Asymmetric Collaborative Distillation for Asymmetric Image Retrieval – Supplementary Materials –

Yi Xie[1] Huaidong Zhang[1*] Xuandi Luo[1] Yan Zhou[2] Shengfeng He[3]
[1]South China University of Technology [2]Foshan University [3]Singapore Management University

A. Experiments

A.1. Implementation Details

Our experiments are implemented in PyTorch 2.0.1 [10] with CUDA 11.8, and conducted on a single NVIDIA GeForce RTX 4090 GPU (48 GB). The training configurations for the gallery network are as follows.

(1) The gallery network is pre-trained on ImageNet [3]. In addition, the last convolutional stride of ResNet [1], ResNet101-IBN [9], and MobileNetV3-Small [2] is set to 1, following [19, 21].

(2) The data augmentation pipeline includes random horizontal flipping, random cropping with zero-padding, random erasing [22], and pixel-wise z-score normalization, following [13, 21]. The probabilities for horizontal flipping and random erasing are both set to 0.5.

(3) We employ mini-batch stochastic gradient descent [3] as the optimizer. For the high-resolution batch $(\mathcal{X}^t, \mathcal{Y}^t)$, the mini-batch size is set to 96, consisting of 16 identities with 6 images each. For $(\mathcal{X}^h, \mathcal{Y}^h)$ and $(\mathcal{X}^l, \mathcal{Y}^l)$, the mini-batch size is set to 256 when the gallery network is ResNet101(256×256), ResNet101-IBN(256×256), or ResNet101(320×160), and reduced to 128 when the gallery network is ResNet101(384×384), ResNet101(480×240), or Swin-Transformer-V2-Small(256×256).

(4) The weight decay is set to 5×10^{-4} and the momentum to 0.9.

(5) We adopt cosine annealing [7] to schedule the learning rate, which is initialized to 1×10^{-2}. The total number of training epochs is set to 20 for CUB-200-2011 [16], In-Shop [5], and SOP [8], and 10 for MSMT17 [15] and VeRi-776 [4].

(6) To ensure reproducibility, the random seed is fixed to 2024 for all experiments.

(7) The hyper-parameters α, β, and τ in Eq. (3) are set to 3, 2, and 2, respectively. The parameter p in Eq. (12) is set to 30, and γ in Eq. (13) is set to 0.05.

All experiments are conducted using the open-source D3Still codebase [20] and evaluated across a diverse set of network backbones, including ResNet [1], ResNet-IBN [9], MobileNetV3-Small [2], and Swin-Transformer V2-Small [6]. All reported results are averaged over three runs.

*Corresponding author: huaidongz@scut.edu.cn.

A.2. Hype-parameter Analysis

We conduct a comprehensive analysis of the three hyperparameters involved in our ACD and RAHO components. To evaluate how each hyperparameter affects both asymmetric and symmetric retrieval, we report two groups of metrics. "mAP (%)" and "R1 (%)" measure asymmetric image retrieval performance, where the gallery network and the query network are ResNet101(256×256) and ResNet18(64×64). In contrast, "SIR-mAP (%)" and "SIR-R1 (%)" measure symmetric image retrieval performance, where both the query and gallery features are extracted using the same ResNet101(256×256). These metrics reflect the preservation of high-resolution feature quality and the stability of the gallery network under symmetric evaluation. The following sections analyze the roles of α, β, and γ, respectively.

α ***in Eq. (3)***. The hyperparameter α controls the strength of the alignment between the gallery logits z^g and the query logits z^q. As shown in Fig. 1, when α is set too small, the gallery network receives insufficient guidance from the query network and thus fails to fully inherit the resolution-invariant semantics encoded in z^q. In contrast, an excessively large α forces the gallery network to overfit the query logits, which may distort the high-resolution semantics inherently preserved in the gallery representation. Empirically, moderate values of α offer the best trade-off: they effectively reduce the semantic gap between the two networks while maintaining the discriminative capability of the gallery network.

