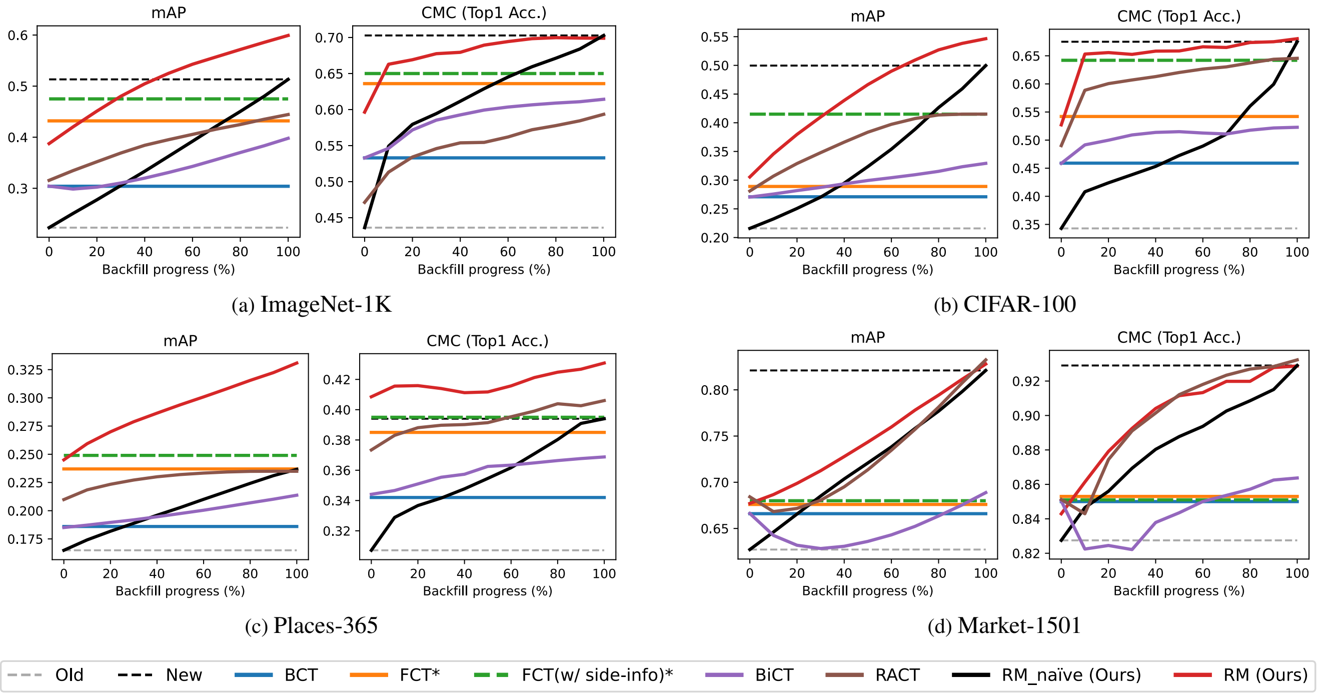
A. Additional Experimental Results

Table A2. Experimental results with heterogeneous model upgrades, from ResNet-18 (old) to ResNet-50 (new), on four standard benchmarks. *Gain* denotes relative gain that each method achieves from old model in terms of AUC_{mAP}, compared to the gain of new model. Our full algorithm (RM) outperforms all other existing approaches with significant margins.

