Cross-Layer and Selective Distillation for Asymmetric Image Retrieval

Shijie Zhang[1][0009−0005−1210−0193], Weiqing Min[2,3][0000−0001−6668−9208], Fangyuan Yao[1][0009−0003−6680−5645], Guorui Sheng[1][0000−0001−6790−0239]✉, and Shuqiang Jiang[2,3][0000−0002−1596−4326]

[1] School of Computer Science and Artificial Intelligence, Ludong University, Yantai 264025, China
shengguorui@ldu.edu.cn
[2] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China
[3] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Existing asymmetric retrieval methods primarily rely on aligning global features to transfer semantic information. However, they often struggle to convey knowledge effectively across different network layers, limiting fine-grained alignment in feature representation spaces. To address this limitation, we propose a Cross-Layer and Selective Distillation (CLSD) framework. It first introduces a semantic-aware cross-layer feature distillation mechanism, where an attention-guided soft layer alignment strategy enables the student model to dynamically select and integrate the most relevant semantic knowledge from multiple intermediate teacher layers, based on its own layer's semantic requirements. This alleviates the knowledge transfer challenges arising from architectural asymmetry. Furthermore, considering the importance of ranking consistency in fine-grained food image retrieval, we propose a decoupled differential relation distillation approach based on unambiguous samples. This method emphasizes the teacher model's discriminative power and ranking behavior on unambiguous samples, while filtering out noisy signals from ambiguous ones. As a result, the student learns more reliable relative relationships between samples, ensuring consistency in ranking order between query and gallery features. Extensive experiments on four benchmark datasets demonstrate that our method consistently surpasses existing state-of-the-art techniques, highlighting its effectiveness in asymmetric fine-grained retrieval tasks.

Keywords: Asymmetric image retrieval · Knowledge distillation · Cross-Layer distillation.

1 Introduction

Current mainstream deep learning image retrieval [6, 7] relies on large networks to extract discriminative features, typically deployed on cloud servers. As shown

in Fig. 1(a), this "cloud extraction, centralized matching" paradigm faces challenges in edge device applications (such as food image retrieval on smartphones), including network latency and server computation burden. Reducing online computation and communication costs is the key to efficient and scalable image retrieval.

Asymmetric image retrieval [2, 21], an emerging solution, effectively balances performance and efficiency by deploying a lightweight network on the client side to extract query features and utilizing features extracted by a large model on the server side for matching. As shown in Fig. 1(b), this architecture enables efficient retrieval with minimal client-side computation. However, this constitutes an implicit knowledge distillation mechanism. The core challenge lies in how the lightweight student model can effectively inherit the semantic capability of the large model under structural heterogeneity. Introducing explicit knowledge distillation strategies is necessary.

In the framework of Knowledge Distillation, the distillation effect largely depends on the chosen "knowledge type"[13]. Feature distillation has made significant progress in recent years and has become one of the state-of-the-art methods [27]. However, feature-based distillation has limited application in fine-grained image retrieval (such as food image retrieval). This is mainly due to: 1) The fine-grained differences in food images require high semantic understanding in intermediate layers, and whether intermediate layer features can accurately express complex semantics directly affects student performance. 2) The significant structural difference between lightweight student models on the edge side and large models on the cloud makes traditional feature distillation's static layer alignment prone to semantic mismatch.

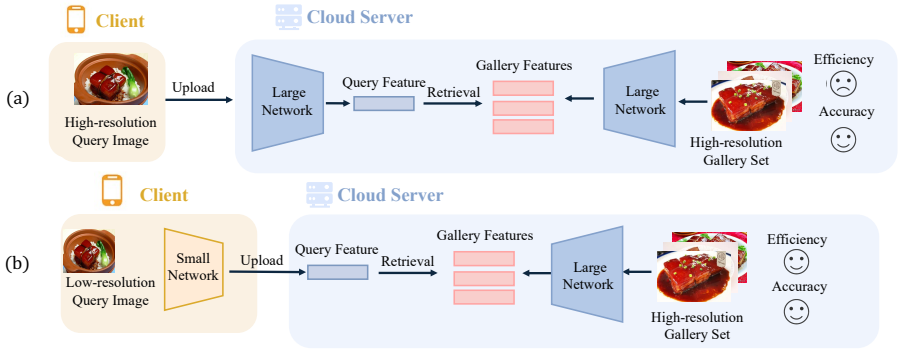

Fig. 1: Comparison of image retrieval architectures: (a) Symmetric image retrieval network architecture. (b) Asymmetric image retrieval network architecture.

Analyzing the limitations of feature alignment in networks of different capacities when handling fine-grained food images, we believe that simply pursuing

feature representation similarity is insufficient to effectively transfer the teacher model's discriminative capability, especially in image retrieval which focuses on relative order. A more practical goal should be to maintain the consistency of the relative order of returned images. Given this, we propose a new distillation objective that focuses on learning the pairwise similarity differential relationships between samples. However, we found that not all samples are beneficial for retrieval order consistency. Extremely similar-looking samples may lead to erroneous differential relationships and mislead the student model.

