A Impact of each contribution:

In this section, we report the effectiveness of each of our contributions. ① Reusing teacher projection head and ② symmetric contrastive learning loss. We verify this via students that are trained with ResNet-50 as a teacher.

Student Model Setting	Eff-b0	Eff-b1	Mob-v3	R-18	R-34
MoCo-V2 (Baseline)	46.8	48.4	36.2	52.2	56.8
DisCo [10]	66.5	66.6	64.4	60.6	62.5
+ ①	66.7	66.9	65.8	62.5	63.4
+ ② (RETRO)	**66.9**	**67.1**	**66.2**	**62.9**	**64.1**

Table 4: ImageNet top-1 accuracy (%) using linear classification on different strategies.

B Computational Complexity:

As illustrated in Figure 3, the computational cost of RETRO is higher compared to SEED [9] and DisCo [10] due to the additional forward propagation required for the mean student. The total number of forward propagation is 6, which is three times higher than SEED and MoCo-V2. However, these additional forward propagations are not used during inference, so there is no overhead at inference time. The results in Table 5 show that RETRO has a lower number of learnable parameters than DisCo. Therefore, the run-time overhead of RETRO is small and negligible compared to DisCo and BINGO. It should be noted that BINGO requires a KNN run to create a bag of positive samples, while RETRO is an end-to-end approach.

Method	Eff-b0	Eff-b1	Mob-v3	R-18	R-34
DisCo [10]	6.57M	8.96M	6.76M	11.91M	21.55M
RETRO	6.32M	8.71M	6.51M	11.66M	21.30M
	(↓ 0.25M)	(↓ 0.25M)	(↓ 0.25M)	(↓ 0.25M)	(↓ 0.25M)

Table 5: Comparison for the number of learnable parameters between DisCo and RETRO.

C Comparison with other Distillation:

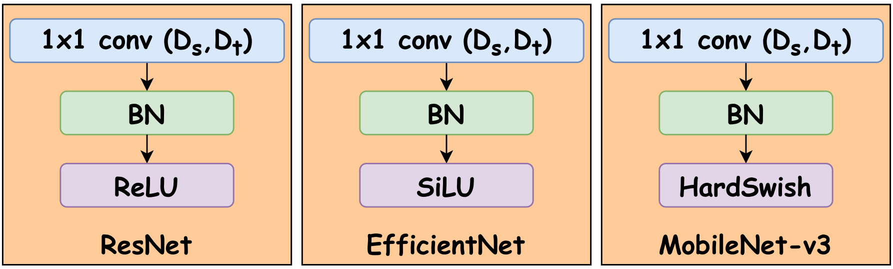
Method	Top-1
MoCo-V2 (Baseline) [14]	52.2
MoCo-V2 + KD [9]	55.3
MoCo-V2 + RKD [20]	61.6
DisCo + KD [10]	60.6
DisCo + RKD [10]	60.6
BINGO [27]	61.4
RETRO	**62.9**

Table 6: Top-1 linear classification accuracy on ImageNet utilizing various distillation techniques on the ResNet-18 student model (ResNet-50 is used as teacher model).

For further verifying the strengths of RETRO, we conducted the comparison against several different distillation strategies. We include feature-based distillation (KD) and relation-based distillation (RKD), following DisCo [10] and BINGO [27]. As shown in Table 6, RETRO shows superior performance compared with other distillation methods and surpasses them by a large margin.

D Adapter settings

We visualize the adapter architecture for different student networks in Figure 5. The adapter for ResNet, EfficientNet, and MobileNet-v3 is a 1-D convolution layer that receives output from the student encoder (D_s) and aligns that for the teacher projection head (D_t), follows by a batch normalization and a non-linear activation layer. Note that the adapter is placed right before the last pooling layer.



(a) ResNet (b) EfficientNet (c) MobileNet-v3

Figure 5: Adapter structure for different student networks.

E Frozen vs unfrozen projection head

We conduct the comparison between training RETRO for 200 epochs with frozen projection head and training for 170 epochs with frozen and 30 epochs with unfrozen projection head. As we can see from Table 7, the latter training scheme produces better performance.

Method	**Eff-b0**	**Eff-b1**	**Mob-v3**	**R-18**	**R-34**
170/30 frozen/unfrozen epochs	66.2	66.7	65.8	62.1	63.8
200 frozen epochs	66.9	67.1	66.2	62.9	64.1

Table 7: Comparison for the training scheme of RETRO.