methods and hyperparameter configurations used in the ML-LD original code with those in our other experiments.

During the training stage of teacher’s SMs, we connect a classifier after each SM (composed of a layer of fully connected), and directly uses category information for supervised learning. In this process, except for SM and classifier, the backbone part of the network remains frozen. On CIFAR100, we train the SMs for a total of 60 epochs, with the learning rate decayed by a factor of 10 at the 30th and 45th epochs. On ImageNet, we train the SMs for 30 epochs, with the learning rate decayed by 10 at the 10th and 20th epoch. Other settings remain consistent with the distillation process.

All experiments are conducted based on the Pytorch framework.

**Architectural Design of SMs.** As discussed in the main paper, we attach one SM after each convolutional stage. The SM is composed of global average pooling(GAP) and two fully-connected(FC) layer. For training teacher SMs, we attach one FC layer for CE loss, where the input dimension is same as the dimension of SM’s output feature (e.g., 128 on CIFAR-100, 1280 on ImageNet) and the output dimension is same as the number of categories.

We illustrate the overall architecture of SMs for various networks on CIFAR-100 on Table 1 and ImageNet on Table 2, including

the family of WRN, ResNet, VGG, MobileNet and ShuffleNet. In the specific case of ResNet34→ResNet18 in the ImageNet classification task, we specifically utilize the output of the second to fourth stages for extracting relational information. Our decision is grounded on the observation that the accuracy of SMP in the first stage is less than 10%. This finding implies that the first stage lacks significant representation information necessary for effective contrastive learning.

## 2 CIFAR-100 AND IMAGENET RESULTS ON OTHER TEACHER-STUDENT PAIRS.

To complement the data presented in Table.1, we conducted additional experiments exploring different network configurations. We observed that, in comparison to network pairs with the same architecture, the distillation improvement achieved by using different network architectures was relatively minor. Furthermore, our distillation method consistently enhanced the accuracy of the student model to a level surpassing that of the teacher model. This outcome validates the effectiveness of knowledge extraction from the teacher model and its successful transfer to the students.

We further evaluated a teacher-student pair on the large-scale ImageNet, using ResNet50 as a teacher and MobileNetV1 as a student. As shown in Table 4, our MDR achieves the best performance in both Top-1 and Top-5 accuracy categories. Specifically, MDR improves the accuracy by 0.42% over SSKD for Top-1 accuracy. These ImageNet results highlight the effectiveness of MDR on large-scale datasets.

## 3 ABLATION STUDIES.

**ADSS with different distance-wise criterion.** Table 6 summarizes the effects of different distance-wise criterion. To ensure a fair comparison, different experiments are carried out under the same angle-wise adaptive stage selection strategy.

In addition to the baseline of not using distance information(w/o L), we take the All Stages’ outputs (AS) and the Penultimate Layer’s outputs (PL), introduced by RKD. We also consider relative longest length (RLL), which means selecting the stage where the length order of the sample is the highest, and the opposite for relative shortest length (RSL). Additionally, we exploit the correct class confidence for auxiliary classifier’s output (CS) to determine the best stage. Furthermore, we believe that although distance and angle are decoupled in learning, angle-wise information needs to be referred to in the judgment of the most appropriate stage of distance, so the angular information between samples is also integrated into the judgment: directly using the Decision of angle-wise adaptive stage selection strategy (DA) and use of Relative Shortest Distance (RSD).