To address these challenges, we propose the following contributions:

- We propose a semantic-aware cross-layer feature distillation method that dynamically selects the most relevant semantics from multiple intermediate teacher layers based on the student model's layer-specific semantic requirements, effectively guiding the student to learn deeper features.
- We further propose a decoupled differential distillation method based on unambiguous samples to ensure retrieval order consistency between the lightweight query network and the large gallery network.
- Experimental results on multiple image retrieval benchmark datasets demonstrate that our proposed method significantly improves retrieval accuracy while maintaining the lightweight nature of the student model, outperforming existing mainstream distillation strategies.

2 Related Work

2.1 Knowledge Distillation

Knowledge Distillation (KD) is a prevalent technique for transferring knowledge from a teacher network to a student network. KD methods are generally categorized into three types: Logit Distillation, Relationship Distillation, and Feature Distillation [13].

Logit Distillation This method aligns the output logits of the student with those of the teacher, allowing the student to learn the teacher's prediction tendencies and inter-class relationships. Methods such as NTCE-KD [9] enhance the role of non-target classes in the logits for more effective knowledge transfer.

Relationship Distillation This method focuses on aligning the relationships derived from the intermediate features of student and teacher networks. Methods such as SPKD [22] transfer pairwise similarity knowledge by minimizing the distance between their cosine similarity matrices.

Feature Distillation This method leverages multi-level intermediate features from the teacher to guide student learning. Methods such as FitNet [20] minimize the distance between the intermediate features of student and teacher networks to improve representational alignment.

2.2 Asymmetric Image Retrieval

In asymmetric retrieval, compatible gallery/query embeddings are vital due to distinct architectures. Feature distillation effectively aligns these features.

AML [2] first applied KD, using query (student) features as anchors for metric learning with gallery (teacher) features (positive/negative). Later methods added pairwise similarity knowledge, e.g., CSD [25] minimizing context similarity matrix differences. Recognizing strict one-to-one constraints for lower-capacity students, recent work favors ranking-based approaches with relaxed constraints. Examples include ROP [24] using sigmoid for binary ranking and D3still [26] optimizing relational similarity differences for consistent retrieval order.

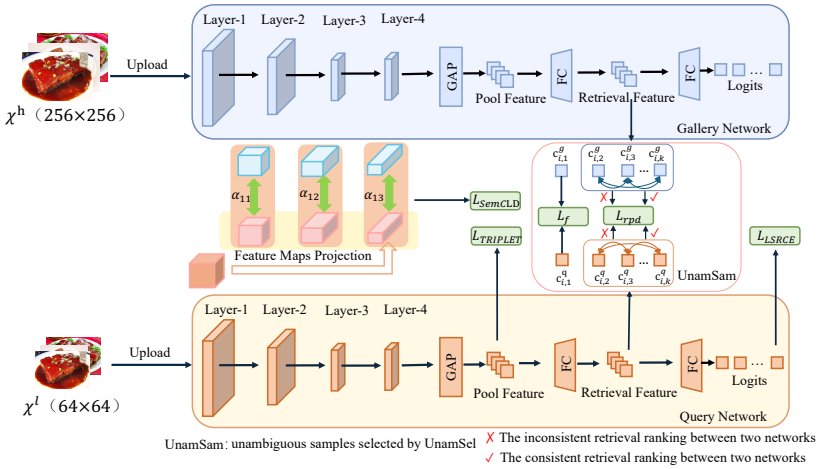

Fig. 2: Overview of the Proposed CLSD Method. Our L_{rpd} comprises the differential distillation loss L_{irpd} for inconsistent pairs and L_{crpd} for consistent pairs.

3 Method

This section details our proposed knowledge distillation method for asymmetric image retrieval. Its core strategies include: 1) semantic alignment formula-guided cross-layer feature distillation; and 2) decoupled differential relation distillation based on unambiguous samples.

3.1 Formulation and Background

Fig. 2 illustrates our method for knowledge transfer using deep features extracted from the query network and the gallery network. First, given a batch of n samples as input $\chi = \{x_1, x_2, ..., x_n\}$, we scale them separately to a low-resolution sample set $\chi^l = \{x_1^l, x_2^l, ..., x_n^l\}$ and a high-resolution sample set $\chi^h = \{x_1^h, x_2^h, ..., x_n^h\}$. Then, we use a lightweight query network $\theta_q(\cdot)$ and a large gallery network $\theta_g(\cdot)$ to convert the low-resolution and high-resolution images into normalized

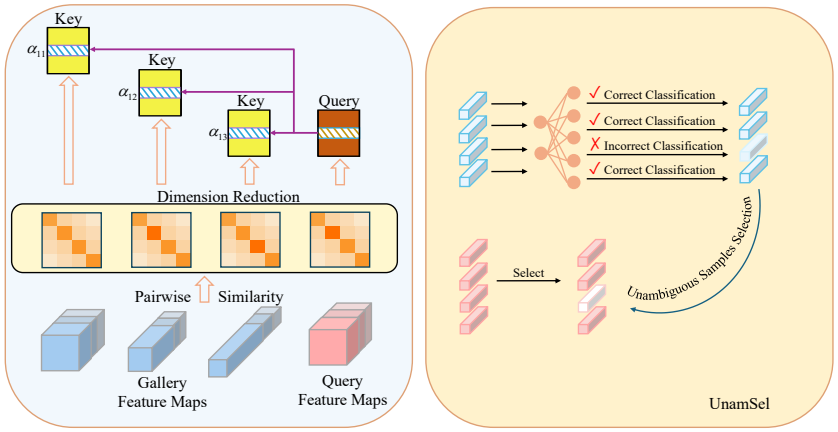

Fig. 3: (a) Attention Weight Calculation for Cross-Layer Distillation. (b) Unambiguous Samples Selection (UnamSel).

