D3still: Decoupled Differential Distillation for Asymmetric Image Retrieval

Yi Xie[1] Yihong Lin[1] Wenjie Cai[2] Xuemiao Xu[1,3,5,6*] Huaidong Zhang[1*] Yong Du[4] Shengfeng He[7]
[1]South China University of Technology [2]Independent Researcher [3]State Key Laboratory of Subtropical Building Science
[4]Ocean University of China [5]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[6]Ministry of Education Key Laboratory of Big Data and Intelligent Robot [7]Singapore Management University

Abstract

Existing methods for asymmetric image retrieval employ a rigid pairwise similarity constraint between the query network and the larger gallery network. However, these one-to-one constraint approaches often fail to maintain retrieval order consistency, especially when the query network has limited representational capacity. To overcome this problem, we introduce the Decoupled Differential Distillation (D3still) framework. This framework shifts from absolute one-to-one supervision to optimizing the relational differences in pairwise similarities produced by the query and gallery networks, thereby preserving a consistent retrieval order across both networks. Our method involves computing a pairwise similarity differential matrix within the gallery domain, which is then decomposed into three components: feature representation knowledge, inconsistent pairwise similarity differential knowledge, and consistent pairwise similarity differential knowledge. This strategic decomposition aligns the retrieval ranking of the query network with the gallery network effectively. Extensive experiments on various benchmark datasets reveal that D3still surpasses state-of-the-art methods in asymmetric image retrieval. Code is available at https://github.com/SCY-X/D3still.

1. Introduction

The predominant image retrieval methods [1, 32, 34] based on deep learning, typically involve mapping both query and gallery images into a shared feature space that is highly discriminative. Within this space, gallery images are then ranked according to their relevance to the query image. However, this feature mapping process often relies on large neural networks, which pose practical challenges for deployment on edge devices in real-world scenarios. Consequently, this necessitates uploading query images to cloud-based platforms for feature extraction, resulting in dependencies on network connectivity and additional computational overhead.

Asymmetric image retrieval [2, 33, 40, 41], has emerged as a compelling alternative, striking an effective balance between performance and efficiency. This method first utilizes a lightweight network deployed on edge devices [10], such as mobile phones, to extract features from query images. Subsequently, these query features are uploaded to the cloud platform, where they are compared for similarity with gallery features extracted offline from gallery images using a large network. Asymmetric image retrieval eliminates the need to upload query images and the dependence on online model inference in cloud platforms, thereby enhancing efficiency. Moreover, to further alleviate the computational load on the query network, recent research [33] advocates for the use of low-resolution images for feature extraction on edge devices. In this setup, while the query features are extracted from low-resolution images using a lightweight network, the gallery features are still derived from high-resolution images using a more robust, heavier network. This contrast in resolution and network capacity enables efficient processing on edge devices while maintaining the quality of feature extraction.

In asymmetric image retrieval, the primary challenge is to synchronize the embedding spaces of the query and gallery

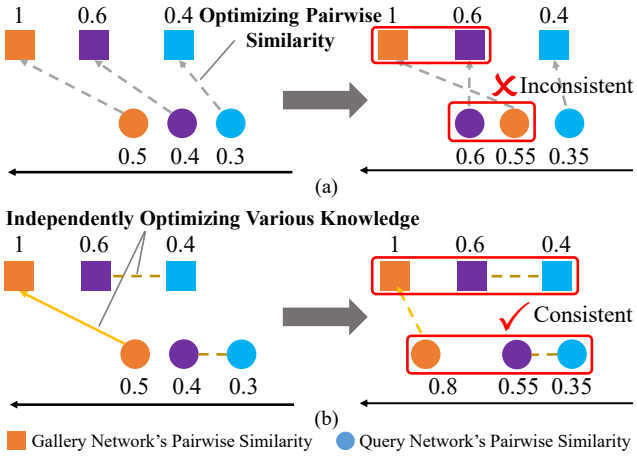

Figure 1. Illustration of existing methods and our pairwise differential distillation (D3still) framework. (a) Existing methods one-by-one optimize similarity pairs to produce inconsistent ranking results. (b) Our method optimizes the relational difference between pairwise similarities and independently transfers various knowledge to generate consistent ranking results.

*Corresponding authors: {xuemx, huaidongz}@scut.edu.cn.

networks. Addressing this, AML [2] pioneered the use of knowledge distillation (KD) alongside contrastive learning to transfer feature knowledge from the gallery network to the query network. This method is instrumental in aligning their respective feature representation spaces. Further developments in this area [33, 39] have enhanced this approach by not only transferring feature knowledge but also integrating pairwise similarity knowledge. This integration aims to enforce consistent neighbor structures across networks. As depicted in Fig. 1 (a), the process involves optimizing the distance between each pairwise similarity in the student (query) network and its corresponding similarity in the teacher (gallery) network. However, a significant issue arises due to the inherently limited representation capacity of the lightweight query network. This limitation hampers its ability to accurately replicate the pairwise similarity knowledge from the gallery network. Consequently, there is often a mismatch in the retrieval orders between the query and gallery networks, as the query network struggles to maintain fidelity to the more complex gallery network's similarity structures.

Our analysis suggests that enforcing strict feature representations on two networks with differing capacities inevitably encounters a notable upper limit. In the realm of image retrieval, the focus should shift from the intricacies of image features or similarity scores to the order in which images are returned. Therefore, we propose a novel distillation objective centered on the differential relationship between pairwise similarities. This approach allows us to focus less on the exact representations and more on maintaining consistent retrieval orders across the two networks. The underlying rationale is that instead of aiming for feature similarity between the networks, we should prioritize preserving a consistent ranking order of samples within their respective spaces. By doing so, we align more closely with the practical requirements of image retrieval, where the relative order of results often holds greater significance than the absolute similarity scores or features.