β ***in Eq. (3)***. The hyperparameter β controls collaborative logit alignment by weighting the KL divergence between the query logits z^q and the collaborative logits z^c. Because z^c is constructed from both query and gallery features, it naturally resides in a semantically smoother and more resolution-consistent space than z^g. Accordingly, β determines how strongly the gallery network is encouraged to follow this intermediate semantic guidance. As shown in Fig. 2, a small β underutilizes the collaborative logits, weakening their role as a semantic bridge for stabilizing cross-resolution alignment. Conversely, an overly large β causes the optimization

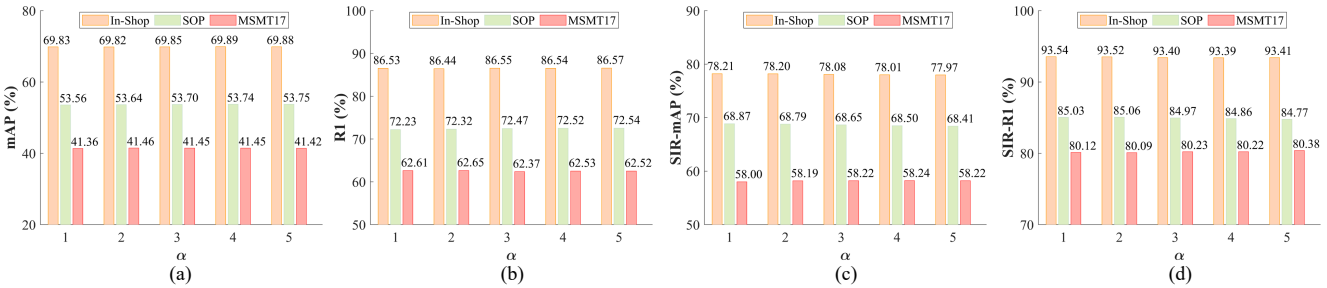

Figure 1. The performance influence of α values. (a) mAP (%) and (b) R1 (%) show asymmetric retrieval results using a ResNet101(256×256) as the gallery network and a ResNet18(64×64) as the query network. (c) SIR-mAP (%) and (d) SIR-R1 (%) show symmetric retrieval results where both query and gallery networks use the same ResNet101(256×256).

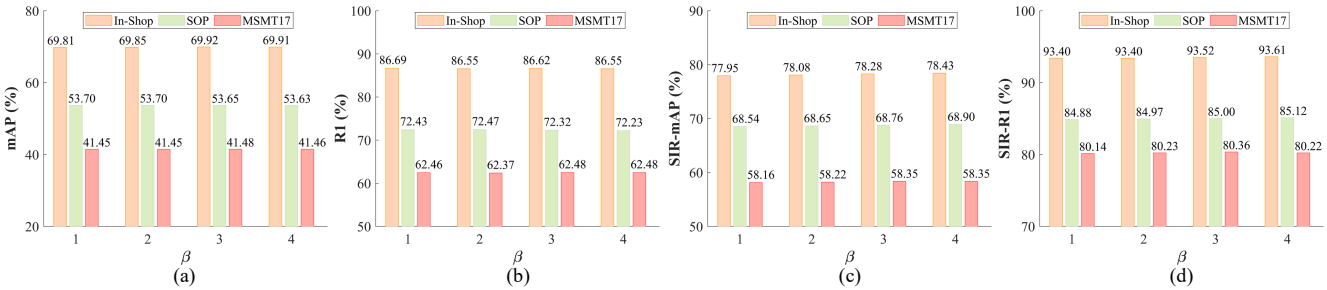

Figure 2. The performance influence of β values. (a) mAP (%) and (b) R1 (%) show asymmetric retrieval results using a ResNet101(256×256) as the gallery network and a ResNet18(64×64) as the query network. (c) SIR-mAP (%) and (d) SIR-R1 (%) show symmetric retrieval results where both query and gallery networks use the same ResNet101(256×256). (%).

to rely too heavily on z^c, diminishing the influence of the original high-resolution logits z^g and potentially biasing the gallery representation toward low-resolution semantics. Our experiments show that a balanced setting of β yields the most favourable performance, indicating that collaborative logits are most effective when they complement—rather than dominate—the standard logit alignment.