two-dimensional vectors as follows:

$$\mathbf{v}_i^q = \theta_q(\chi_i^l), \mathbf{v}_i^g = \theta_g(\chi_i^h), i = 1, 2, ..., n, \tag{1}$$

where $\mathbf{v}_i$ represents a normalized D-dimensional vector extracted from the i-th image. Additionally, to achieve effective knowledge transfer between the query network and the gallery network, we also extract intermediate layer features of the model. The multi-layer features of the student network are represented as $\mathbf{F}^s = [f_{s_1}^s, f_{s_2}^s, ..., f_{s_L}^s]$, and the teacher network extracts corresponding multi-layer features $\mathbf{F}^t = [f_{t_1}^t, f_{t_2}^t, ..., f_{t_L}^t]$.

3.2 Cross-Layer Feature Distillation

As shown in Fig. 3(a), we propose an attention allocation mechanism for student layers to establish soft associations with semantically relevant teacher layers. For each student layer $f_{s_l}^{\mathrm{s}}$, an attention mechanism calculates its weight $\alpha(s_l, t_l)$ with each teacher layer, ensuring $\sum_{t_l=1}^{t_L} \alpha(s_l, t_l) = 1, \forall s_l \in [1, ..., s_L]$. To align spatial dimensions for distance calculation, student feature maps are projected into t_L independent forms:

$$f_{s_l}^{s'} = Proj(f_{s_l}^s \in \mathbb{R}^{b \times c_{s_l} \times h_{s_l} \times w_{s_l}}, t_l), t_l \in [1, ..., t_L], \tag{2}$$

where $f_{s_l}^{s'} \in \mathbb{R}^{b \times c_{t_l} \times h_{t_l} \times w_{t_l}}$. Each $Proj(\cdot, \cdot)$ uses stacked 1×1, 3×3, and 1×1 convolutions for feature transformation.

To optimize distillation, pairwise similarity matrices effectively measure intrinsic semantic similarity for layer association:

$$A_{s_l}^{\mathrm{s}} = R(f_{s_l}^{\mathrm{s}}) \cdot R(f_{s_l}^{\mathrm{s}})^{\mathrm{T}}, \quad A_{t_l}^{\mathrm{t}} = R(f_{t_l}^{\mathrm{t}}) \cdot R(f_{t_l}^{\mathrm{t}})^{\mathrm{T}}, \tag{3}$$

where $R(\cdot) : \mathbb{R}^{b\times c\times h\times w} \rightarrow \mathbb{R}^{b\times chw}$ is a reshaping operation. Based on self-attention, these matrices are projected via MLPs into query and key vectors to alleviate noise/sparsity:

$$Q_{s_l} = MLP_Q(A^{\mathrm{s}}_{s_l}) \quad K_{t_l} = MLP_K(A^{\mathrm{t}}_{t_l}). \tag{4}$$

The attention weight $\alpha_{(s_l,t_l)}$ is then calculated:

$$\alpha_{(s_l,t_l)} = \frac{e^{Q^T_{s_l} K_{t_l}}}{\sum_{t_j=1}^{t_L} e^{Q^T_{s_l} K_{t_j}}}. \tag{5}$$

This attention mechanism alleviates hierarchical mismatch and integrates positive guidance from multiple teacher layers. The complete training process is summarized in Algorithm 1.

Algorithm 1: Calculate Cross-Layer Distillation Loss

Input: A mini-batch $\mathcal{B}$ of size b; a pre-trained teacher model with parameters θ_g; a student model with randomly initialized parameters θ_q

Output: Calculated cross-layer distillation loss L_{SemCLD}

Forward $\mathcal{B}$ through θ_g and θ_q to obtain intermediate features $f^t_{t_l}$ and $f^s_{s_l}$ across layers;

Construct pairwise similarity matrices $A^{\mathrm{t}}_{t_l}$ and $A^{\mathrm{s}}_{s_l}$ (Eq. 3);

Perform attention allocation (Eq. 4-5);

Align feature maps via projection (Eq. 2);

Calculate cross-layer distillation loss L_{SemCLD} using $f^t_{t_l}$, $f^{s'}_{s_l}$, and $\alpha_{(s_l,t_l)}$ (Eq. 6);

return L_{SemCLD};

The cross-layer distillation loss L_{SemCLD} is calculated using Mean Squared Error (MSE) after semantic layer association and dimension projection:

$$L_{SemCLD} = \sum_{i=1}^{n} \sum_{s_l=1}^{s_L} \sum_{t_l=1}^{t_L} \alpha_{i,(s_l,t_l)} MSE(f^t_{i,t_l}, f^{s'}_{i,s_l}). \tag{6}$$

