Learning from the Unambiguous: Selective Relation Distillation for Asymmetric Fine-Grained Retrieval

Shijie Zhang, Weiqing Min, Guorui Sheng, Tao Yao, and Shuqiang Jiang

***Abstract*—Most deep learning-based image retrieval methods rely on cloud-deployed models, leading to substantial transmission latency and high computational demands for users on edge devices. This limitation is critical in fine-grained retrieval tasks, such as food recognition, which require both high semantic discrimination and low-latency response. Asymmetric image retrieval addresses this issue by deploying lightweight models on the client side. However, transferring knowledge from a powerful teacher network to a structurally distinct lightweight student remains challenging due to semantic misalignment and negative transfer. To overcome this, we propose Multi-layer and Selective Distillation (MLSD), an asymmetric retrieval framework with two key components: (1) a multi-layer distillation mechanism that fuses intermediate teacher–student features to bridge semantic gaps and enable accurate cross-layer transfer; and (2) a decoupled differential relation distillation method that filters ambiguous samples via a binary mask and distills pairwise similarity relations only from unambiguous data, ensuring ranking consistency and reducing noisy supervision. Extensive experiments on benchmark datasets show that MLSD consistently outperforms state-of-the-art asymmetric retrieval methods.**

***Index Terms*—Asymmetric image retrieval, Knowledge distillation, Fine-grained image retrieval, Edge device deployment.**

I. Introduction

MOST deep learning-based image retrieval methods [1]–[4] project query and database images into a shared discriminative feature space using large, parameter-heavy networks. In practical deployment, these models are usually hosted on cloud servers. When a query is issued from an edge device (e.g., a smartphone) [5], the image must be transmitted to the server, where feature extraction and centralized matching are performed. This paradigm suffers from two main drawbacks: (1) transmission is highly dependent on network conditions, leading to latency and instability, and (2) each uploaded query requires a computationally expensive forward pass on the server, creating heavy overhead under large-scale concurrent requests. Reducing the online computational burden is therefore a critical issue for efficient and scalable image retrieval.

These challenges are particularly critical in food image retrieval, where images span a large number of fine-grained categories with substantial intra-class variations [6]–[9]. Such tasks demand high discriminative capability from the model, while practical applications—such as menu identification, nutritional analysis, and dish recommendation—are typically deployed on mobile devices [10], [11]. These use cases impose stringent requirements on both low-latency response and flexible deployment. Thus, designing retrieval systems that reduce online computational and communication costs while preserving semantic discrimination is essential.

To address this, Asymmetric Image Retrieval (AIR) [12]–[14] has emerged as a promising paradigm. AIR deploys a lightweight network on the client side to extract query features, which are then transmitted to the server for matching against gallery features pre-computed by a large-scale model. This design reduces both transmission cost and server-side inference load, making it more suitable for real-world deployment. Conceptually, AIR can be regarded as implicit knowledge distillation: the gallery features are generated by a powerful teacher model, while the query features are produced by a lightweight student model. The key challenge is how to transfer semantic capability from the teacher to the student under structural heterogeneity and representational asymmetry [15]. Knowledge distillation has made notable progress in both relation-based and feature-based strategies [16]–[18]. Relation-based distillation [19], [20] transfers inter-sample ranking structures, but existing methods often replicate teacher rankings indiscriminately, ignoring differences in guidance value and transferring noisy relational knowledge from ambiguous samples that even the teacher finds difficult to distinguish. This can impair fine-grained discrimination. Feature-based distillation [21], [22], which aligns intermediate features, faces another challenge: the structural disparity between large cloud models and lightweight edge models. This semantic gap means that features from corresponding layers of teacher and student do not align well, limiting transfer effectiveness.

To overcome these limitations, this paper makes the following contributions:

- We propose a novel Multi-layer and Selective Distillation framework for asymmetric retrieval that integrates multi-layer feature distillation with a decoupled differential relation distillation method based on unambiguous samples, and validate its effectiveness through real-world deployment on mobile devices.
- We propose a simple yet effective multi-layer feature distillation method designed to bridge the semantic gap caused by structural disparities between teacher and student networks, thereby enabling effective knowledge transfer.
- We further propose a decoupled differential relation dis-

Shijie Zhang, Guorui Sheng and Tao Yao are with the School of Computer and Artificial Intelligence, Ludong University, Yantai 264025, China. E-mail:zhangshijie@m.lde.edu.cn,shengguorui@ldu.edu.cn,yaotao@ldu.edu.cn

Weiqing Min and Shuqiang Jiang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100190, China. E-mail:minweiqing@ict.ac.cn,sqjiang@ict.ac.cn

Corresponding author: Guorui Sheng E-mail: shengguorui@ldu.edu.cn.

tillation method based on unambiguous samples to ensure ranking order consistency between the lightweight query network and the large-scale gallery network.
- Experiments on multiple image retrieval benchmarks demonstrate that our proposed method significantly boosts retrieval accuracy and outperforms existing state-of-the-art distillation strategies, all while maintaining the lightweight nature of the student model.

II. Related Work

A. Traditional Asymmetric Image Retrieval Methods

Asymmetric image retrieval (AIR) has emerged as an effective paradigm for reducing online computation and communication overhead in practical retrieval systems [12]–[14], [23]. Unlike conventional symmetric retrieval, AIR deploys a lightweight query encoder on the client side while using a large gallery encoder on the server side, where gallery features can be pre-computed offline. This design is particularly attractive for edge-device applications, since it significantly reduces query-side latency and server-side inference cost while preserving compatibility with high-quality gallery representations.

Early AIR methods mainly focused on improving the compatibility between lightweight query features and powerful gallery features through direct representation transfer. AML is one of the pioneering works in this direction, which teaches the lightweight student to mimic the teacher's retrieval representation via metric learning [12]. Later methods further recognized that retrieval depends not only on individual feature quality but also on inter-sample structure. Therefore, CSD and RAML extended AIR from feature imitation to joint feature and relation transfer by additionally distilling pairwise similarity knowledge [13], [24]. More recently, ranking-oriented AIR methods such as ROP and D3still relaxed strict similarity matching and instead emphasized preserving retrieval order consistency, which is often more achievable for lightweight students under asymmetric settings [14], [23]. Although these methods have substantially advanced AIR, most of them mainly operate on global representations or global ranking relations, and thus still have limited ability to handle the semantic mismatch caused by large structural disparities between teacher and student networks.