Table. 5 shows that utilizing the RSD as the criterion for distance-wise adaptive stage selection yields the most effective performance overall. Apart from RSD, CS exhibits a positive influence on the

**Table 1: Architectural details of SMs for various networks for CIFAR-100 classification.**

Network Name	$SM_1GAP$	$SM_2GAP$	$SM_3GAP$	$SM_4GAP$	SS Modules input dim	Classifier
WRN-40-2	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[512, 256, 128, -]	(128, 100)
WRN-40-1	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[256, 128, 64, -]	(128, 100)
WRN-16-2	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[512, 256, 128, -]	(128, 100)
resnet56	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[256, 128, 64, -]	(128, 100)
resnet20	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[256, 128, 64, -]	(128, 100)
ResNet32×4	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[1024, 512, 256, -]	(128, 100)
ResNet8×4	$(8 \times 8)$	$(8 \times 8)$	$(8 \times 8)$	-	[1024, 512, 256, -]	(128, 100)
ResNet-50	$(8 \times 8)$	$(8 \times 8)$	$(4 \times 4)$	$(2 \times 2)$	[1024, 512, 1024, 2048]	(128, 100)
VGG-13	$(16 \times 16)$	$(8 \times 8)$	$(4 \times 4)$	$(4 \times 4)$	[128, 256, 512, 512]	(128, 100)
VGG-8	$(16 \times 16)$	$(8 \times 8)$	$(4 \times 4)$	$(4 \times 4)$	[128, 256, 512, 512]	(128, 100)
MobileNetV2	$(8 \times 8)$	$(8 \times 8)$	$(4 \times 4)$	$(2 \times 2)$	[48, 16, 48, 160]	(128, 100)
ShuffleNetV1	$(8 \times 8)$	$(8 \times 8)$	$(4 \times 4)$	-	[960, 480, 960, -]	(128, 100)
ShuffleNetV2	$(8 \times 8)$	$(8 \times 8)$	$(4 \times 4)$	-	[464, 232, 464, -]	(128, 100)

**Table 2: Architectural details of SMs for various networks for ImageNet classification.**

Network Name	$SM_1GAP$	$SM_2GAP$	$SM_3GAP$	$SM_4GAP$	SM's input dim	Classifier
ResNet-34	-	$(28 \times 28)$	$(14 \times 14)$	$(7 \times 7)$	[-, 128, 256, 512]	(1280, 1000)
ResNet-18	-	$(28 \times 28)$	$(14 \times 14)$	$(7 \times 7)$	[-, 128, 256, 512]	(1280, 1000)
ResNet-50	-	$(28 \times 28)$	$(14 \times 14)$	$(7 \times 7)$	[-, 512, 1024, 1024]	(1280, 1000)
MobileNetV1	-	$(14 \times 14)$	$(7 \times 7)$	$(1 \times 1)$	[-, 512, 1024, 2048]	(1280, 1000)

**Table 3: Top-1 accuracy (%) comparison of SOTA distillation methods across various teacher-student pairs on CIFAR-100 (as a supplement to Table 1). The numbers in **Bold** and underline denote the best and the second-best results, respectively.**

Teacher-Student pair	KD	FitNet	AT	RKD	CRD	DKD	ML-LD	SSKD	ReviewKD	Ours
ResNet32×4→ShuffleV1	74.52	73.76	76.37	74.00	75.66	76.68	76.81	<u>78.37</u>	77.60	<b>78.92</b>
ResNet50→VGG8	73.51	73.29	73.88	73.84	74.10	75.46	75.54	<u>76.02</u>	75.41	<b>76.52</b>
WRN40-2→ShuffleV1	75.55	76.19	77.03	75.71	77.22	76.91	76.52	77.21	<u>77.39</u>	<b>78.47</b>

majority of network pairs. Notably, RSL has attained superior outcomes on VGG13→VGG8 and ResNet50→MobileV2, albeit with greater instability compared to RSD. The other criterion demonstrate considerably smaller positive effects than RSD. Therefore, we ultimately adopt RSD as the criterion for distance-wise adaptive stage selection.

**Effect of Adaptive Stage Selection.** We extended the experiment in Fig.3. On different teacher-student pairs, we experimented

by adding only multi-stage angle-wise information (MS\_A), multi-stage angle-wise and length-wise information (MS), and adding ADSS strategy only on angle-wise information (MS + ADSS\_A), and using ADSS strategy for angle-wise and length-wise information at the same time (MS + ADSS\_A + ADSS\_D).