3.3 Decoupled Differential Relation Distillation

We measure query-gallery similarity using cosine similarity. Given asymmetric network capabilities, cosine similarity matrices G^q and G^g are computed in the gallery network's representation space:

$$G^q = \mathbf{v}^q(\mathbf{v}^g)^T \in \mathbb{R}^{n\times n}, \quad G^g = \mathbf{v}^g(\mathbf{v}^g)^T \in \mathbb{R}^{n\times n}. \tag{7}$$

For Top-k retrieval, we obtain Top-k indices $R \in \mathbb{R}^{n\times k}$ from G^g:

$$R = \mathrm{argsort}(G^g, \mathrm{dim} = 2), \tag{8}$$

where argsort$(\cdot)$ returns the top-k indices based on the descending order of cosine similarity along the second dimension. Then, Top-k retrieval similarity matrices $C^q \in \mathbb{R}^{n\times k}$ and $C^g \in \mathbb{R}^{n\times k}$ are constructed:

$$C^q = \text{sort}(G^q, \text{index} = R), \quad C^g = \text{sort}(G^g, \text{index} = R), \tag{9}$$

where sort$(\cdot)$ represents a function that sorts the cosine similarity matrix based on the top-k indices. As depicted in Fig. 3(b), our method focuses on leveraging unambiguous knowledge. We construct a binary mask m to evaluate sample features using the trained gallery network's unbiased fully connected layer, thereby discarding ambiguous samples:

$$m_i = \begin{cases} 1 & \hat{y}_i = y_i \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $\hat{y}_i = \text{argmax}\left(W^g\left(\theta_g(\chi_i^h)\right)\right)$, y_i is the i-th sample's label, and $W^g \in \mathbb{R}^{M\times D}$ are gallery network FC layer weights.

For feature representation knowledge, the feature distillation loss L_f aligns query and gallery network representation spaces based on the Top-k retrieval similarity matrix:

$$L_f = \frac{1}{\sum_{i=1}^n m_i}\left(\sum_{i=1}^n m_i (C_{i,1}^q - C_{i,1}^g)^2\right)^{\frac{1}{2}}. \tag{11}$$

To model similarity difference knowledge between sample pairs, two difference similarity matrices $M^q \in \mathbb{R}^{n\times(k-1)\times(k-1)}$ and $M^g \in \mathbb{R}^{n\times(k-1)\times(k-1)}$ are constructed from the Top-k retrieval similarity matrix:

$$M_{i,j,l}^q = C_{i,j+1}^q - C_{i,l+1}^q, \quad 1 \le j,l \le k-1. \tag{12}$$

$$M_{i,j,l}^g = C_{i,j+1}^g - C_{i,l+1}^g, \quad 1 \le j,l \le k-1. \tag{13}$$

The difference distillation loss is then decomposed into L_{irpd} for inconsistent and L_{crpd} for consistent sample pairs:

$$L_{irpd} = \frac{1}{\sum_{i=1}^n m_i}\sum_{i=1}^n m_i\left(\sum_{j=1}^{k-1}\sum_{l=1}^{k-1}\mathcal{H}\left(-\frac{M_{i,j,l}^q}{M_{i,j,l}^g}\right)\left(\frac{M_{i,j,l}^q - M_{i,j,l}^g}{\mu + |M_{i,j,l}^g|}\right)^2\right)^{\frac{1}{2}}. \tag{14}$$

$$L_{crpd} = \frac{1}{\sum_{i=1}^n m_i}\sum_{i=1}^n m_i\left(\sum_{j=1}^{k-1}\sum_{l=1}^{k-1}\mathcal{H}\left(\frac{M_{i,j,l}^q}{M_{i,j,l}^g}\right)\left(\frac{M_{i,j,l}^q - M_{i,j,l}^g}{\mu + |M_{i,j,l}^g|}\right)^2\right)^{\frac{1}{2}}. \tag{15}$$

where, μ is a constant set to 0.1 to avoid a small denominator. $\mathcal{H}(\cdot)$ represents a Heaviside step function.

During the training phase, the total loss function of the query network is $L_{student}$ as follows:

$$L_{student} = \alpha L_f + \beta L_{irpd} + \gamma L_{crpd} + \delta L_{SemCLD} + \epsilon L_{TRIPLET} + \zeta L_{LSRCE}. \tag{16}$$

To further optimize the feature learning effect of the query network in asymmetric image retrieval, we apply the triplet loss $L_{TRIPLET}$ to the features after pooling, and calculate the cross-entropy loss L_{LSRCE} on the Logits output. Where, $\alpha, \beta, \gamma, \delta, \epsilon$ and ζ are all hyperparameters, used to adjust their contributions in the total loss function.

4 Experiment

In this section, we conduct experiments on four widely used datasets to verify the superiority of our method. These datasets include ETH Food-101 [1], Vireo Food-172 [3], In-Shop Clothes Retrieval [12], and Stanford Online Products [14].

4.1 Datasets and Performance Metrics

ETH Food-101 (Food-101) [1] is a dataset of Western food images, containing 101,000 images from 101 categories. The training set includes 750 images per category, totaling 75,750 images. The test set includes 250 images per category, totaling 25,250 images.