B. Feature-based Distillation Methods

Feature-based distillation transfers intermediate representations from the teacher to the student and has become a fundamental paradigm in knowledge distillation [25]. Starting from FitNet, this line of work aims to provide richer supervision than final outputs by aligning hidden features from intermediate layers [25]. In retrieval scenarios, feature distillation is particularly important because the quality of the learned embedding space directly determines matching performance. For AIR, feature-based distillation is widely used to align the lightweight query encoder with the large gallery encoder, so that the student can produce retrieval-compatible embeddings [12].

Early feature-based methods such as Zhai et al. [15] and Li et al. [26] have shown that multi-layer feature alignment can improve knowledge transfer when the teacher and student have heterogeneous architectures. These works highlight the importance of capturing intermediate semantic representations across layers, a principle which MLSD extends further by combining multi-layer feature distillation with selective relation distillation to address noisy or ambiguous samples.

C. Relation-based Distillation Methods

Relation-based distillation transfers higher-order structural knowledge by preserving the relative relationships among samples rather than matching exact feature values [27]–[29]. Representative methods such as RKD and PKT showed that transferring pairwise similarity relations can effectively improve the structural consistency of the student embedding space [27], [28]. In addition, CCKD demonstrated that correlation-aware supervision can further enhance knowledge transfer by jointly considering feature and relation information [29].

Recent works in response- and ranking-based distillation, such as Song et al. [30] and He et al. [31], extend relation-based supervision by considering ranking consistency and response alignment. However, these methods generally do not filter noisy pairwise relations or consider differences across multiple feature layers, which are critical in AIR for fine-grained tasks. MLSD addresses these limitations by combining multi-layer feature distillation with selective relation distillation based on unambiguous samples, enhancing both reliability and discriminative capability.

In summary, although existing AIR methods have achieved promising results, two challenges are still not fully addressed. On the one hand, feature-based distillation methods often perform direct alignment between teacher and student representations, which may be insufficient when the two networks are highly heterogeneous. On the other hand, relation-based or ranking-based methods usually treat all sample pairs equally, which may introduce noisy supervision from ambiguous samples. Motivated by these observations, we propose MLSD, a unified framework that combines multi-layer feature distillation and selective relation distillation for asymmetric fine-grained retrieval. A summarized comparison of representative AIR methods and our MLSD framework is provided in Table I.

III. Methodology

A. Formulation and Background

Existing asymmetric distillation methods mainly focus on either feature transfer or relation transfer. However, two issues remain insufficiently addressed in AIR: 1) the cross-layer semantic mismatch caused by the structural heterogeneity between teacher and student networks, and 2) the noisy relational supervision introduced by ambiguous samples. To address these limitations, we propose a Multi-layer and Selective Distillation (MLSD) framework, which transfers knowledge from the teacher network to the student network from two complementary aspects: semantic alignment at intermediate layers and reliable relation transfer at the retrieval level. Specifically, MLSD combines a multi-layer feature distillation strategy for intermediate representations with a selective differential similarity distillation strategy for final retrieval features. Fig. 1

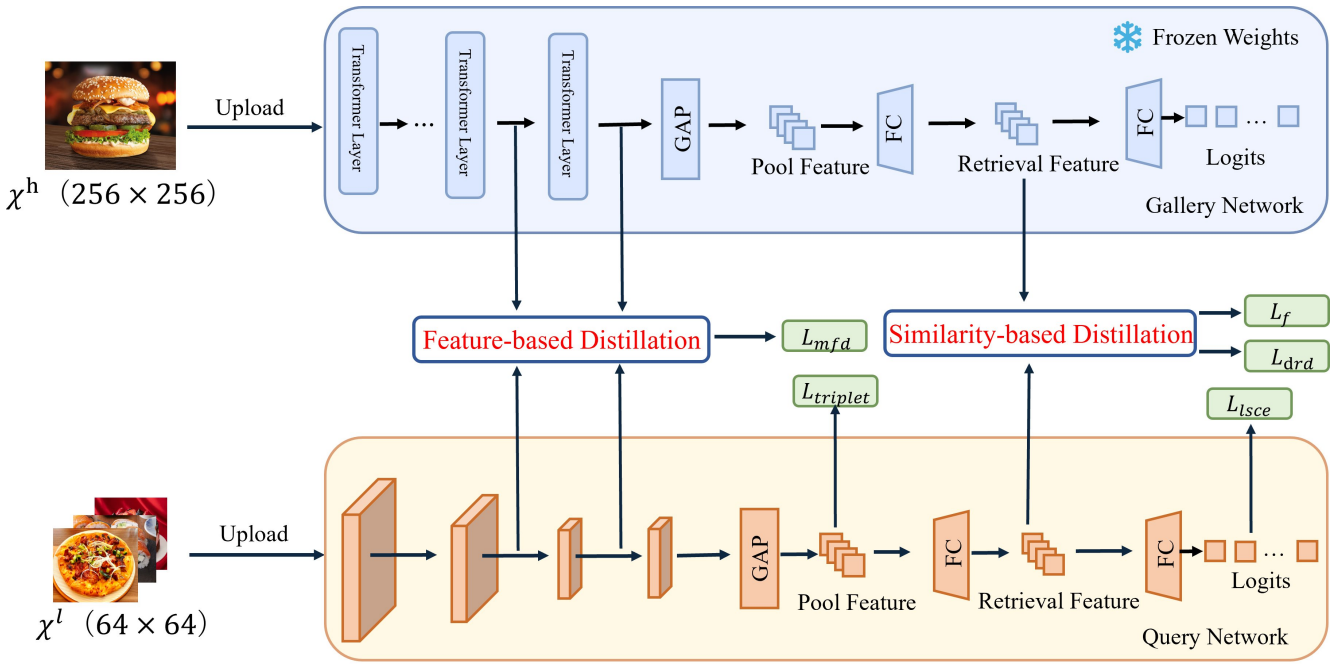


Fig. 1. Our Multi-layer and Selective Distillation framework. Here, L_{drd} denotes the overall decoupled relation distillation loss, which consists of the loss for inconsistent pairs (L_{drc}) and the loss for consistent pairs (L_{dca}).

TABLE I
COMPARISON OF REPRESENTATIVE AIR METHODS AND OUR MLSD FRAMEWORK.