**Statistics on the number of samples in each stage with ADSS.** We performed a statistical analysis of the number of stage selections based on angle and distance for various teacher networks in the CIFAR-100 training set. The results are illustrated in Fig. 1.

**Table 4: Top-1 and Top-5 accuracy (%) comparisons of SOTA distillation methods on ImageNet. Part of the compared results are from [3] and [1]. \* means the result of our reproduction.**

Acc.	Teacher	Student	KD	AT	OFD	CRD	SSKD*	ReviewKD	DKD	ML-LD*	Ours
Top-1	76.16	68.87	68.58	69.56	71.25	71.37	72.48	<u>72.56</u>	72.05	72.04	<b>72.90</b>
Top-5	92.86	88.76	88.98	89.33	90.34	90.41	91.07	91.00	91.05	<u>91.10</u>	<b>91.21</b>

**Table 5: Ablations on distance-wise criterion.**

Teacher-Student	w/o L	AS	PL	RLL	RSL	CS	DA	RSD
WRN40-2→WRN40-1	76.47	76.43	76.40	76.59	76.44	76.55	76.68	<b>76.79</b>
WRN40-2→WRN16-2	76.81	76.76	76.80	76.90	76.81	76.79	76.88	<b>77.09</b>
Res56→Res20	72.34	72.33	72.31	72.30	72.39	72.55	72.42	<b>72.77</b>
Res110→Res32	74.86	74.70	74.99	75.10	75.01	74.89	74.87	<b>75.18</b>
VGG13→VGG8	75.69	75.66	75.70	75.61	<b>75.99</b>	75.74	75.67	75.97
ResNet32×4→ResNet8×4	77.60	77.65	77.69	77.75	77.73	77.50	77.59	<b>77.94</b>
ResNet32×4→ShuffleV2	78.89	78.80	79.05	79.09	78.88	79.02	79.01	<b>79.27</b>
ResNet50→MobileV2	72.37	72.21	72.31	72.41	<b>72.55</b>	72.20	72.30	72.52
ResNet32×4→ShuffleV1	78.61	78.77	78.71	78.66	78.57	78.90	78.44	<b>78.92</b>
ResNet50→VGG8	76.33	76.23	76.31	76.44	76.31	76.33	76.33	<b>76.52</b>
WRN40-2→ShuffleV1	78.23	78.29	78.26	78.31	78.44	78.21	78.38	<b>78.47</b>

**Table 6: Ablations on Adaptive Stage Selection.**

Teacher-Student	MS_A	MS	MS+ADSS_A	MS+ADSS_A+ADSS_D
WRN40-2→WRN40-1	75.97	76.10	76.68	<b>76.79</b>
WRN40-2→WRN16-2	76.23	76.40	76.92	<b>77.09</b>
Res56→Res20	71.66	71.89	72.50	<b>72.77</b>
Res110→Res32	74.08	74.31	74.92	<b>75.18</b>
VGG13→VGG8	75.13	75.21	75.66	<b>75.97</b>
ResNet32×4→ResNet8×4	76.67	76.95	77.69	<b>77.94</b>
ResNet32×4→ShuffleV2	78.61	78.69	79.11	<b>79.27</b>
ResNet50→MobileV2	71.88	71.93	72.21	<b>72.52</b>
ResNet32×4→ShuffleV1	78.45	78.51	78.84	<b>78.92</b>
ResNet50→VGG8	76.09	76.10	76.44	<b>76.52</b>
WRN40-2→ShuffleV1	77.48	77.63	78.20	<b>78.47</b>

As shown in Fig. 1a, regarding the angle-wise ADSS, notable disparities exist in the number of choices across different stages. However, a consistent trend is observed: the proportion of choices is generally higher in the initial and final stages, while relatively lower in the intermediate stages. A plausible explanation is that in the initial stage, the feature extraction capability is limited, primarily capturing shallow features of the original input. As the network

deepens, subsequent stages exhibit stronger feature extraction abilities, with a heightened focus on local characteristics. The final stage, which connects to the classifier, is essential to discern significant feature difference among distinct categories, resulting in the strongest representation capability.