To this end, we introduce the Decoupled Differential Distillation (D3still) framework, as illustrated in Fig. 1 (b). Our approach begins with a thorough analysis of the optimization objectives in asymmetric image retrieval, leading to the conclusion that focusing on the differential between pairwise similarities in the gallery representation space is key to ensuring a consistent retrieval order between the lightweight query network and the larger gallery network. Based on this foundation, we develop a pairwise differential distillation loss function specifically designed to optimize the ranking order of pairwise similarities within the gallery domain. This involves computing pairwise similarities using features from both query and gallery images in the gallery representation space, followed by calculating the differentials between these similarities to form a pairwise similarity differential matrix. This matrix is then refined through knowledge transfer, enhancing the query network's retrieval performance.

A critical aspect of asymmetric retrieval, as highlighted in previous studies [2, 39], is the compatibility between the query and gallery networks. Consequently, we emphasize the transfer of feature representation knowledge as a dominant factor in the optimization process. To this end, we decouple the pairwise similarity differential matrix, separating feature representation knowledge from pairwise similarity differential knowledge. Additionally, we further divide the remaining matrix to distinguish between inconsistent and consistent pairwise similarity differential knowledge for distillation. This distinction enables the query network to concentrate on optimizing "hard samples", which are those whose rankings vary significantly from the gallery network, thereby improving overall retrieval accuracy.

The main contributions of this paper are threefold:

- We design a pairwise differential distillation loss function to ensure consistent retrieval order between the lightweight query network and the large gallery network.
- We propose a decoupled differential distillation (D3still) framework for asymmetric image retrieval, which transfers feature representation knowledge, inconsistent pairwise similarity differential knowledge, and consistent pairwise similarity differential knowledge to the query network.
- Extensive experiments demonstrate that our method is superior to state-of-the-art approaches.

2. Related Works

2.1. Knowledge Distillation

Knowledge distillation (KD) [9, 11, 28, 29] improves the accuracy performance of lightweight student networks by transferring knowledge from larger teacher networks. KD can be broadly categorized into two research streams: feature distillation [31, 45] and relationship distillation [25, 35].

Feature Distillation. Feature distillation methods [7, 12, 13, 31, 45] typically minimize the distance between the output features of the student network and the teacher network to guide the student network to generate similar features to those of the teacher network. For example, FitNet [31] minimizes the distance between the intermediate features of the student and teacher networks to achieve feature space alignment. Since feature knowledge can align the feature representation space between two networks, KD based asymmetric image retrieval methods [33, 39, 40] usually transfer feature representation knowledge to the query network.

Relationship Distillation. Relationship distillation methods [24, 25, 35] enforce the correlation generated by intermediate features of the student network consistent be the teacher network. For example, SPKD [35] first calculates a cosine similarity matrix between features and itself. Then, SPKD minimizes the distance between the cosine similarity matrix of the student network and the teacher network to transfer pairwise similarity knowledge. However, since

the transferring of relationship knowledge can't align the feature representation space between two networks, relationship knowledge is usually transferred together with feature knowledge in asymmetric image retrieval.

2.2. Asymmetric Image Retrieval

Ensuring embedding compatibility between gallery and query features is crucial in asymmetric image retrieval, where different networks process gallery and query examples. To address this challenge, AML [2] introduces a knowledge distillation approach to align the representation space between query and gallery networks. AML treats the lightweight query network and the large gallery network as student and teacher networks, respectively. Building upon AML's framework, subsequent works [33, 39, 40] have focused on designing distillation methods. For instance, CSD [39] minimizes the probability distribution between a contextual similarity matrix of the query and gallery networks to transfer pairwise similarity knowledge. However, due to a significant capacity gap between query and gallery networks, the query network struggles to faithfully preserve pairwise similarity knowledge from the gallery network, causing inconsistent retrieval orders between them. Moreover, ROP [40] uses the sigmoid function to replace the Heaviside step function [1] to acquire a binary smooth retrieval ranking. Then, ROP minimizes the probability distribution between the smooth retrieval ranking of the query and gallery networks. However, the binary ranking matrix may encounter the vanishing gradient problem when the predictions are close to 0 or 1. In contrast, we overcome this problem with the proposed pairwise similarity differential matrix, which can more accurately measure the retrieval ranking order.

3. Method

3.1. Formulation and Background

Assume that $\theta_q(\cdot)$ denotes a lightweight query network, which converts a low-resolution query image into a normalized d-dimensional vector. Correspondingly, $\theta_g(\cdot)$ denotes a large gallery network, which converts a high-resolution gallery image into a normalized d-dimensional vector. Given a batch of n samples $\mathcal{X} = [x_1, x_2, ..., x_n]$, we resize $\mathcal{X}$ to obtain a low-resolution sample set $\mathcal{X}^l = [x_1^l, x_2^l, ..., x_n^l]$ and a high-resolution sample set $\mathcal{X}^h = [x_1^h, x_2^h, ..., x_n^h]$, respectively. Then, we utilize query and gallery networks to extract query features $\mathcal{V}_i^q$ and gallery features $\mathcal{V}_i^g$ as follows:

$$\mathcal{V}_i^q = \theta_q(\mathcal{X}_i^l), \quad \mathcal{V}_i^g = \theta_g(\mathcal{X}_i^h), \quad i = 1, 2, ..., n. \tag{1}$$

For ranking-oriented asymmetric retrieval, the query network needs to learn ranking knowledge from the gallery network to maintain consistent retrieval orders. To accomplish this, a strict ranking distillation loss function L_{rank} can be formulated as follows:

$$L_{rank} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\substack{j=1\\l=1}}^{n}\Big(\mathcal{H}(\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q) - \mathcal{H}(\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g)\Big)^2, \tag{2}$$

where $\mathcal{S}_{i,j}^q$ denote the cosine similarity between $\mathcal{V}_i^q$ and $\mathcal{V}_j^g$; $\mathcal{S}_{i,j}^g$ denote the cosine similarity between $\mathcal{V}_i^g$ and $\mathcal{V}_j^g$; $\mathcal{H}(\cdot)$ represents a Heaviside step function [1]. However, Eq. (2) is infeasible for gradient-based optimization [1] due to the non-differentiability of the Heaviside step function. Therefore, approximating the optimization of Eq. (2) is a key challenge in enhancing asymmetric retrieval performance.