Method	Feature Distillation	Relation Distillation	Multi-layer Alignment
VanillaKD [32]			
RKD [27]		✓	
PKT [28]		✓	
FitNet [25]	✓		
AML [12]	✓		
CSD [24]	✓	✓	
RAML [13]	✓	✓	
ROP [23]		✓	
CCKD [29]	✓	✓	
D3still [14]		✓	
MLSD (Ours)	✓	✓	✓

illustrates the overall knowledge transfer process between the query and gallery networks.

Compared with prior AIR distillation methods such as CSD and D3still, MLSD differs in both conceptual design and optimization objective. CSD mainly combines feature imitation and relation transfer in a holistic manner, but does not explicitly address the semantic mismatch between heterogeneous teacher and student layers. In contrast, our multi-layer feature distillation explicitly aligns complementary intermediate representations to bridge cross-layer semantic gaps. D3still mainly focuses on preserving ranking consistency, while applying relation transfer uniformly over samples. Different from this design, our method introduces a teacher-guided masking mechanism to filter ambiguous samples and further decouples relation supervision into inconsistent-pair correction and consistent-pair calibration, thereby reducing noisy supervision and negative transfer.

Given an input batch $\chi = [x_1, x_2, \ldots, x_n]$ containing n samples, we resize them into a low-resolution sample set $\chi^l = [x_1^l, x_2^l, \ldots, x_n^l]$ and a high-resolution sample set $\chi^h = [x_1^h, x_2^h, \ldots, x_n^h]$, respectively. Subsequently, we use a lightweight query network $\theta_q(\cdot)$ and a large-scale gallery network $\theta_g(\cdot)$ to transform the low-resolution and high-resolution images into normalized d-dimensional vectors, as shown below:

$$V^q = \theta_q(x^l) \in \mathbb{R}^{n \times d}, \quad V^g = \theta_g(x^h) \in \mathbb{R}^{n \times d}, \tag{1}$$

where V^q and V^g are the matrices containing the normal-

ized d-dimensional feature vectors for the query and gallery batches, respectively.

B. Multi-layer Feature Distillation

In our proposed method, we argue that an effective student network should not only mimic the teacher network in the final embedding space, but should also learn how the teacher progressively aggregates semantic information across intermediate stages. This is particularly important in AIR, where the student and teacher are structurally heterogeneous and direct final-output matching alone may be insufficient to bridge their semantic gap.

To implement multi-layer feature distillation, we select the last two intermediate layers from both the student and teacher networks and extract their corresponding feature maps. We choose these two layers because they provide complementary information: the relatively shallower layer retains richer local visual patterns, while the deeper layer captures more category-level semantic abstraction. Distilling both layers therefore helps the student align with the teacher at different representation levels.

For each pair of student and teacher feature maps, f_j^S and f_j^T, we first align their resolutions using a pooling layer, as detailed below:

$$\hat{f}_j^T = \text{AvgPool}(f_j^T) \in \mathbb{R}^{n\times k}, \quad \hat{f}_j^S = \text{AvgPool}(f_j^S) \in \mathbb{R}^{n\times k}, \tag{2}$$

where $\text{AvgPool}(\cdot)$ denotes Global Average Pooling, and $\hat{f}_j^T$ and $\hat{f}_j^S$ represent the batch of pooled feature vectors from layer j. This operation compresses spatial feature maps into compact semantic descriptors, making intermediate representations from heterogeneous backbones more comparable with negligible extra computation. These vectors are then passed through projection modules before calculating the loss:

$$L_{mfd} = \frac{1}{n \cdot L} \sum_{i=1}^{n} \sum_{j=L-1}^{L} \|\phi_g(f_{i,j}^T) - \phi_q(f_{i,j}^S)\|_2^2, \tag{3}$$

where $\phi_g(\cdot)$ and $\phi_q(\cdot)$ are projection modules, each consisting of a Batch Normalization layer, a linear layer without bias, and an activation function. The role of the projection modules is to reduce the representation gap between the teacher and student features before alignment. In this way, L_{mfd} does not simply enforce feature imitation, but encourages the student to absorb semantic cues from multiple levels, which helps bridge the cross-layer semantic mismatch caused by network heterogeneity.

C. Decoupled Differential Distillation based on Unambiguous Samples

Figure 2 illustrates the decoupled differential distillation process based on unambiguous samples. It shows the query features (c_i^q) and gallery features (c_i^g), the teacher-based masking process, and the application of decoupled differential relation distillation (L_{drc} for inconsistent pairs and L_{dca} for consistent pairs).

Although multi-layer feature distillation improves semantic alignment, retrieval performance is ultimately determined by the relative similarity structure in the embedding space. Therefore, besides feature-level transfer, the student network also needs to learn the teacher's ranking behavior. To achieve this, we compute the cosine similarity matrices between the query and gallery features within the gallery network's representation space:

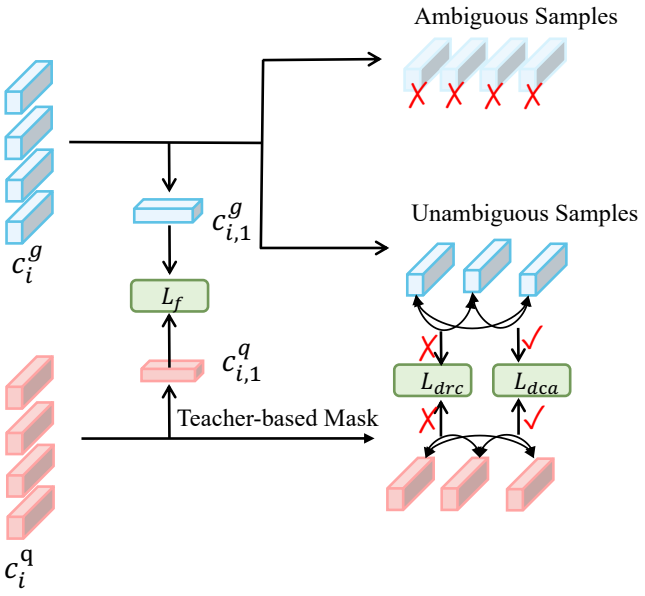


Fig. 2. Decoupled Differential Distillation based on Unambiguous Samples. Query (c_i^q) and gallery (c_i^g) features are filtered by a teacher-based binary mask, and the selected unambiguous samples are used for feature distillation (L_f) and decoupled differential relation distillation (L_{drc} for inconsistent pairs, L_{dca} for consistent pairs).

$$G^q = V^q(V^g)^T \in \mathbb{R}^{n\times n}, \quad G^g = V^g(V^g)^T \in \mathbb{R}^{n\times n}, \tag{4}$$

where G^q measures the cross-network similarity between query features and teacher-generated gallery features, and G^g characterizes the teacher's self-consistent similarity structure. Since the teacher embedding space is more stable and discriminative, it is used as the reference space for decoupled differential relation distillation.