Vireo Food-172 (Food-172) [3] is a large-scale food image dataset, containing 110,241 images from 172 categories. The training set includes 172 categories, 66,071 images, and the test set includes 172 categories, 44,170 images.

In-Shop Clothes Retrieval (In-Shop) [12] is a clothes retrieval dataset, containing 72,712 images, covering 7,986 categories. The training set includes 3,997 categories, totaling 25,882 images. The query set consists of 14,218 images, belonging to 3,985 categories. The gallery set includes 3,985 categories, totaling 12,612 images.

Stanford Online Products (SOP) [14] is a widely used product recognition dataset, containing a large number of 120,053 product images, covering 22,634 categories. The training set includes 59,551 training images, belonging to 11,318 categories, while the test set includes 60,502 images, belonging to 11,316 categories.

In the inference stage, we rank gallery images by cosine similarity between query and gallery features, with higher scores indicating higher rank. We evaluate accuracy using mean Average Precision (mAP) [15] and Rank-1 (R1) [10].

4.2 Implementation Details

We set hyperparameters ($\alpha = 200, \beta = 5, \gamma = 1, \delta = 8, \epsilon = 1, \zeta = 1$) and use PyTorch 2.0.1 [18] with CUDA 11.8 on an NVIDIA A800 GPU. For network training: The gallery network processes inputs at a resolution of 256×256, while the query network processes inputs at 64×64. Data augmentation includes z-score normalization, random cropping [23], erasing (p=0.5), and horizontal flipping (p=0.5). We use mini-batch SGD [8] (batch size 512, weight decay 5×10^{-4}, momentum 0.9). The learning rate (initial 1×10^{-3}, warmed up to 1×10^{-2} over 10 epochs, dropping at 40th epoch) follows a cosine annealing and linear warm-up strategy for 120 total epochs.

Table 1: Comparison of different methods on Food-101 and Food-172 datasets.

METHOD	Query net	Query input	Gallery Net	Gallery Input	Food-101		Food-172	
					mAP(%)	R1(%)	mAP(%)	R1(%)
(A)Training without the gallery network								
ResNet101 [5]	ResNet101	256 × 256	ResNet101	256 × 256	82.45	86.14	80.1	84.88
SwinV2 [11]	SwinV2-T	256 × 256	SwinV2-T	256 × 256	76.86	87.13	71.64	86.63
(B) Training with ResNet101 as the gallery network								
VanillaKD [4]					1.78	0.99	1.02	1.74
RKD [16]					1.41	0.99	0.83	0.58
PKT [17]					1.47	0.99	0.94	0.00
FitNet [20]					41.16	49.5	46.18	55.81
CSD [25]	ResNet18	64 × 64	ResNet101	256 × 256	53.24	53.47	63.98	66.86
RAML [21]					50.13	51.49	65.10	68.02
ROP [24]					50.87	51.49	58.12	63.37
CCKD [19]					51.38	50.5	65.73	70.35
D3still [26]					55.16	56.44	64.65	70.32
Ours					**56.23**	**60.40**	**67.35**	**70.93**
(C) Training with Swin-Transformer-V2 as the gallery network								
VanillaKD [4]					1.68	0.99	0.88	0.58
RKD [16]					2.32	0.99	0.89	1.16
PKT [17]					1.35	0.99	0.91	0.00
FitNet [20]					44.59	53.47	51.73	61.05
CSD [25]	ResNet18	64 × 64	SwinV2-T	256 × 256	53.15	58.42	62.02	70.35
RAML [21]					53.85	56.44	62.5	69.19
ROP [24]					49.86	58.42	60.14	70.93
CCKD [19]					54.59	61.39	61.28	70.35
D3still [26]					52.36	62.38	61.65	70.35
Ours					**57.01**	**63.37**	**62.82**	**74.42**
(D) Training with ResNet101 as the gallery network								
VanillaKD [4]					1.76	0.99	0.82	0.58
RKD [16]					2.21	0.99	0.81	0.58
PKT [17]					1.67	0.99	1.12	1.16
FitNet [20]					27.64	43.56	28.34	27.91
CSD [25]	MobileNetV3	64 × 64	ResNet101	256 × 256	50.02	51.49	**57.06**	56.4
RAML [21]					46.44	47.52	55.96	58.72
ROP [24]					43.30	46.53	39.5	45.35
CCKD [19]					46.89	49.5	47.11	49.42
D3still [26]					46.08	48.51	46.84	51.74
Ours					**51.92**	**55.45**	55.36	**58.72**

Table 2: Comparison of different methods on In-Shop and SOP datasets.