In what follows, we analyze the limitations of pairwise similarity knowledge when it is used to approximate L_{rank}. Then, we design a pairwise differential distillation loss function L_{pd} to approximate L_{rank} more appropriately. Finally, we show a complete distillation framework, which can independently transfer various knowledge to the query network.

3.2. Pairwise Similarity Knowledge Analysis

Recent asymmetric retrieval works [33, 39] design a pairwise similarity distillation loss function L_{pair} to transfer pairwise similarity knowledge to query networks as follows:

$$L_{pair} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\big(\mathcal{S}_{i,j}^q - \mathcal{S}_{i,j}^g\big)^2\Big)^{\frac{1}{2}}. \tag{3}$$

However, the pairwise similarity distillation loss function could not ensure consistent retrieval rankings between a lightweight query network and a gallery network due to the inconsistency in the capabilities between the query and gallery networks. More details are discussed as follows.

In Eq. (2), the retrieval ranking is determined by the sign of the difference between two pairwise cosine similarities (i.e., for any two pairwise cosine similarities $\mathcal{S}_{i,j}$ and $\mathcal{S}_{i,l}$, $\mathcal{H}(\mathcal{S}_{i,j} - \mathcal{S}_{i,l}) = 1$ if and only if $\mathcal{S}_{i,j} - \mathcal{S}_{i,l} > 0$). Therefore, to maintain consistency in the retrieval ranking of query and gallery networks, we should promise that: (1) $\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q > 0$ when $\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g > 0$; (2) $\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q < 0$ when $\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g < 0$; These two solutions can be unified into Eq. (4), as follows:

$$\frac{\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q}{\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g} > 0. \tag{4}$$

Obviously, Eq. (4) involves the relationship between two sample pairs (i.e., i and j, i and l). As shown in Eq. (3), there are n^2 pairs to be optimized, but each pair is independent, which could not take into account the relationship between the two pairs. To make matters worse, since the representation capacity of the query network is much lower than the gallery network, there is a significant semantic difference between query and gallery features. Thus, it is extremely difficult for a lightweight query network to generate the same similarity score as a well-trained teacher network for each pair, resulting in significant disparities in the optimization of different pairs. As a result, optimizing the pairwise similarity distillation loss function proves inadequate to ensure Eq. (4) holds, limiting the lightweight query network's performance.

3.3. Pairwise Similarity Differential Knowledge

In this work, we focus on the difference between each pairwise similarity to transfer pairwise similarity differential

knowledge to ensure the retrieval ranking consistency between the query and the gallery. In what follows, we introduce pairwise similarity differential knowledge.

Essentially, Eq. (4) can be rewritten as:

$$\frac{\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l}}{\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l}}=1+\frac{(\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l})-(\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l})}{\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l}}. \quad (5)$$

From Eq. (5), we observe that the retrieval ranking of the query network aligns with that of the gallery network when the second term of Eq. (5) exceeds -1, i.e., Eq. (5) exceeds 0. However, given the high level of indeterminacy associated with such an objective, it poses challenges in ensuring the validity of the inequality. To address this, we opt to relax the inequality, setting Eq. (5) to be as close to 1 as possible. This is done by ensuring consistency in retrieval rankings between the query and gallery networks via optimizing the second term of Eq. (5) towards 0. Notably, this relaxation is reasonable as it not only maintains the consistency of retrieval rankings in both networks but also enforces that the query network generates a similar pairwise similarity difference relationship as the gallery network. Based on the above analysis, we design a pairwise differential distillation loss function L_{pd} to approximate optimize the second terms of Eq. (5) toward 0 as follows:

$$L_{pd}=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\substack{j=1\\l=1}}^{n}\left(\frac{(\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l})-(\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l})}{m+|\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l}|}\right)^2\right)^{\frac{1}{2}}, \quad (6)$$

where m is a constant set to 0.1 to avoid a small denominator.

From Eq. (6), our method involves the relationship between two sample pairs and simultaneously optimizes two sample pairs, i.e., $(\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l})$ and $(\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l})$. Thus, our method can greatly reduce mismatched signs between two sample pairs to constrain the order of pairwise similarities between two networks to be consistent. Moreover, different from pairwise similarity knowledge that one-by-one strict constraints pairwise similarity between two networks, our method loosely constrains the relational difference of pairwise similarities between two networks to be consistent. This is because even if the query network with low capacity cannot strictly constrain the relational difference between the pairwise similarities of the two networks to be consistent, the retrieval ranking of the query network is still consistent with that of the gallery network. Consequently, in asymmetric retrieval, our pairwise similarity differential knowledge would ensure a consistent retrieval ranking between the query and gallery networks better than pairwise similarity knowledge.

3.4. Decouple Similarity Differential Knowledge

With further analysis of Eq. (6), it becomes evident that minimizing L_{pd} is equivalent to minimizing $\frac{(S^q_{i,i}-1)+(S^g_{i,l}-S^q_{i,l})}{m+|1-S^g_{i,l}|}$ when $i=j$. It is worth noting that minimizing $(S^q_{i,i}-1)$ is equivalent to narrowing the gap of the output feature between the query network and the gallery network. In other words, L_{pd} both transfers feature representation knowledge and pairwise similarity differential knowledge. Since a crucial requirement of asymmetric retrieval is that the query and gallery networks should be compatible [4], we should independently transfer feature representation knowledge to ensure its dominance, rather than coupling it with relationship knowledge, as done in previous studies. [39, 40]. Thus, we decouple the pairwise differential distillation loss function L_{pd} to form a feature distillation loss function L_f and a pure pairwise differential distillation loss function L_{rpd}. The formula for L_f is as follows:

$$L_f=\frac{1}{n}\Big(\sum_{i=1}^{n}\big(S^q_{i,i}-1\big)^2\Big)^{\frac{1}{2}}. \quad (7)$$

Table 1. Ablation study on the decoupling of feature representation knowledge and pairwise similarity differential knowledge.