Considering that image retrieval typically returns the Top-K relevant results, we further construct Top-K retrieval similarity matrices. Based on the similarity matrix G^g generated by the gallery network, we obtain the Top-k indices of the retrieval results, $R \in \mathbb{R}^{n\times k}$:

$$R = \text{argsort}(G^g, \text{dim} = 2), \tag{5}$$

where $\text{argsort}(\cdot)$ is a function that returns the Top-k indices by sorting the cosine similarity values along the second dimension in descending order. We use the teacher-defined Top-K neighborhood because retrieval mainly depends on highly ranked candidates rather than the full similarity matrix, and such local supervision is more consistent with the AIR objective.

Then, using the indices R, we construct the Top-k retrieval similarity matrices $C^q \in \mathbb{R}^{n\times k}$ and $C^g \in \mathbb{R}^{n\times k}$:

$$C^q = \text{sort}(G^q, \text{index} = R), \quad C^g = \text{sort}(G^g, \text{index} = R), \tag{6}$$

where sort$(\cdot)$ denotes a function that reorders the cosine similarity matrix according to the top-k indices. In this way, the student is supervised under the teacher-induced retrieval neighborhood rather than under global similarity matching.

As shown in Fig. 1, our method focuses on learning unambiguous knowledge from the teacher model. Specifically, we utilize a well-trained, unbiased Fully Connected (FC) layer from the gallery network to evaluate sample features. Here, correct teacher classification is not treated as a strict guarantee that all pairwise relations of a sample are perfectly reliable. Instead, we regard it as an efficient and practical proxy for identifying samples whose representations are more likely to lie in a class-consistent semantic region and therefore provide relatively stable local neighborhood structure for distillation. By discarding ambiguous samples, we aim to transfer more reliable semantic information from the gallery network to the query network. To implement this evaluation, we construct a binary mask m:

$$m_i = \begin{cases} 1 & \text{if } \hat{y}_i = y_i \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where y_i is the ground-truth label of the i-th sample, and $\hat{y}_i$ is the predicted label from the Fully Connected layer of the gallery network, calculated as follows:

$$\hat{y}_i = \text{argmax}\left(W^g(\theta_g(x_i^h))\right), \tag{8}$$

where argmax$(\cdot)$ is a function that returns the index of the maximum value, and $\hat{y}_i$ represents the predicted class label. $W^g \in \mathbb{R}^{M\times D}$ denotes the weights of the gallery network's Fully Connected layer, where M is the total number of classes in the training set.

Correctly classified samples are more likely to be located close to the corresponding class manifold in the embedding space and therefore preserve more reliable pairwise similarity relationships within their local neighborhood. In contrast, misclassified samples tend to lie near decision boundaries and may introduce unstable relational structures, which can negatively affect the distillation process. Consequently, filtering samples based on teacher classification correctness provides a practical criterion for selecting reliable supervision signals and stabilizing local similarity structure during relation distillation.

For the feature knowledge, we design a feature distillation loss function, L_f, based on the Top-K retrieval similarity matrices to align the representation spaces of the query and gallery networks. It is defined as follows:

$$L_f = \frac{1}{\sum_{i=1}^{n} m_i}\left(\sum_{i=1}^{n} m_i (C_{i,1}^q - C_{i,1}^g)^2\right)^{\frac{1}{2}}. \tag{9}$$

We align only the Top-1 similarity rather than the entire Top-K set for two reasons. First, the primary objective of retrieval is to rank the most relevant sample at the top position, so Top-1 alignment provides the most direct supervision for improving retrieval precision. Second, the subsequent differential relation distillation already models the relative structure within the Top-K neighborhood. Therefore, L_f focuses on the most critical anchor similarity, while the following relation losses further preserve local ranking consistency.

To model the similarity difference knowledge between sample pairs, we build upon the Top-k retrieval similarity matrices. First, we construct two differential similarity matrices, $M^q \in \mathbb{R}^{n\times(k-1)\times(k-1)}$ and $M^g \in \mathbb{R}^{n\times(k-1)\times(k-1)}$, which are defined as follows:

$$M_{i,j,l}^q = C_{i,j+1}^q - C_{i,l+1}^q, \quad 1 \le j,l \le k-1, \tag{10}$$

$$M_{i,j,l}^g = C_{i,j+1}^g - C_{i,l+1}^g, \quad 1 \le j,l \le k-1, \tag{11}$$

where each element describes the relative similarity gap between two candidates within the local Top-K neighborhood. Compared with directly matching absolute similarity values, such differential modeling is more suitable for retrieval because it explicitly captures local ranking behavior.

Subsequently, we separately calculate the differential distillation loss functions for sample pairs with inconsistent and consistent ranking directions. Specifically, we decouple the differential distillation loss into a loss for inconsistent pairs, L_{drc}, and a loss for consistent pairs, L_{dca}. They are defined as follows:

$$L_{drc} = \frac{1}{n}\sum_{i=1}^{n} m_i \left(\sum_{j=1}^{k-1}\sum_{l=1}^{k-1} \mathcal{H}\left(-\frac{M_{i,j,l}^q}{M_{i,j,l}^g}\right)\left(\frac{M_{i,j,l}^q - M_{i,j,l}^g}{\mu + |M_{i,j,l}^g|}\right)^2\right)^{\frac{1}{2}}, \tag{12}$$

$$L_{dca} = \frac{1}{n}\sum_{i=1}^{n} m_i \left(\sum_{j=1}^{k-1}\sum_{l=1}^{k-1} \mathcal{H}\left(\frac{M_{i,j,l}^q}{M_{i,j,l}^g}\right)\left(\frac{M_{i,j,l}^q - M_{i,j,l}^g}{\mu + |M_{i,j,l}^g|}\right)^2\right)^{\frac{1}{2}}, \tag{13}$$

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

This decoupled design distinguishes two different learning situations. For consistent ranking pairs, the student network already preserves the same ordering direction as the teacher, and the optimization mainly serves to refine the relative similarity magnitudes; this contribution is measured and optimized via L_{dca}. In contrast, inconsistent ranking pairs correspond to cases where the student fails to preserve the teacher's ordering, thereby providing stronger corrective supervision, which is measured and optimized via L_{drc}. Since these inconsistent pairs carry more informative hard knowledge for correcting ranking errors, L_{drc} is assigned a larger weight in the total loss, while consistent pairs remain useful for stabilizing local similarity calibration and are thus preserved with a smaller weight via L_{dca}. In this sense, our relation distillation is not a uniform ranking-consistency constraint as in D3still, but a selective and decoupled correction–calibration mechanism that explicitly accounts for supervision reliability and pairwise learning difficulty.