METHOD	Query net	Query input	Gallery Net	Gallery Input	In-Shop		SOP	
					mAP(%)	R1(%)	mAP(%)	R1(%)
(A) Training without the gallery network								
ResNet101 [5]	ResNet101	256 × 256	ResNet101	256 × 256	81.96	95.42	72.06	86.92
SwinV2 [11]	SwinV2-T	256 × 256	SwinV2-T	256 × 256	80.28	94.55	74.22	88.00
(B) Training with ResNet101 as the gallery network								
VanillaKD [4]					0.15	0.02	0.04	0.00
RKD [16]					0.15	0.06	0.03	0.00
PKT [17]					0.13	0.02	0.04	0.02
FitNet [20]					65.99	80.50	48.87	65.35
CSD [25]	ResNet18	64 × 64	ResNet101	256 × 256	66.64	81.00	49.43	65.96
RAML [21]					67.18	81.85	49.46	66.24
ROP [24]					65.58	80.24	48.03	64.66
CCKD [19]					66.60	81.21	49.11	66.05
D3still [26]					68.56	83.96	51.12	68.42
Ours					**69.43**	**84.79**	**52.16**	**69.58**
(C) Training with Swin-Transformer-V2 as the gallery network								
VanillaKD [4]					0.13	0.03	0.03	0.01
RKD [16]					0.14	0.04	0.04	0.02
PKT [17]					0.16	0.04	0.03	0.00
FitNet [20]					56.35	65.77	37.06	46.57
CSD [25]	ResNet18	64 × 64	SwinV2-T	256 × 256	57.58	67.87	40.50	51.63
RAML [21]					57.34	67.42	40.64	51.97
ROP [24]					53.87	63.52	37.90	49.45
CCKD [19]					56.55	65.51	40.27	51.94
D3still [26]					60.19	72.07	43.32	56.98
Ours					**62.13**	**74.93**	**44.00**	**57.47**
(D) Training with ResNet101 as the gallery network								
VanillaKD [4]					0.16	0.06	0.03	0.00
RKD [16]					0.15	0.03	0.03	0.00
PKT [17]					0.15	0.08	0.04	0.01
FitNet [20]					60.41	74.61	44.80	60.30
CSD [25]	MobileNetV3	64 × 64	ResNet101	256 × 256	62.27	76.48	44.98	60.72
RAML [21]					62.29	76.13	45.52	61.48
ROP [24]					61.43	75.91	43.67	59.35
CCKD [19]					61.53	76.25	44.37	60.09
D3still [26]					63.58	79.43	45.80	62.72
Ours					**64.29**	**80.27**	**46.68**	**63.67**

4.3 Comparison with Existing State-of-the-Art Methods

In this section, we conduct comparative experiments between the CLSD framework and state-of-the-art methods to evaluate the superiority of our proposed method in asymmetric image retrieval. To ensure a fair comparison, we re-implement eight prior KD techniques for asymmetric image retrieval, as they have different training configurations. Detailed comparative analysis is as follows.

As shown in Tables 1 and 2, pure relation distillation methods (e.g., RKD [16], PKT [17], and VanillaKD [4]) fail to effectively improve query network performance in asymmetric image retrieval. This is because they transfer only relational knowledge, neglecting feature knowledge, leading to misaligned feature spaces between networks. For instance, on ETH Food-101 [1], VanillaKD [4] only achieves 0.15% mAP and 0.02% R1 with ResNet18 as the query and ResNet101 as the gallery network.

Our method significantly outperforms distillation techniques that primarily optimize relational similarity differences, namely D3still [26], CSD [25], and RAML [21]. For instance, when employing ResNet18 [5] as the query network and ResNet101 [5] as the gallery network, our approach surpasses D3still on the Food-101 dataset [1] by 1.07% mAP and 3.96% R1. On the Food-172 dataset [3], our method achieves a higher mAP of 1.62% and R1 of 0.58% compared to the previous best distillation method. On the In-shop dataset [12], it yields a mAP gain of 0.87% and R1 gain of 0.83%. Additionally, it achieves a higher mAP of 1.04% and R1 of 1.16% on the SOP dataset [14]. Our method maintains superior performance even when increasing the semantic gap between the query and gallery networks. For example, when using ResNet18 [5] as the query network and SwinTransformerV2-Tiny [11] as the gallery network, our method on the Food-101 dataset [1] exceeds CCKD by 2.42% mAP and 1.98% R1. Similarly, on the In-shop dataset [12], it surpasses D3still by 1.94% mAP and 2.86% R1. It is noteworthy that the performance of our method on the Food-172 dataset [3] is not optimal when using MobileNetV3 as the query network and ResNet101 [5] as the gallery network. This suggests that effectively balancing and transferring knowledge of different granularities remains a challenge in certain datasets and network configurations. Nevertheless, these experimental results demonstrate that our method achieves state-of-the-art performance in most cases across various network architectures.

Table 3: Ablation study on In-shop and ETH Food-101.

METHOD	FLOPs (G)	In-Shop		ETH Food-101	
		mAP (%)	R1 (%)	mAP (%)	R1 (%)
Gallery	12.99	81.96	95.42	82.45	86.14
L_f	0.25	67.24	81.91	51.18	52.48
$L_f + L_{SemCLD}$	0.25	68.77	83.61	52.78	56.93
$L_f + L_{SemCLD} + L_{irpd} + L_{crpd}$	0.25	69.43	84.79	57.01	63.37

4.4 Ablation Experiments

As shown in Table 3, we conducted ablation experiments on the ETH Food-101 [1] and In-shop datasets [12] to evaluate the effectiveness of each proposed component. "Gallery" denotes the baseline of direct ResNet101 retrieval performance without knowledge transfer. The base model on the query side is trained with a combination of triplet loss $L_{TRIPLET}$ and cross-entropy loss L_{LSRCE}. We progressively added three types of losses to the query network: L_f as an auxiliary loss for learning feature representation knowledge from the gallery network; L_{SemCLD} to capture cross-layer semantic knowledge ; and finally, $L_{\text{irpd}} + L_{\text{crpd}}$ to transfer inconsistent and consistent pairwise similarity difference knowledge for enhancing fine-grained relational features.