METHOD	In-Shop		SOP	
	mAP (%)	R1 (%)	mAP (%)	R1 (%)
L_{pd}	0.85	0.58	25.73	35.61
L_f	61.55	75.37	49.46	66.55
L_f+L_{rpd}	**63.88**	**78.92**	**51.00**	**68.65**

As shown in Table 1, we conduct an ablation study of the decoupling of feature knowledge and pairwise similarity differential knowledge on In-Shop Clothes Retrieval (In-Shop) [19] and Stanford Online Products (SOP) [22] datasets. From Table 1, we can find that L_{pd} cannot improve the retrieval performance of the query network because feature representation knowledge in L_{pd} does not dominate, causing the feature representation space of the query network to be misaligned with the gallery network. Moreover, as the loss function L_{pd} is decoupled, the retrieval performance of the query network is significantly improved. These results demonstrate the necessity of decoupling feature representation knowledge and relationship knowledge.

Furthermore, the query network must decide whether to focus on learning pairwise similarity differential knowledge from hard or simple samples due to the limitation of representation capacity. To evaluate the sample difficulty, we calculate the ranking consistency between samples as the indicator. Thus, the pairwise relationship distillation loss function L_{rpd} can be decoupled as follows:

$$L_{rpd}=L_{irpd}+L_{crpd}, \quad (8)$$

where L_{irpd} is an inconsistent pairwise differential distillation loss function as follows:

$$L_{irpd}=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\substack{j=1,j\neq i\\l=1}}^{n}\mathcal{H}\Big(-\frac{\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l}}{\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l}}\Big)\Big(\frac{(\mathcal{S}^q_{i,j}-\mathcal{S}^q_{i,l})-(\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l})}{m+|\mathcal{S}^g_{i,j}-\mathcal{S}^g_{i,l}|}\Big)^2\right)^{\frac{1}{2}}. \quad (9)$$

And L_{crpd} is a consistent pairwise differential distillation

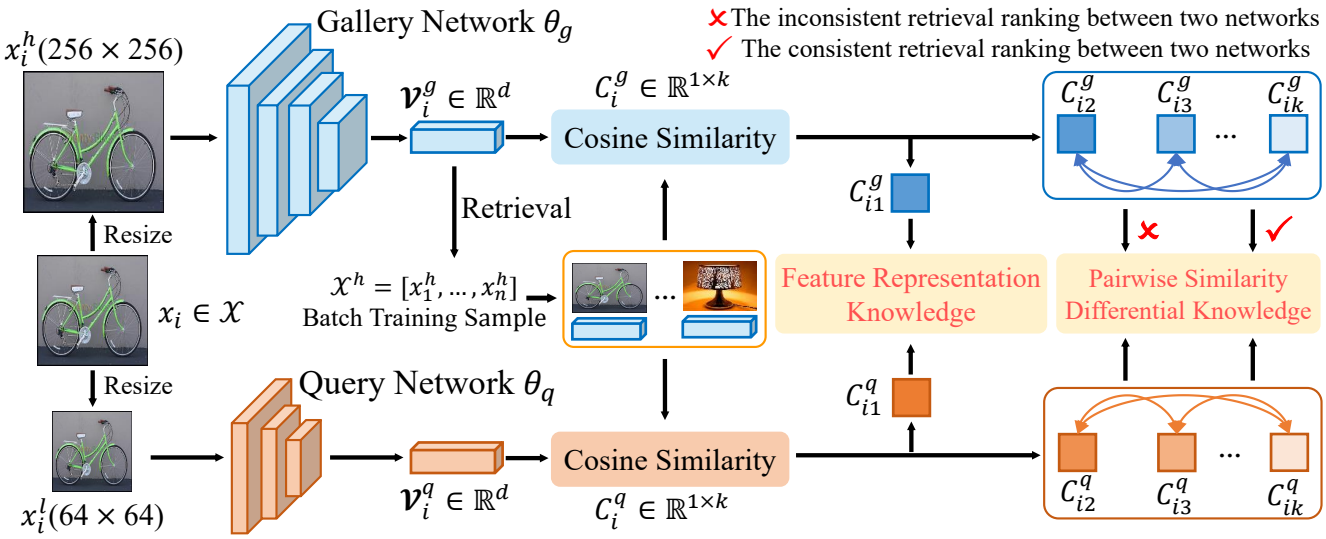

Figure 2. An overview of our decoupled differential distillation (D3still) framework. Our framework can independently transfer feature representation knowledge, inconsistent pairwise similarity differential knowledge, and consistent pairwise similarity differential knowledge.

Table 2. Ablation study about different distillation loss functions.

METHOD	In-Shop		SOP	
	mAP (%)	R1 (%)	mAP (%)	R1 (%)
$L_f + L_{irpd}$	63.38	78.01	50.57	68.04
$L_f + L_{crpd}$	63.46	78.33	50.89	68.63
$L_f + L_{rpd}$	63.88	78.92	51.00	68.65
$L_f + L_{drpd}$	**64.31**	**79.60**	**51.33**	**69.18**

loss function as follows:

$$L_{crpd} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\substack{j=1,j\neq i\\ l=1}}^{n}\mathcal{H}\left(\frac{\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q}{\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g}\right)\left(\frac{(\mathcal{S}_{i,j}^q - \mathcal{S}_{i,l}^q) - (\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g)}{m + |\mathcal{S}_{i,j}^g - \mathcal{S}_{i,l}^g|}\right)^2\right)^{\frac{1}{2}}. \tag{10}$$

The reformulation of the above Eq. (8) inspires us to investigate the individual effects of L_{iprd} and L_{cprd}. More details are discussed as follows.