γ in Eq. (13). The hyperparameter γ regulates the learning rates assigned to the important modules of the gallery network. As shown in Fig. 3, when γ is small, these modules receive relatively low learning rates, limiting the network's ability to adapt to the low-resolution feature space and weakening cross-resolution compatibility. As γ increases, the important modules progressively obtain larger learning rates, enabling them to align more closely with the query-oriented representation space and thereby improving asymmetric retrieval performance. However, when γ becomes excessively large, this over-adaptation enhances asymmetric retrieval, it simultaneously disrupts the high-resolution feature space of the gallery network, ultimately degrading symmetric retrieval performance.

A.3. Evaluation on VeRi-776

We evaluate the effectiveness of our proposed method on VeRi-776 [4], a public challenging open-set vehicle re-identification dataset, as shown in Table 1. The dataset collects images from real-world traffic scenes using 20 surveillance cameras and contains a total of 776 vehicle identities. The training set consists of 37,746 images of 576 vehicles, while the remaining 200 identities are reserved for testing. In addition, the test set includes a probe set with 1,678 images and a gallery set containing 11,579 images.

Table 1 summarizes the performance of several baseline distillation methods, including FitNet [12], CC [11], CSD [17], RAML [14], ROP [18], and D3still [20], with and without our approach. Across all network architectures, incorporating our method consistently improves both mAP and Rank-1 accuracy. For instance, with a ResNet101(256×256) as the gallery network and a ResNet18(128×128) as the query network, our method boosts FitNet [12] by +0.77% mAP and +1.55% R1, CSD [17] by +1.49% mAP and +2.80% R1, and RAML [14] by +1.26% mAP and +1.55% R1. Similar improvements are observed for higher-resolution and alternative backbone galleries: with a ResNet101(384×384) as the gallery network, FitNet [12] improves by +0.37% mAP and +0.39% R1, CSD [17] by +1.45% mAP and +2.92% R1, and RAML [14] by +0.96% mAP and +0.83% R1; with a ResNet101-IBN (256×256) as the gallery network, FitNet [12] gains +0.36% mAP and +2.74% R1, CSD [17] gains +1.65% mAP and +2.03% R1, and RAML [14] improves by +0.93% mAP and +1.49% R1. These results demonstrate that our method effectively enhances cross-

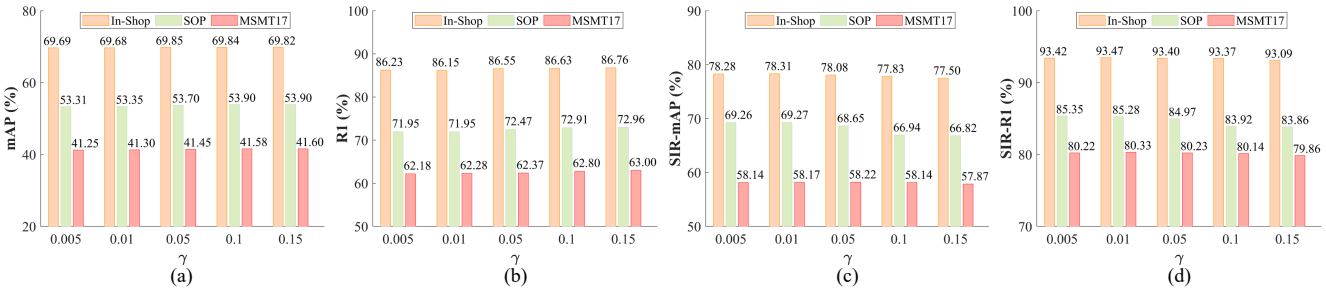

Figure 3. The performance influence of γ values. (a) mAP (%) and (b) R1 (%) show asymmetric retrieval results using a ResNet101(256×256) as the gallery network and a ResNet18(64×64) as the query network. (c) SIR-mAP (%) and (d) SIR-R1 (%) show symmetric retrieval results where both query and gallery networks use the same ResNet101(256×256).

Table 1. Performance improvement various network architectures on VeRi-776 [4].