By employing the distance-wise ADSS, we observed a relatively small variation in the number of samples across each stage. This

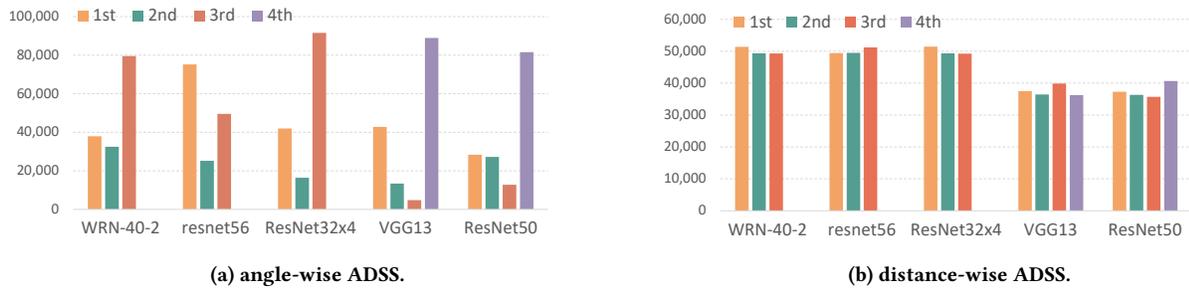


Figure 1: Statistics on the number of samples in each stage with ADSS on CIFAR-100.

Table 7: Ablations on angle-wise and length-wise weight.

$\lambda_3$	50	100	300	500	700	900
Top-1	77.71 / 72.40	77.83 / <b>72.56</b>	<b>77.94</b> / 72.52	77.81 / 72.45	77.80 / 72.33	77.66 / 72.23
$\lambda_4$	0.2	0.5	1.0	1.5	2.0	4.0
Top-1	77.69 / 72.33	77.85 / 72.46	<b>77.94</b> / <b>72.52</b>	77.80 / 72.40	77.71 / 72.33	77.60 / 72.21

finding suggests that the distance-wise representation space exhibits weaker correlation with the network depth, as compared to the angle-wise one.

**Sensitivity analysis for angle-wise and length-wise loss.** We performed a sensitivity analysis for angle-wise and length-wise loss for various teacher-student networks in the CIFAR-100. ResNet32 $\times$ 4 $\rightarrow$ ResNet8 $\times$ 4 (left) and ResNet50 $\rightarrow$ MobileNetV2 (right) are set as the teacher and the student, respectively. As shown in Table. 7, the best results were achieved when  $\lambda_3$  is around 300 and  $\lambda_4$  is 1.0. Our other experiments were conducted under this set of hyperparameter settings.

## 4 VISUALIZATIONS

In this part, we present some visualizations to show that our MDR does bridge the teacher-student gap in the relation-level. In particular, we visualize the MSE loss of relational matrix between ResNet32 $\times$ 4 and ResNet8 $\times$ 4 in Fig. 2. We find that our MDR significantly improves the similarity of angle-wise, length-wise and distance-wise relational matrix between the student and the teacher.

## REFERENCES

- [1] Ying Jin, Jiaqi Wang, and Dahua Lin. 2023. Multi-Level Logit Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24276–24285.
- [2] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. 2021. Hierarchical self-supervised augmented knowledge distillation. *arXiv preprint arXiv:2107.13715* (2021).
- [3] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 11953–11962.