Table 2 presents a detailed comparison experiment utilizing different loss functions on In-shop [19] and SOP [22]. In Table 2, the retrieval performance of L_{rpd} outperforms that of L_{crpd} and L_{irpd}, suggesting that integrating knowledge from both consistent and inconsistent samples enhances the performance of the query network. However, the limited capacity of the query network increases the likelihood of getting trapped in sub-optimal solutions dictated by the training data distribution. Furthermore, previous studies [8, 15] have indicated that emphasizing hard samples during training mitigates the risk of the model overfitting to simpler samples and improves its generalization. Thus, to further enhance the generalization ability of the query network, we propose amplifying the loss of inconsistent samples by decoupling the pairwise relationship distillation loss function L_{rpd} to obtain the decoupled pairwise differential distillation loss function L_{drpd} as follows:

$$L_{drpd} = \beta L_{irpd} + \gamma L_{crpd}, \tag{11}$$

where β and γ are hyper-parameters to balance L_{irpd} and L_{crpd}, whose default values are 0.2 and 0.1, respectively.

3.5. Decoupled Differential Distillation Framework

As shown in Fig. 2, we construct a decoupled pairwise similarity differential distillation (D3still) framework for asymmetric retrieval, which transfers feature representation knowledge and pairwise similarity differential knowledge from the gallery to query networks. Specifically, we first calculate two cosine similarity matrices G^q and G^g in the representation space of the gallery network as follows:

$$G^q = \mathcal{V}^q\mathcal{V}^g \in \mathbb{R}^{n\times n}, \quad G^g = \mathcal{V}^g\mathcal{V}^g \in \mathbb{R}^{n\times n}. \tag{12}$$

Second, since image retrieval usually returns top-k gallery images relevant to the query images, we further construct two top-k retrieval similarity matrices C^q and C^g. We acquire a retrieval result top-k index $\mathcal{R}$ based on G^g because the gallery network has been well-trained, as follows:

$$\mathcal{R} = \mathbf{argsort}(G^g, \mathbf{dim} = 2) \in \mathbb{R}^{n\times k}, \tag{13}$$

where $\mathbf{argsort}(\cdot)$ is a function that returns the top-k index corresponding to the descending order of the cosine similarity value according to the second dimension.

Then, two top-k retrieval similarity matrices $C^q \in \mathbb{R}^{n\times k}$ and $C^g \in \mathbb{R}^{n\times k}$ are constructed as follows:

$$\begin{aligned} C^q &= \mathbf{sort}(G^q, \mathbf{index} = \mathcal{R}),\\ C^g &= \mathbf{sort}(G^g, \mathbf{index} = \mathcal{R}), \end{aligned} \tag{14}$$

where $\mathbf{sort}(\cdot)$ represents a sort function that sorts the cosine similarity matrix according to the top-k index.

For feature representation knowledge, based on the top-k

retrieval similarity matrices, we design a feature distillation loss function L_f to align the representation space between query and gallery networks as follows:

$$L_f = \frac{1}{n}\Big(\sum_{i=1}^{n}\big(C^q_{i,1} - C^g_{i,1}\big)^2\Big)^{\frac{1}{2}}. \tag{15}$$

For pairwise similarity differential knowledge, based on the top-k retrieval similarity matrices, we first construct two pairwise similarity difference matrices $\mathcal{M}^q \in \mathbb{R}^{n\times(k-1)\times(k-1)}$ and $\mathcal{M}^g \in \mathbb{R}^{n\times(k-1)\times(k-1)}$ as follows:

$$\begin{aligned}\mathcal{M}^q_{i,j,l} &= C^q_{i,j+1} - C^q_{i,l+1}, \quad 1 \le j,l \le k-1.\\ \mathcal{M}^g_{i,j,l} &= C^g_{i,j+1} - C^g_{i,l+1}, \quad 1 \le j,l \le k-1.\end{aligned} \tag{16}$$

Then, the inconsistent pairwise differential distillation loss function L_{iprd} and the consistent pairwise differential distillation loss function L_{cprd} is calculated as follows:

$$L_{irpd} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{\substack{j=1\\l=1}}^{k-1}\mathcal{H}(-\frac{\mathcal{M}^q_{i,j,l}}{\mathcal{M}^g_{i,j,l}})(\frac{\mathcal{M}^q_{i,j,l}-\mathcal{M}^g_{i,j,l}}{m+|\mathcal{M}^g_{i,j,l}|})^2\Big)^{\frac{1}{2}}.$$

$$L_{crpd} = \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{\substack{j=1\\l=1}}^{k-1}\mathcal{H}(\frac{\mathcal{M}^q_{i,j,l}}{\mathcal{M}^g_{i,j,l}})(\frac{\mathcal{M}^q_{i,j,l}-\mathcal{M}^g_{i,j,l}}{m+|\mathcal{M}^g_{i,j,l}|})^2\Big)^{\frac{1}{2}}. \tag{17}$$

Finally, in the training phase, the total loss function $L_{student}$ for the query network is as follows:

$$L_{student} = \alpha L_f + \beta L_{irpd} + \gamma L_{crpd}, \tag{18}$$

where α is a hyper-parameter and used to control the contribution of L_f. The default value of α is set to 100.

4. Experiments

In this section, we present the implementation details and experimental results of our method on three publicly available image retrieval datasets: In-Shop Clothes Retrieval (In-shop) [19], Stanford Online Products (SOP) [22] and MSMT17 [37]. We first provide a brief introduction to the datasets and performance metrics. Then, we conduct ablation experiments to evaluate the effectiveness of our method and compare our method with state-of-the-art methods. Finally, we analyze the impact of hyper-parameters on performance.

4.1. Datasets and Performance Metric

In-Shop Clothes Retrieval (In-Shop) [19] is a clothes retrieval database with 72,712 images across 7,986 categories. The training set includes 3,997 classes with 25,882 images. The query set comprises 14,218 images of 3,985 classes. The gallery set has 3,985 classes with 12,612 images.