	ImageNet-1K			CIFAR-100			Places-365			Market-1501		
	AUmAP	AuCMC	Gain	AUmAP	AuCMC	Gain	AUmAP	AuCMC	Gain	AUmAP	AuCMC	Gain
Old	22.3	43.6	0%	21.6	34.3	0%	16.5	30.7	0%	62.7	82.7	0%
New	51.3	70.3	100%	50.0	67.5	100%	23.4	39.4	100%	82.1	92.9	100%
$\text{RM}_{\text{naïve}}$ (Ours)	36.5	61.5	49%	33.8	48.9	43%	20.3	35.6	55%	72.2	82.3	49%
BCT [19]	30.4	53.3	28%	27.1	45.9	19%	18.6	34.2	30%	66.6	85.0	20%
FCT [17]	43.2	63.6	72%	28.9	54.2	26%	23.7	38.5	104%	67.6	85.3	25%
FCT (w/ side-info) [17]	47.5	65.0	87%	41.5	64.2	70%	24.9	39.5	122%	68.0	85.1	27%
BiCT [20]	33.8	58.8	40%	29.9	50.7	29%	21.4	36.7	71%	65.1	84.4	12%
RACT [26]	38.9	55.1	57%	36.9	60.9	54%	22.8	39.3	89%	73.0	89.9	53%
RM (Ours)	**50.9**	**67.3**	**99%**	**45.6**	**64.5**	**85%**	**29.2**	**41.8**	**184%**	**74.7**	**90.0**	**62%**



(a) ImageNet-1K

(b) CIFAR-100

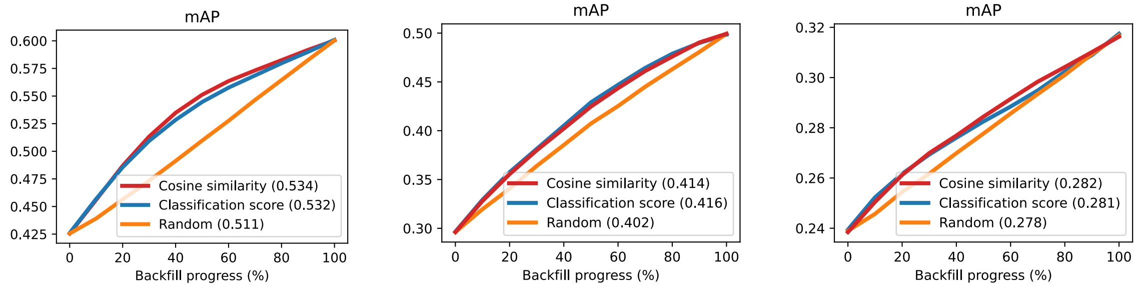
(c) Places-365

(d) Market-1501

Figure A10. mAP and CMC (Top-1 Acc.) results of our algorithms in comparison to existing approaches under heterogenous model upgrades. The numbers in the legend indicate either AUC_{mAP} or AUC_{CMC} scores.

Full experimental results with heterogenous model upgrades Table A2 and Figure A10 present the full experimental results in comparison to existing compatible learning approaches [17, 19, 20, 26] in a heterogeneous model upgrade scenario, which supplements Figure 8 of the main paper. In this scenario, *Old* and *New* models employ ResNet-18 and ResNet-50 architectures, respectively. As in the homogeneous model upgrade (Table 1 and Figure 7 of the main paper), our naïve distance rank merge framework between old and new models still provides monotonically increasing results throughout the backfilling process. Our final algorithm, dubbed as RM, significantly outperforms all other methods on all datasets, which validates its excellence in a more challenging scenario.

Backfilling strategy Another possible improvement we can make is by selecting the right samples to be backfilled first. With the results so far, even random sample selection is sufficient to achieve strong and robust retrieval merge, but there may



(a) ImageNet-1K

(b) CIFAR-100

(c) Places-365

Figure A11. Ablation study of the backfilling order on ImageNet-1K, CIFAR-100, and Places-365 datasets. To determine the backfilling order, compared to random selection, adopting either cosine similarity or classification score gives better performance at the same backfilling cost.

be room for further improvement. If we select less reliable samples to be backfilled first among gallery images, then those will benefit more than other samples from the model upgrade, improving the performance at the same backfilling cost.