As can be clearly seen from Table 3, asymmetric image retrieval significantly reduces the computational burden of the query network compared to symmetric methods. Specifically, the inference computation for the query network is greatly reduced from 12.99 GFLOPs to 0.25 GFLOPs. In addition, we observe that introducing cross-layer knowledge can effectively improve retrieval performance. For example, on the In-Shop dataset [12], after adding L_{SemCLD}, the mAP increased by 1.53%, and R1 increased by 1.7%. On the ETH Food-101 dataset [1], the mAP increased by 1.6%, and R1 increased by 4.45%. By further adding L_{irpd} and L_{crpd}, the mAP on the In-Shop dataset [12] increased by 0.66%, and R1 increased by 1.18%; on the ETH Food-101 dataset [1], the mAP increased by 4.23%, and R1 increased by 6.44%.

4.5 Visualization

To analyze method performance, we conducted Grad-CAM, t-SNE, and retrieval visualization on the ETH Food-101 dataset [1]. As shown in Fig. 4, Grad-CAM indicates our method more accurately focuses on target objects, effectively suppressing background interference, and significantly improves retrieval performance. As shown in Fig. 5, t-SNE visualization demonstrates our method forms tighter clusters for the same class and maintains good inter-class separation in the feature space, validating its effectiveness in semantic modeling and class discrimination. As shown in Fig. 6, retrieval results further show our method exhibits good generalization ability with accurate retrieval across multiple subcategories. Although minor confusion exists for some fine-grained categories, this set of visualizations strongly supports the effectiveness of our method in image understanding and retrieval.