GALLERY	ResNet101(256×256)		ResNet101(384×384)		ResNet101(256×256)		ResNet101-IBN(256×256)	
QUERY	ResNet18(128×128)		ResNet18(128×128)		MobileNetV3-Small(128×128)		ResNet18(128×128)	
ACC	mAP (%)	R1 (%)	mAP (%)	R1 (%)	mAP (%)	R1 (%)	mAP (%)	R1 (%)
FitNet [12]	67.26	82.36	67.24	81.97	60.64	73.66	68.02	82.60
+ Ours	68.03(△0.77)	83.91(△1.55)	67.61(△0.37)	82.36(△0.39)	62.72(△2.08)	76.82(△3.16)	68.38(△0.36)	85.34(△2.74)
CC [11]	66.85	81.05	66.83	80.04	61.37	76.10	67.81	83.25
+ Ours	68.22(△1.37)	83.85(△2.80)	67.51(△0.68)	80.75(△0.71)	62.80(△1.43)	79.26(△3.16)	68.48(△0.67)	83.79(△0.54)
CSD [17]	66.30	81.82	67.03	80.93	62.70	77.65	67.79	83.49
+ Ours	67.79(△1.49)	84.62(△2.80)	68.48(△1.45)	83.85(△2.92)	64.91(△2.21)	81.76(△4.11)	69.44(△1.65)	85.52(△2.03)
RAML [14]	67.75	83.13	68.37	83.02	62.32	77.29	68.81	84.62
+ Ours	69.01(△1.26)	84.68(△1.55)	69.33(△0.96)	83.85(△0.83)	64.02(△1.70)	80.51(△3.22)	69.74(△0.93)	86.11(△1.49)
ROP [18]	65.53	83.02	65.82	81.76	57.44	73.90	65.93	83.37
+ Ours	66.98(△1.45)	83.49(△0.47)	66.83(△1.01)	83.19(△1.43)	61.46(△4.02)	78.25(△4.35)	67.57(△1.64)	84.51(△1.14)
D3still [20]	67.97	83.55	67.15	80.75	63.03	77.59	68.35	84.09
+ Ours	68.53(△0.56)	84.45(△0.90)	68.23(△1.08)	82.12(△1.37)	66.11(△3.08)	82.66(△5.07)	69.06(△0.71)	85.28(△1.19)

resolution feature alignment across different gallery resolutions and backbone variants, consistently benefiting asymmetric retrieval performance.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[2] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 1

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Neural Information Processing Systems Conference*, pages 1106–1114, 2012. 1

[4] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2016. 1, 2, 3

[5] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 1

[6] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 1

[7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*, 2017. 1

[8] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 1

[9] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479, 2018. 1

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An im-

perative style, high-performance deep learning library. 2019. 1

[11] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 2, 3

[12] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of the International Conference on Learning Representations*, 2015. 2, 3

[13] Fei Shen, Jianqing Zhu, Xiaobin Zhu, Yi Xie, and Jingchang Huang. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8793–8804, 2022. 1

[14] Pavel Suma and Giorgos Tolias. Large-to-small image resolution asymmetry in deep metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1451–1460, 2023. 2, 3

[15] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018. 1

[16] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical report, California Institute of Technology, 2010. 1

[17] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9489–9498, 2022. 2, 3

[18] Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. A general rank preserving framework for asymmetric image retrieval. In *Proceedings of the International Conference on Learning Representations*, 2023. 2, 3

[19] Yi Xie, Hanxiao Wu, Fei Shen, Jianqing Zhu, and Huanqiang Zeng. Object re-identification using teacher-like and light students. In *Proceedings of the British Machine Vision Conference*, 2021. 1

[20] Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, and Shengfeng He. D3still: Decoupled differential distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17181–17190, 2024. 1, 2, 3

[21] Yi Xie, Hanxiao Wu, Jianqing Zhu, Huanqiang Zeng, and Jing Zhang. Expanding and refining hybrid compressors for efficient object re-identification. *IEEE Transactions on Image Processing*, 33:3793–3808, 2024. 1

[22] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. 1