Stanford Online Products (SOP) [22] is a widely used product recognition dataset, including an extensive collection of 120,053 product images across 22,634 classes. The training set includes 59,551 training images of 11,318 classes, and the test set includes 60,502 images of 11,316 classes.

MSMT17 [37] is a widely recognized pedestrian retrieval database, including 126,441 images of 4,101 pedestrian identities captured by a total of 15 cameras (3 indoor and 12 outdoor). The training set includes 32,621 images with 1,041 unique pedestrian identities. The test set consists of 11,659 query images and a staggering 82,161 gallery images, which showcase 3,060 distinct pedestrian identities.

Table 3. Ablation study on In-shop [19] and SOP [22].

METHOD	FLOPs (G)	In-Shop		SOP	
		mAP (%)	R1 (%)	mAP (%)	R1 (%)
Gallery	12.99	81.96	95.42	72.06	86.92
L_f	0.25	61.55	75.37	49.46	66.55
$L_f + L_{pair}$	0.25	62.71	76.94	49.96	66.94
$L_f + L_{dprd}$	0.25	**64.31**	**79.60**	**51.33**	**69.18**

The **Cosine** distance between features as the retrieval algorithm, i.e., the more similar the gallery image, the higher the ranking. The mean average precision (mAP) [23, 30, 33] and rank-1 identification rate (R1) [16, 33, 38] are both applied to evaluate the retrieval accuracy performance. The floating-point of operations (FLOPs) is used to measure the models' computational complexity.

4.2. Implementation Details

The software tools are Pytorch 2.0.1 [26], CUDA 11.8. The hardware device is one GeForce RTX 4090 GPU 24G. The network training configurations are as follows. (1) We use ImageNet pre-trained ResNet101 [6] and ResNet18 [6] as the gallery network and the query network, respectively. Besides, the last stride of ResNet is set to 1, as done in [42, 44]. (2) The input image resolutions for the query network and the gallery network are 64×64 and 256×256, respectively. (3) The data augmentation includes z-score normalization, random cropping, random erasing [36, 49], and random horizontal flip operations, as done in [3, 43]. The probabilities of horizontal flip and random erasing operations are both set to 0.5. (4) We use the mini-batch stochastic gradient descent method [14] as an optimizer. The mini-batch size is set to 512. (5) We set weight decays [46–48] as 5×10^{-4} while momentums as 0.9. (6) The cosine annealing strategy [17, 18, 21] and linearly warmed strategy [5] are applied to adjust learning rates. Specifically, the learning rates are initialized to 1×10^{-3}, then linearly warmed up to 1×10^{-2} in the first 10 epochs. The drop point for learning rates is the 40-th epoch and the total training epoch is 120. (7) The hyper-parameter k in Eq. (13) is set to 10.

4.3. Ablation Experiments

As shown in Table 3, we conduct ablation experiments on In-shop [19] and SOP [22] datasets. "Gallery" denotes that we directly evaluate the retrieval performance of ResNet101 [6]. "L_f" represents that the query network learns feature representation knowledge from the gallery network. "$L_f + L_{pair}$" means that the query network learns feature representation knowledge and pairwise similarity knowledge. "$L_f + L_{dprd}$" denotes that we transfer feature representation knowledge,

Table 4. Performance (%) comparison across various network structures structures on three datasets.

METHOD	QUERY NET	QUERY INPUT	GALLERY NET	GALLERY INPUT	In-Shop		SOP		MSMT17	
					mAP (%)	R1 (%)	mAP (%)	R1 (%)	mAP (%)	R1 (%)
(A) Training without the gallery network										
ResNet101	ResNet101	256×256	ResNet101	256×256	81.96	95.42	72.06	86.92	59.81	81.91
SwinV2	SwinV2-T	256×256	SwinV2-T	256×256	80.28	94.55	74.22	88.00	56.14	78.76
ResNet18	ResNet18	64×64	ResNet18	64×64	60.16	80.23	41.84	65.76	13.85	30.17
MobileV2	MobileNetV2	64×64	MobileNetV2	64×64	62.53	82.64	44.94	68.39	12.13	26.70
(B) Training with ResNet101 as the gallery netowrk										
RKD [24]	ResNet18	64×64	ResNet101	256×256	0.15	0.10	0.03	0.01	0.07	0.06
PKT [25]					0.14	0.06	0.04	0.01	0.07	0.06
FitNet [31]					62.84	77.21	49.66	66.44	14.40	21.65
CCKD [27]					62.42	76.54	49.18	65.97	16.58	24.79
CSD [39]					26.00	29.83	39.28	54.85	6.19	9.04
RAML [33]					63.35	77.06	49.41	66.21	16.87	24.92
ROP [40]					35.16	39.38	28.79	37.71	7.65	11.60
D3still (Ours)					**64.31**	**79.60**	**51.33**	**69.18**	**18.64**	**28.79**
(C) Training with ResNet101 as the gallery netowrk										
FitNet [31]	MobileNetV2	64×64	ResNet101	256×256	66.52	81.09	50.91	67.57	15.38	22.45
CCKD [27]					63.74	77.63	49.88	66.63	14.71	20.67
CSD [39]					37.07	45.37	42.71	59.09	8.35	12.37
RAML [33]					62.85	75.98	50.72	67.38	14.69	20.33
ROP [40]					45.00	52.59	35.28	47.11	8.46	13.00
D3still (Ours)					**67.40**	**83.39**	**53.19**	**71.00**	**19.45**	**29.34**
(D) Training with SwinTransformerV2 as the gallery netowrk										
FitNet [31]	ResNet18	64×64	SwinV2-T	256×256	54.60	67.32	44.27	59.96	16.87	26.08
CCKD [27]					45.11	77.63	39.22	53.32	14.78	22.73
CSD [39]					17.36	18.80	32.81	46.45	4.80	7.11
RAML [33]					52.13	63.73	44.93	60.89	16.64	25.65
ROP [40]					22.38	22.99	23.47	30.01	7.95	12.24
D3still (Ours)					**56.08**	**70.03**	**46.77**	**64.22**	**18.29**	**28.78**

inconsistent pairwise similarity differential knowledge, and consistent pairwise similarity differential knowledge.