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

$$L_{\text{student}} = \alpha L_f + \beta L_{drc} + \gamma L_{dca} + \delta L_{mfd} + \epsilon L_{\text{triplet}} + \zeta L_{\text{lsce}}, \tag{14}$$

To further optimize the feature learning effectiveness of the student network in asymmetric image retrieval, we apply a triplet loss, L_{triplet}, on the post-pooling features, and compute a cross-entropy loss, L_{lsce}, on the logits output. Here,

TABLE II
OVERVIEW OF DATASETS USED IN OUR EXPERIMENTS

Dataset	**Classes**	**Total Samples**	**Train/Test Split**	**Characteristics**
ETH Food-101 [6]	101	101,000	75,750 / 25,250	Fine-grained food recognition; relatively balanced classes
Vireo-Food172 [7]	172	110,241	77,576 / 32,665	Real-world food images; large intra-class variation
Clothes Retrieval (In-Shop) [33]	11,735	52,712	34,997 / 17,715	Product image retrieval; background noise and viewpoint variation
Stanford Online Products (SOP) [34]	22,634	120,053	59,551 / 60,502	Large-scale product retrieval; high intra-class similarity, low inter-class variance

TABLE III
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
$L_f + L_{mfd}$	0.25	68.67	83.65	53.68	56.98
$L_f + L_{mfd} + L_{dca}$	0.25	68.98	84.03	54.57	58.11
$L_f + L_{mfd} + L_{drc}$	0.25	69.24	84.41	55.02	58.96
$L_f + L_{mfd} + L_{dca} + L_{drc}$	0.25	69.39	84.82	56.13	60.51

$\alpha, \beta, \gamma, \delta, \epsilon,$ and ζ are hyperparameters used to tune the contribution of each component in the total loss function.

Among them, β and γ are specifically related to the decoupled differential relation distillation, where β controls the weight of inconsistent ranking pairs through L_{drc} and γ controls the weight of consistent ranking pairs through L_{dca}. Therefore, we provide a detailed analysis of these two parameters. The remaining coefficients, α, δ, ϵ, and ζ, correspond to the feature distillation loss, multi-layer feature distillation loss, triplet loss, and cross-entropy loss, respectively. These terms mainly serve as standard balancing coefficients to ensure stable feature alignment and task-oriented learning. Based on validation experiments, these coefficients are fixed because they yield stable performance. Consequently, in the following analysis, we focus on β and γ, while keeping the other parameters constant to ensure experimental reproducibility and stability.

IV. EXPERIMENTS

In this section, we conduct experiments on four widely-used datasets to verify the superiority of our method. These datasets include ETH Food-101 [6], Vireo-Food172 [7], Clothes Retrieval (In-Shop) [33], and Stanford Online Products (SOP) [34].

We selected these datasets because they provide standard benchmarks widely used in the retrieval community and also represent practical real-world applications, including fine-grained food recognition and product retrieval. Table II summarizes key characteristics of each dataset, including the number of classes, total samples, class imbalance, and training/testing splits.

All images are normalized and augmented with random cropping and horizontal flipping. Query and gallery images use asymmetric resolutions to reflect real deployment scenarios (e.g., 64×64 for query, 256×256 for gallery). Hyperparameters α, β, γ, δ, ϵ, and ζ were tuned on the validation set. Evaluation metrics include Top-1 accuracy and mean Average Precision (mAP). Baselines for comparison include VanillaKD, RKD, PKT, FitNet, AML, CSD, RAML, ROP, and D3still, representing classical distillation approaches and recent asymmetric retrieval methods.

Beyond reporting numerical results, we analyze the strengths and limitations of MLSD. Selective relation distillation improves stability for non-ambiguous samples, while multi-layer feature alignment helps capture semantic structures for classes with high intra-class variance. For queries in sparse classes or with extreme intra-class variability, performance improvement is less pronounced. Real-world deployment on Pixel 4 demonstrates that MLSD achieves high retrieval accuracy while maintaining efficiency.

A. Datasets and Performance Metrics

1) Product Retrieval Datasets: ETH Food-101 [6] is a dataset of Western food, containing 101,000 images from 101 classes. The training set consists of 750 images per class, for a total of 75,750 images. The test set consists of 250 images per class, for a total of 25,250 images.

Vireo Food-172 [7] is a large-scale food image dataset containing 110,241 images from 172 classes. The training set consists of 66,071 images across 172 classes, and the test set consists of 44,170 images across the same 172 classes.

In-Shop Clothes Retrieval (In-Shop) [33] is a clothes retrieval dataset containing 72,712 images covering 7,986 classes. The training set consists of 25,882 images from 3,997 classes. The test set is divided into a query set, composed of 14,218 images belonging to 3,985 classes, and a gallery set, which contains 12,612 images from the same 3,985 classes.

Stanford Online Products (SOP) [34] is a widely-used product recognition dataset, containing 120,053 product images

TABLE IV
COMPARISON OF DIFFERENT METHODS ON FOOD101 AND VIREO FOOD-172 DATASETS.