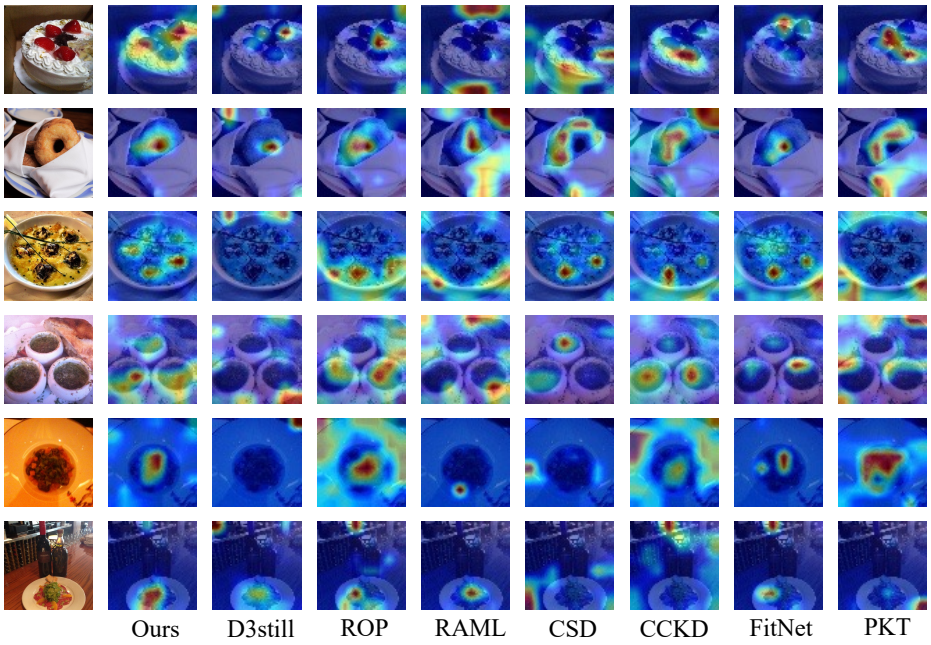

Fig. 4: Grad-CAM Visualization of Different Methods on ETH Food-101

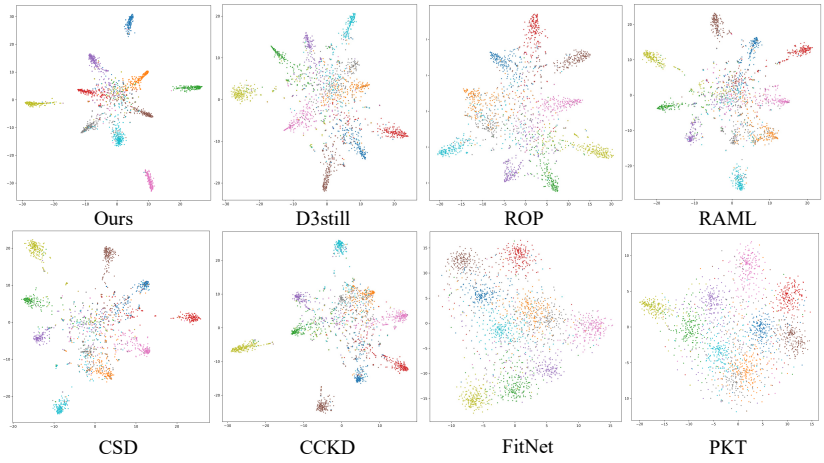

Fig. 5: t-SNE Visualization of Different Methods on ETH Food-101

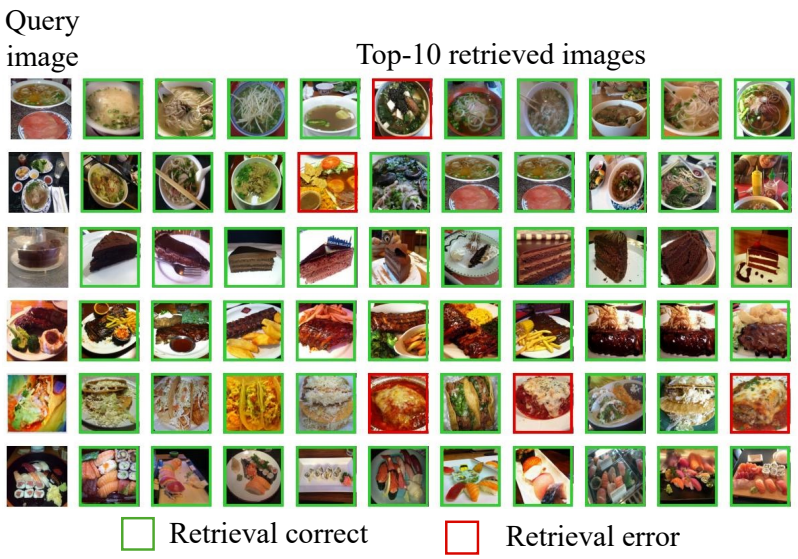

Fig. 6: Top-10 Retrieval Results on ETH Food-101

5 Conclusion

In conclusion, this paper presents CLSD, a novel framework for asymmetric image retrieval. By introducing semantic-aware cross-layer feature distillation, CLSD effectively bridges the structural gap between teacher and student networks. Additionally, the decoupled differential relation distillation based on unambiguous samples enhances the student model's retrieval consistency, especially in fine-grained scenarios. Experimental results demonstrate that CLSD outperforms existing methods in both retrieval accuracy and lightweightness, making it suitable for deployment on edge devices.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13. pp. 446–461. Springer (2014)
2. Budnik, M., Avrithis, Y.: Asymmetric metric learning for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2021)
3. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 32–41 (2016)

4. Hao, Z., Guo, J., Han, K., Hu, H., Xu, C., Wang, Y.: Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale. arXiv preprint arXiv:2305.15781 (2023)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Jiang, X., Tang, H., Li, Z.: Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. IEEE Transactions on Knowledge and Data Engineering (2024)
7. Jiang, X., Tang, H., Yan, R., Tang, J., Li, Z.: Dvf: Advancing robust and accurate fine-grained image retrieval with retrieval guidelines. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2379–2388 (2024)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
9. Li, C., Teng, X., Ding, Y., Lan, L.: Ntce-kd: Non-target-class-enhanced knowledge distillation. Sensors **24**(11), 3617 (2024)
10. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: 2016 IEEE international conference on multimedia and expo (ICME). pp. 1–6. IEEE (2016)
11. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019 (2022)
12. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1096–1104 (2016)
13. Mansourian, A.M., Ahmadi, R., Ghafouri, M., Babaei, A.M., Golezani, E.B., Ghamchi, Z.Y., Ramezanian, V., Taherian, A., Dinashi, K., Miri, A., et al.: A comprehensive survey on knowledge distillation. arXiv preprint arXiv:2503.12067 (2025)
14. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016)
15. Pan, W., Huang, L., Liang, J., Hong, L., Zhu, J.: Progressively hybrid transformer for multi-modal vehicle re-identification. Sensors **23**(9), 4206 (2023)
16. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
17. Passalis, N., Tzelepi, M., Tefas, A.: Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Transactions on Neural Networks and learning systems **32**(5), 2030–2039 (2020)
18. Paszke, A.: Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
19. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5007–5016 (2019)
20. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)

21. Suma, P., Tolias, G.: Large-to-small image resolution asymmetry in deep metric learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1451–1460 (2023)
22. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
23. Wang, P., Ding, C., Tan, W., Gong, M., Jia, K., Tao, D.: Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. IEEE Transactions on Multimedia **25**, 2624–2635 (2022)
24. Wu, H., Wang, M., Zhou, W., Li, H.: A general rank preserving framework for asymmetric image retrieval. In: The Eleventh International Conference on Learning Representations (2023)
25. Wu, H., Wang, M., Zhou, W., Li, H., Tian, Q.: Contextual similarity distillation for asymmetric image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9489–9498 (2022)
26. Xie, Y., Lin, Y., Cai, W., Xu, X., Zhang, H., Du, Y., He, S.: D3still: Decoupled differential distillation for asymmetric image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17181–17190 (2024)
27. Zhang, H., Xie, Y., Zhang, H., Xu, C., Luo, X., Chen, D., Xu, X., Zhang, H., Heng, P.A., He, S.: Unambiguous granularity distillation for asymmetric image retrieval. Neural Networks **187**, 107303 (2025)