We introduce two ways to measure the confidence of each old gallery sample: classification score and cosine similarity from centroid. The former uses the old classifier's final score, which reveals samples that have a lower confidence among the old gallery samples. However, it requires the classification layer of old model, which may not be applicable in some training regimes like as self-supervised learning. For the latter, we first get classwise centroid embedding by calculating the average of feature embeddings for each class, and measure the cosine similarity between each sample and its class's centroid. Low scoring samples are backfilled first based on one of these measurements, and we adopt the former in our experiments.

We conduct ablative study on the backfilling order with different strategies using our framework in Figure A11. 'Cosine similarity' indicates that the ordering is sorted by the cosine similarity between each sample and its class's centroid, while 'Classification score' denotes the order is based on the final score of classification layer in old model. In each case, samples with low scores are backfilled first. Compared to random selection, both strategies improves the average performance during backfilling, while the gain is more significant in the large-scale dataset, ImageNet.

B. Experimental Details

B.1. Datasets

ImageNet-1K [18] is a large-scale image recognition dataset introduced in ILSVRC 2012 challenge, which contains 1000 image classes with 1.2M of training images. The evaluation set has 50,000 images, 50 images per each class. Following previous works, we use the first 500 classes to train the old model, while the whole classes is used to train the new model. For evaluation, we employ the full validation split for both query and gallery sets.

CIFAR-100 [9] consists of 100 classes with 50K of training and 10K of validation images. The old model is trained with the images of the first 50 classes, and the new model utilizes the whole classes. We use the entire set of validation images as query and gallery sets for evaluation, as in ImageNet-1K.

Places-365 [29] is a large-scale scene recognition dataset, which contains 1.8M of training images with 365 scene categories. We use the first 182 categories to train the old model, while use the whole images to train the new model. The evaluation set contains 36,500 images, 100 images per each category, all of which are used for evaluation.

Market-1501 [28] is a re-identification dataset, which consists of training, gallery, and query splits, with a total of 32,886 images and 1,501 classes. The training split has 751 classes with 12,936 images. Among them, the first 375 classes are used to train the old model, while the whole 751 classes are used to train the new model. For evaluation, we follow the standard re-identification testing process with the pre-defined query and gallery splits, each of which contains 3,368 and 15,913 images, respectively.

Google Landmarks V2 (GLDv2) [25] is a large-scale landmark retrieval dataset, which consists of 1.5M of training images with 81,313 classes. The test set contains 750 query images and 760K gallery images. We randomly sample 30% of classes for training the old model while the rest of the classes are used for training the new model. We adopted mAP@100 as an evaluation metric following prior works.

B.2. Setup

We adopt the cosine distance as the distance metric for training and evaluation. The input image is resized to 224×224 for both training and evaluation on all datasets. In our frameworks, all transformation modules, $\psi(\cdot)$ and $\rho(\cdot)$, consist of 2 lightweight network blocks for (CIFAR-100, Market-1501, GLDv2) and 5 blocks for (ImageNet-1K, Places-365), where each block is composed of a sequence of operations, (Linear → BatchNorm → ReLU), except for the last block that only has a linear layer. Following the previous work [17], we set the feature dimension to 512 for the Places-365 dataset and 128 for the others. For a fair comparison, we decide the backfilling order based on the score of old classifier for all online backfilling methods. We reproduced BCT, FCT, BiCT, and RACT based on their official codes[4,5,6], and papers [17, 19, 20]. All our experiments are conducted with 4 NVIDIA A100 GPUs.

B.3. Cost Analysis

In Section 4 of the main paper, our framework introduces lightweight transformation modules for metric compatible training. Given that the backbone network is ResNet-18, which takes 1.8G of multiply-accumulate operations (MACs), the computational overhead for the transformation modules at the inference stage is only 132.8K of MACs, which is negligible ($< 0.01\%$).

[4] https://github.com/apple/ml-fct

[5] https://github.com/YantaoShen/openBCT

[6] https://github.com/TencentARC/OpenCompatible