METHOD	Query net	Query input	Gallery net	Gallery input	MP(M)	FLOPs(G)	ETH Food-101		Vireo-Food172	
							mAP(%)	R1(%)	mAP(%)	R1(%)
(A) Symmetric Baselines (No Distillation)										
ResNet-101	ResNet-101	256×256	ResNet-101	256×256	43.55	10.27	82.45	86.14	80.10	84.88
SwinT-V2	SwinT-V2	256×256	SwinT-V2	256×256	49.36	8.49	76.86	87.13	71.64	86.63
(B) Distillation from ResNet-101 (Teacher) to ResNet-18 (Student)										
VanillaKD	ResNet-18	64×64	ResNet-101	256×256	11.44	0.25	1.78	0.99	1.02	1.74
RKD					11.44	0.25	1.41	0.99	0.83	0.58
PKT					11.44	0.25	1.47	0.99	0.94	0.00
FitNet					11.44	0.25	41.16	49.50	46.18	55.81
CSD					11.44	0.25	53.24	53.47	63.98	66.86
RAML					11.44	0.25	50.13	51.49	65.10	68.02
ROP					11.44	0.25	50.87	51.49	58.12	63.37
CCKD					11.44	0.25	51.38	50.50	65.73	70.35
D3still					11.44	0.25	55.16	56.44	64.65	70.35
Ours					11.44	0.25	**56.13**	**60.51**	**67.29**	**70.98**
(C) Distillation from SwinT-V2 (Teacher) to ResNet-18 (Student)										
VanillaKD	ResNet-18	64×64	SwinT-V2	256×256	11.44	0.25	1.68	0.99	0.88	0.58
RKD					11.44	0.25	2.32	0.99	0.89	1.16
PKT					11.44	0.25	1.35	0.99	0.91	0.00
FitNet					11.44	0.25	44.59	53.47	51.73	61.05
CSD					11.44	0.25	53.15	58.42	62.02	70.35
RAML					11.44	0.25	53.85	56.44	62.50	69.19
ROP					11.44	0.25	49.86	58.42	60.14	70.93
CCKD					11.44	0.25	54.59	61.39	61.28	70.35
D3still					11.44	0.25	52.36	62.38	61.65	70.35
Ours					11.44	0.25	**57.11**	**63.28**	**62.72**	**74.55**
(D) Distillation from ResNet-101 (Teacher) to MobileNetV3 (Student)										
VanillaKD	MobileNetV3	64×64	ResNet-101	256×256	1.22	0.009	1.76	0.99	0.82	0.58
RKD					1.22	0.009	2.21	0.99	0.81	0.58
PKT					1.22	0.009	1.67	0.99	1.12	1.16
FitNet					1.22	0.009	27.64	43.56	28.34	27.91
CSD					1.22	0.009	50.02	51.49	**57.06**	56.40
RAML					1.22	0.009	46.44	47.52	55.96	**58.72**
ROP					1.22	0.009	43.30	46.53	39.50	45.35
CCKD					1.22	0.009	46.89	49.50	47.11	49.42
D3still					1.22	0.009	46.08	48.51	46.84	51.74
Ours					1.22	0.009	**51.81**	**55.35**	55.21	58.63

covering 22,634 classes. The training set consists of 59,551 images from 11,318 classes, while the test set contains 60,502 images from the remaining 11,316 classes.

2) Performance Metrics: During the inference phase, following prior retrieval work, we rank the corresponding gallery images using the cosine similarity between the query features and the gallery features. Specifically, gallery images with higher cosine similarity scores are ranked higher. Furthermore, we use mean Average Precision (mAP) and Rank-1 accuracy (R1) to evaluate the performance.

B. Implementation Details

In Equation (14), we set the hyperparameters as follows: $\alpha = 200, \beta = 5, \gamma = 1, \delta = 8, \epsilon = 1$, and $\zeta = 1$. The software tools used for the experiments are PyTorch 2.0.1 and CUDA 11.8. The hardware platform is a single NVIDIA A800 80GB PCIe GPU. For network training, we select various models pre-trained on ImageNet (e.g., ResNet101 [35], SwinT-V2) as the teacher networks, and correspondingly select another set of lightweight models (e.g., ResNet18 [35], MobileNetV3) as the query networks.

The input image resolutions for the query and teacher networks are 64×64 and 256×256, respectively. This asymmetric input design is crucial for achieving computational efficiency in the asymmetric image retrieval (AIR) framework, allowing the lightweight query network to perform fast inference on the client side while leveraging the high-capacity gallery network on the server side. Moreover, this design reduces transmission bandwidth since only the compact query features need to be sent to the server for retrieval.

Data augmentation includes z-score normalization, random cropping, random erasing [36], [37], and random horizontal flipping, as in [38], [39]. The probabilities for both horizontal flipping and random erasing are set to 0.5. We use mini-batch stochastic gradient descent [40] as the optimizer, with a batch size of 512. The weight decay [41], [42] is set to 5×10^{-4} and the momentum to 0.9. We employ a cosine annealing schedule [43] with a linear warm-up strategy [44] to adjust the learning rate. Specifically, the learning rate is initialized to 1×10^{-3}, then linearly warmed up to 1×10^{-2} over the first 10 epochs. The total number of training epochs is 120.

TABLE V
COMPARISON OF DIFFERENT DISTILLATION METHODS ON THE IN-SHOP AND STANFORD ONLINE PRODUCTS (SOP) [34] DATASETS. OUR METHOD IS EVALUATED AGAINST BASELINES AND SOTA TECHNIQUES ACROSS VARIOUS TEACHER-STUDENT BACKBONE CONFIGURATIONS.

METHOD	Query net	Query input	Gallery net	Gallery input	MP(M)	FLOPs(G)	In-Shop		SOP	
							mAP(%)	R1(%)	mAP(%)	R1(%)
(A) Symmetric Baselines (No Distillation)										
ResNet-101	ResNet-101	256×256	ResNet-101	256×256	43.55	10.27	81.96	95.42	72.06	86.92
SwinT-V2	SwinT-V2	256×256	SwinT-V2	256×256	49.36	8.49	80.28	94.55	74.22	88.00
(B) Distillation from ResNet-101 (Teacher) to ResNet-18 (Student)										
VanillaKD	ResNet-18	64×64	ResNet-101	256×256	11.44	0.25	0.15	0.02	0.04	0.00
RKD					11.44	0.25	0.15	0.06	0.03	0.00
PKT					11.44	0.25	0.13	0.02	0.04	0.02
FitNet					11.44	0.25	65.99	80.50	48.87	65.35
CSD					11.44	0.25	66.64	81.00	49.43	65.96
RAML					11.44	0.25	67.18	81.85	49.46	66.24
ROP					11.44	0.25	65.58	80.24	48.03	64.66
CCKD					11.44	0.25	66.60	81.21	49.11	66.05
D3still					11.44	0.25	68.56	83.96	51.12	68.42
Ours					11.44	0.25	**69.39**	**84.82**	**52.25**	**69.63**
(C) Distillation from SwinT-V2 (Teacher) to ResNet-18 (Student)										
VanillaKD	ResNet-18	64×64	SwinT-V2	256×256	11.44	0.25	0.13	0.03	0.03	0.01
RKD					11.44	0.25	0.14	0.04	0.04	0.02
PKT					11.44	0.25	0.16	0.04	0.03	0.00
FitNet					11.44	0.25	56.35	65.77	37.06	46.57
CSD					11.44	0.25	57.58	67.87	40.50	51.63
RAML					11.44	0.25	57.34	67.42	40.64	51.97
ROP					11.44	0.25	53.87	63.52	37.90	49.45
CCKD					11.44	0.25	56.55	65.51	40.27	51.94
D3still					11.44	0.25	60.19	72.07	43.32	56.98
Ours					11.44	0.25	**62.23**	**74.96**	**44.10**	**57.57**
(D) Distillation from ResNet-101 (Teacher) to MobileNetV3 (Student)										
VanillaKD	MobileNetV3	64×64	ResNet-101	256×256	1.22	0.009	0.16	0.06	0.03	0.00
RKD					1.22	0.009	0.15	0.03	0.03	0.00
PKT					1.22	0.009	0.15	0.08	0.04	0.01
FitNet					1.22	0.009	60.41	74.61	44.80	60.30
CSD					1.22	0.009	62.27	76.48	44.98	60.72
RAML					1.22	0.009	62.29	76.13	45.52	61.48
ROP					1.22	0.009	61.43	75.91	43.67	59.35
CCKD					1.22	0.009	61.53	76.25	44.37	60.09
D3still					1.22	0.009	63.58	79.43	45.80	62.72
Ours					1.22	0.009	**64.39**	**80.36**	**46.75**	**63.77**