From Table 3, it is evident that the asymmetric image retrieval method significantly decreases the computational burden of the query network compared to symmetric image retrieval. Specifically, the asymmetric image retrieval method reduces the inference consumption of the query network from 12.99G FLOPs to 0.25G FLOPs, making it feasible to deploy the query network on edge devices.

Moreover, we observe that pairwise similarity knowledge slightly improves the query network's retrieval performance because pairwise similarity knowledge partially maintains retrieval ranking consistency between the query and gallery networks. For example, on the SOP dataset [22], "$L_f + L_{pair}$" outperforms "L_f" by 0.50% mAP and 0.39% R1.

Finally, the retrieval performance of the query network has been significantly improved when decoupled pairwise similarity differential knowledge is transferred from the gallery network to the query network. For example, on the SOP dataset [22], "$L_f + L_{dprd}$" outperforms "L_f" by a large margin, i.e., 1.87% mAP and 2.63% R1. These results demonstrate the effectiveness of our decoupled pairwise similarity differential knowledge in asymmetric image retrieval.

4.4. Comparison with State-of-the-art Methods

In this section, we conduct a comparative experiment between the D3still framework and state-of-the-art methods to assess the advantages of our proposed approach for asymmetric image retrieval. To ensure a fair comparison, we re-implement seven previous KD techniques for asymmetric image retrieval, as they exhibit varying training configurations. The detailed comparison analyses are presented below.

From Table 4, we can find that pure relationship distillation methods (i.e., RKD [24] and PKT [25]) that perform well in symmetric image retrieval cannot improve the performance of the query network in asymmetric image retrieval. This limitation arises from these methods only transferring relationship knowledge while neglecting feature knowledge, resulting in a misalignment of feature representation spaces between the two networks. For example, when using ResNet18 as query networks and ResNet101 as gallery networks, on the SOP dataset [22], RKD [24] only acquires a very poor performance, i.e., 0.03% mAP and 0.01% R1.

Moreover, theoretically, a distillation method that simultaneously transfers feature representation knowledge and relationship knowledge outperforms a pure feature distillation method. However, CSD [39] and ROP [40] are significantly

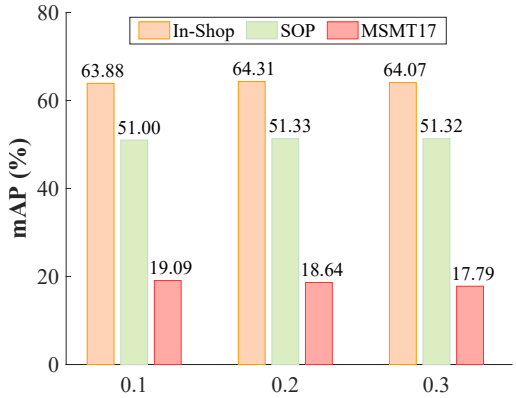

Figure 3. The influence of β value on mAP (%) performance.

Figure 4. The influence of γ value on mAP (%) performance.

worse than pure feature distillation methods (i.e., FitNet [31]). For example, when using ResNet18 as query networks and ResNet101 as gallery networks, on the SOP dataset [22], CSD [39] is 12.05% mAP and 14.33% R1 lower than FitNet [31]. This is because CSD [39] and ROP [40] fail to decouple feature representation knowledge and relationship knowledge, leading to feature representation knowledge taking a secondary role in the knowledge transfer process. Consequently, they struggle to align the feature representation space between the query network and the gallery network. These results show that it is necessary to decouple feature representation knowledge and relationship knowledge.

Finally, D3still significantly outperforms these distillation works (i.e., CSD [39] and RAML [33]) that transfer pairwise similarity knowledge. For example, when using ResNet18 as query networks and ResNet101 as gallery networks, on the MSMT17 dataset [37], D3still outperforms RAML [33] by 1.82% mAP and 4.20% R1. Besides, our method is superior to the best previous distillation method by 0.96% mAP on In-shop [19], 1.67% mAP on SOP [22], and 1.46% mAP on MSMT17 [37]. Even when widening the semantic gap between the query network and the gallery network, our method consistently achieves superior performance. For example, when using ResNet18 as query networks and SwinTransformerV2-Tiny (SwinV2-T) [20] as gallery networks, D3still exceeds RAML [33] by 1.84% mAP and 3.33% R1 on the SOP dataset [22]. These experiment results can demonstrate that our D3still framework achieves state-of-the-art performances across all various network structures on three benchmark datasets.

4.5. Hyper-parameter Analysis

The inconsistent pairwise differential knowledge weight (i.e., β in Eq. (11)). The hyper-parameter β crucially regulates inconsistent pairwise differential knowledge contributions in the distillation process. From Fig. 3, we observe a slight impact on retrieval performance due to β.

The consistent pairwise differential knowledge weight (i.e., γ in Eq. (11)). The hyper-parameter γ crucially regulates consistent pairwise differential knowledge contributions in the distillation process. From Fig. 4, we observe that γ does not significantly impact the retrieval performance.

5. Conclusion

In this paper, we introduce the Decoupled Differential Distillation (D3still) framework to alleviate the problem of conventional knowledge transfer with one-to-one pairwise similarity, which improves the performance of the query network suffering limited representation capacity. We decouple the pairwise similarity differential matrix in the gallery domain into three components, which focus on dealing with feature information, consistent ranking information, and inconsistent ranking information, respectively. With the more pertinent and relaxed learning objectives, our method reduces the representational capability requirements of the query network, leading to a significant improvement in accuracy.

Limitation. One notable limitation of our study is its specific focus on ranking properties within the context of image retrieval tasks, as well as the assumption of a significant representational capacity gap between the teacher network and the student network. Consequently, the applicability of our method to other tasks, such as object detection or image classification, may be limited.