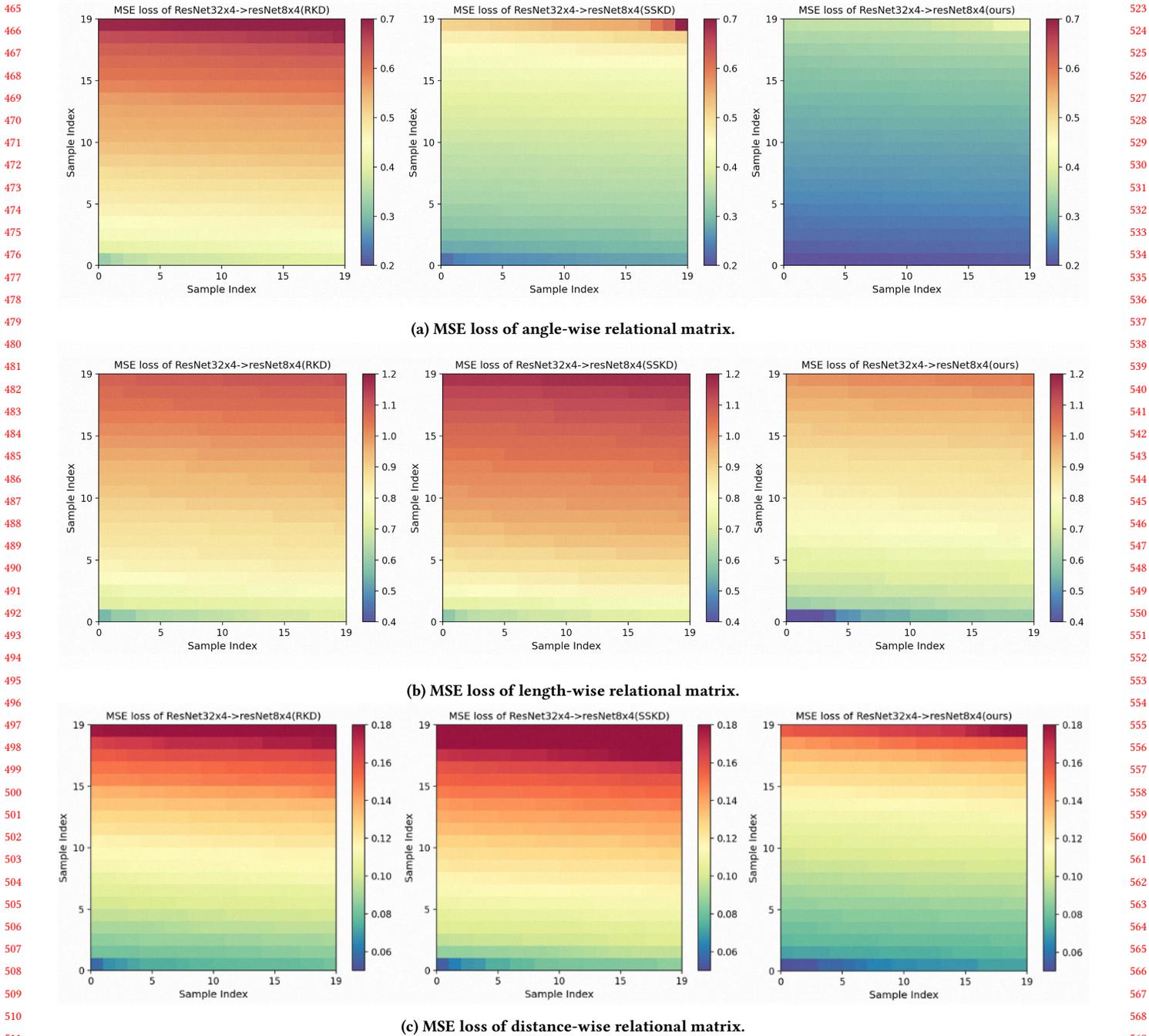


Figure 2: MSE loss of relational matrix between ResNet32x4 and ResNet8x4. We visualize the MSE loss of angle-wise (top), length-wise (middle) and distance-wise (bottom) relational matrix between the models trained by RKD (left), the models trained by SSKD (middle), and the models trained by our MDR (right). The experiments are conducted on the sampled CIFAR-100 validation set (10,000 samples). We compute relational matrix with a batch size of 25 for the penultimate stage, so these are 400 values for each experiment. For better presentation, we rank these values and organize them as the heatmap representation. The smaller the value, the more similar the matrix are.