C. Ablation Study

As shown in Table. III, we conducted ablation studies on the ETH Food-101 [6] and In-shop [33] datasets.

"Gallery" refers to the direct evaluation of the ResNet101's retrieval performance. "L_f" denotes the student network learning feature representation knowledge from the teacher network. "$L_f + L_{mfd}$" indicates the student network learning both feature representation knowledge and Multi-layer knowledge. "$L_f + L_{mfd} + L_{drc} + L_{dca}$" represents the transfer of feature representation knowledge, Multi-layer knowledge, and the relational knowledge of both inconsistent and consistent pairwise similarity differences.

As can be clearly seen from Table. III, compared to the symmetric image retrieval method, the asymmetric approach significantly reduces the computational burden of the query network. Specifically, our asymmetric method drastically reduces the inference computation of the query network from 12.99 GFLOPs to 0.25 GFLOPs, making it possible to deploy the query network on computationally constrained edge devices. Furthermore, we observe that introducing multi-layer knowledge effectively improves retrieval performance. For instance, on the In-Shop [33] dataset, adding L_{mfd} improves mAP by 1.53% and R1 by 1.7%. On ETH Food-101 [6], mAP improves by 2.6% and R1 by 4.45%. By further incorporating L_{drc} and L_{dca}, mAP on the In-Shop [33] dataset increases by an additional 0.66% and R1 by 1.18%; on the ETH Food-101 dataset [6], mAP increases by 2.45% and R1 by 3.53%.

In this section, we conduct comparative experiments between our MLSD framework and state-of-the-art methods to evaluate the advantages of our proposed approach in asymmetric image retrieval. To ensure a fair comparison, we re-implemented eight prior knowledge distillation techniques for asymmetric image retrieval, as they originally had different training configurations. A detailed comparative analysis follows.

As seen from Table. IV and V, purely relational distillation methods (such as RKD [27], PKT [28], and VanillaKD [32]) show initial promise in improving the performance of the query network in asymmetric image retrieval. However, their

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3718659

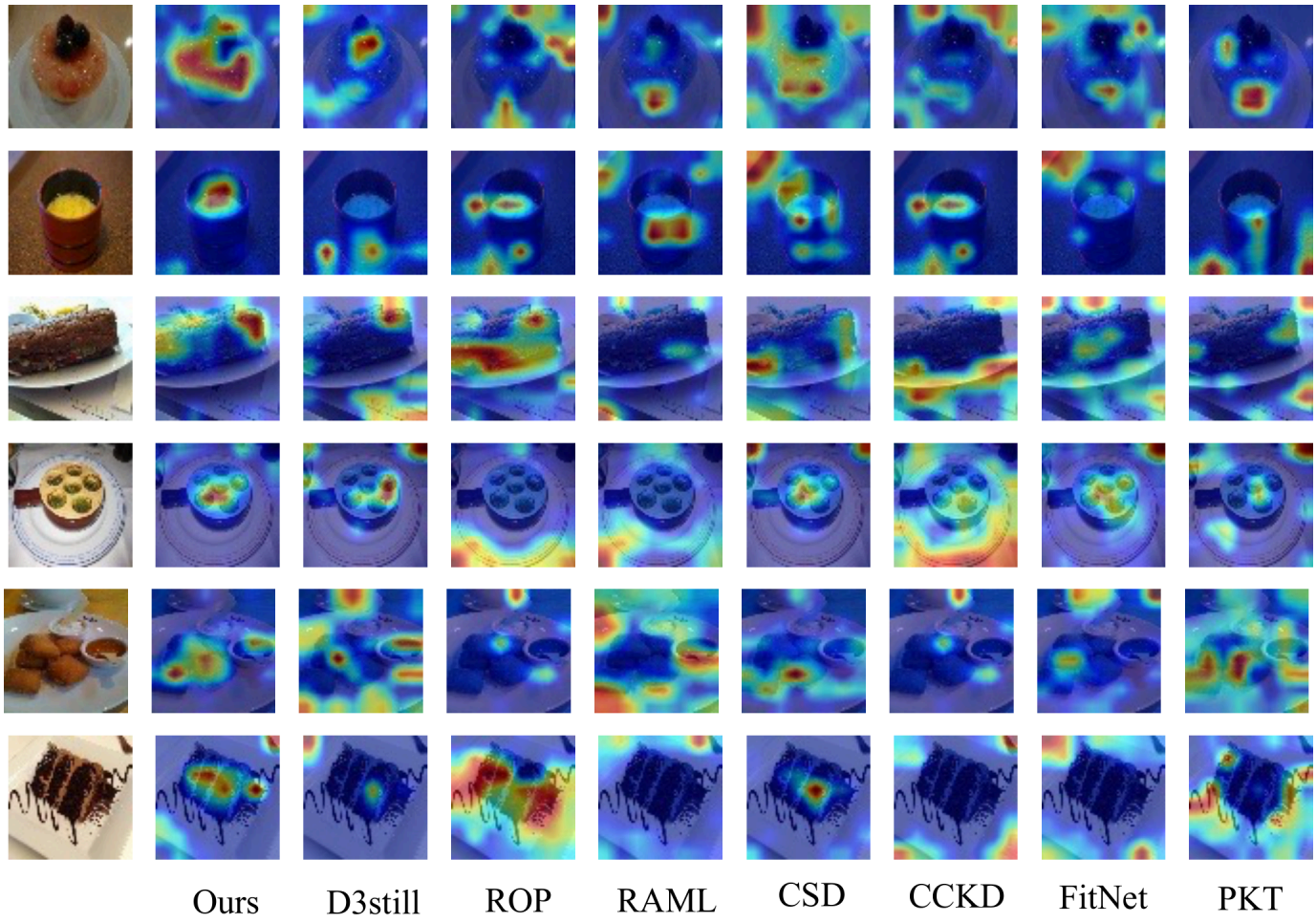


Fig. 3. Grad-CAM visualizations of different methods on the ETH Food-101 dataset.