Broader Impact. Our method highlights the effectiveness of relative knowledge in the context of asymmetric image retrieval, particularly when the query network has limited representation capacity. This insight may inspire further exploration of relative knowledge techniques tailored to specific tasks within the research community.

Acknowledgement. The work is supported by the China National Key R&D Program (No. 2023YFE0202700), Guangdong International Technology Cooperation Project (No.2022A0505050009), Key-Area Research and Development Program of Guangzhou City (No. 2023B01J0022), National Natural Science Foundation of China (No.62302170), Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097), Singapore MOE Tier 1 Funds (MSS23C002), and the NRF Singapore under the AI Singapore Programme (No. AISG3-GV-2023-011).

References

[1] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694, 2020. 1, 3

[2] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 8228–8238, 2021. 1, 2, 3

[3] Weiwei Cai, Huaidong Zhang, Xuemiao Xu, Shengfeng He, Kun Zhang, and Jing Qin. Contextual-assisted scratched photo restoration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 6

[4] Rahul Duggal, Hao Zhou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Compatibility-aware heterogeneous visual search. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 10723–10732, 2021. 4

[5] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 770–778, 2016. 6

[7] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 9161–9171, 2022. 2

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *ArXiv*, 1703.07737, 2017. 5

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[10] Hualong Huang, Wenhan Zhan, Geyong Min, Zhekai Duan, and Kai Peng. Mobility-aware computation offloading with load balancing in smart city networks using mec federation. *IEEE Transactions on Mobile Computing*, pages 1–17, 2024. 1

[11] Guanzhou Ke, Bo Wang, Xiao-Li Wang, and Shengfeng He. Rethinking multi-view representation learning via distilled disentangling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[12] Jangho Kim, Seonguk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *Conference on Neural Information Processing Systems*, pages 2765–2774, 2018. 2

[13] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems*, pages 1106–1114, 2012. 6

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

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

[17] Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, and Kai-Kuang Ma. Deep rank cross-modal hashing with semantic consistent for image-text retrieval. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4828–4832, 2022. 6

[18] Xiaoqing Liu, Huanqiang Zeng, Yifan Shi, Jianqing Zhu, Chih-Hsien Hsia, and Kai-Kuang Ma. Deep cross-modal hashing based on semantic consistent ranking. *IEEE Transactions on Multimedia*, 25:9530–9542, 2023. 6

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 1096–1104, 2016. 4, 5, 6, 8

[20] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 8

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 6

[22] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 4004–4012, 2016. 4, 5, 6, 7, 8

[23] Wenjie Pan, Linhan Huang, Jianbao Liang, Lan Hong, and Jianqing Zhu. Progressively hybrid transformer for multimodal vehicle re-identification. *Sensors*, 23(9):4206, 2023. 6

[24] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 3967–3976, 2019. 2, 7

[25] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and learning systems*, 32(5):2030–2039, 2020. 2, 7

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Conference on Neural Information Processing Systems*, 32, 2019. 6

[27] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *International Conference on Computer Vision*, pages 5007–5016, 2019. 7

[28] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333, 2021. 2

[29] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16782, 2022. 2

[30] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35, 2016. 6

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. 2, 7, 8

[32] Fei Shen, Yi Xie, Jianqing Zhu, Xiaobin Zhu, and Huanqiang Zeng. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing*, 32:1039–1051, 2023. 1

[33] Pavel Suma and Giorgos Tolias. Large-to-small image resolution asymmetry in deep metric learning. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1451–1460, 2023. 1, 2, 3, 6, 7, 8

[34] Wentao Tan, Changxing Ding, Pengfei Wang, Mingming Gong, and Kui Jia. Style interleaved learning for generalizable person re-identification. *IEEE Transactions on Multimedia*, 2023. 1

[35] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *International Conference on Computer Vision*, pages 1365–1374, 2019. 2

[36] Pengfei Wang, Changxing Ding, Wentao Tan, Mingming Gong, Kui Jia, and Dacheng Tao. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. *IEEE Transactions on Multimedia*, 2022. 6

[37] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 79–88, 2018. 6, 8

[38] Hanxiao Wu, Fei Shen, Jianqing Zhu, Huanqiang Zeng, Xiaobin Zhu, and Zhen Lei. A sample-proxy dual triplet loss function for object re-identification. *IET Image Processing*, 16(14):3781–3789, 2022. 6

[39] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 9489–9498, 2022. 2, 3, 4, 7, 8

[40] Hui Wu, Min Wang, Wengang Zhou, and Houqiang Li. A general rank preserving framework for asymmetric image retrieval. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 7, 8

[41] Hui Wu, Min Wang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. Asymmetric feature fusion for image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092, 2023. 1

[42] Yi Xie, Jianqing Zhu, Huanqiang Zeng, Canhui Cai, and Lixin Zheng. Learning matching behavior differences for compressing vehicle re-identification models. In *IEEE International Conference on Visual Communications and Image Processing*, pages 523–526, 2020. 6

[43] Yi Xie, Fei Shen, Jianqing Zhu, and Huanqiang Zeng. Viewpoint robust knowledge distillation for accelerating vehicle re-identification. *EURASIP Journal on Advances in Signal Processing*, 2021(1):1–13, 2021. 6

[44] Yi Xie, Hanxiao Wu, Fei Shen, Jianqing Zhu, and Huanqiang Zeng. Object re-identification using teacher-like and light students. In *British Machine Vision Conference*, 2021. 6

[45] Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, and Shengfeng He. Towards a smaller student: Capacity dynamic distillation for efficient image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2023. 2

[46] Yi Xie, Hanxiao Wu, Jianqing Zhu, and Huanqiang Zeng. Distillation embedded absorbable pruning for fast object re-identification. *Pattern Recognition*, 152:110437, 2024. 6

[47] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2023.

[48] Weiying Zheng, Cheng Xu, Xuemiao Xu, Wenxi Liu, and Shengfeng He. Ciri: Curricular inactivation for residue-aware one-shot video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13012–13022, 2023. 6

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. 6