limitation is that they only transfer relational knowledge while neglecting feature knowledge, which prevents the feature representation spaces of the two networks from being aligned. For example, when using ResNet-18 [35] as the query network and ResNet-101 [35] as the gallery network on ETH Food-101 [6], VanillaKD [32] only achieves 1.78% in mAP and 0.99% in R1.

In contrast, MLSD exhibits superior performance. For instance, with ResNet-101 [35] as the gallery network and ResNet-18 [35] as the query network, our method improves mAP by 0.97% and R1 by 4.07% over the second-best method on the ETH Food-101 [6] dataset; on the Vireo-Food172 [7] dataset, mAP is improved by 1.56% and R1 by 0.63%. When the gallery network is SwinT-V2 [45] and the query network is ResNet-18 [35], our method boosts mAP by 2.52% and R1 by 1.89% on ETH Food-101 [6]; on Vireo-Food172 [7], mAP is improved by 0.22% and R1 by 5.36%. Furthermore, when using ResNet-101 [35] as the gallery network and MobileNetV3 [46] as the query network, our method improves mAP and R1 on ETH Food-101 by 1.79% and 3.86%, respectively. It is noteworthy that the performance of our method on the Food-172 dataset [7] is not optimal when using MobileNetV3 as the query network and ResNet101 [35] as the gallery network. This suggests that effectively balancing and transferring knowledge of different granularities remains a challenge in certain datasets and network configurations.

We also validated the effectiveness of our method on two general-purpose image retrieval datasets: In-Shop [33] and SOP [34]. For instance, with a ResNet-101 gallery network and a ResNet-18 [35] query network, our method improves mAP by 0.83% and R1 by 0.86% on In-Shop [33], and by 1.13% and 1.21% on SOP [34], respectively. When the gallery network is SwinT-V2 [45] and the query network is ResNet-18, the improvements on In-Shop [33] are 2.04% for mAP and 2.89% for R1, while on SOP [34], the gains are 0.78% and 0.59%, respectively. Furthermore, using a ResNet-101 [35] gallery and a MobileNetV3 [46] query network, our approach boosts mAP and R1 by 0.81% and 0.93% on In-Shop, and by 0.95% and 1.05% on SOP [34]. These consistent results on general-purpose datasets strongly demonstrate the powerful generalization ability and leading performance of our proposed framework.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3718659

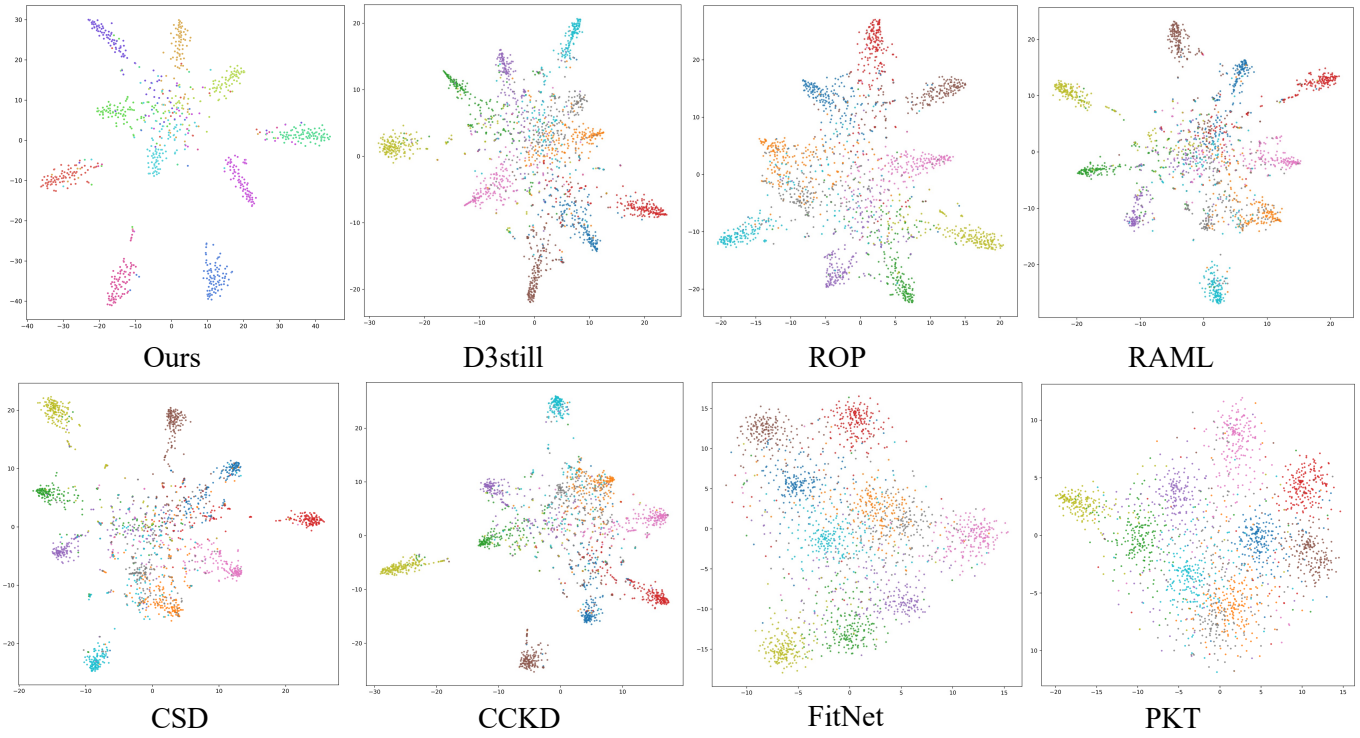


Fig. 4. t-SNE visualizations of different methods on the ETH Food-101 dataset.

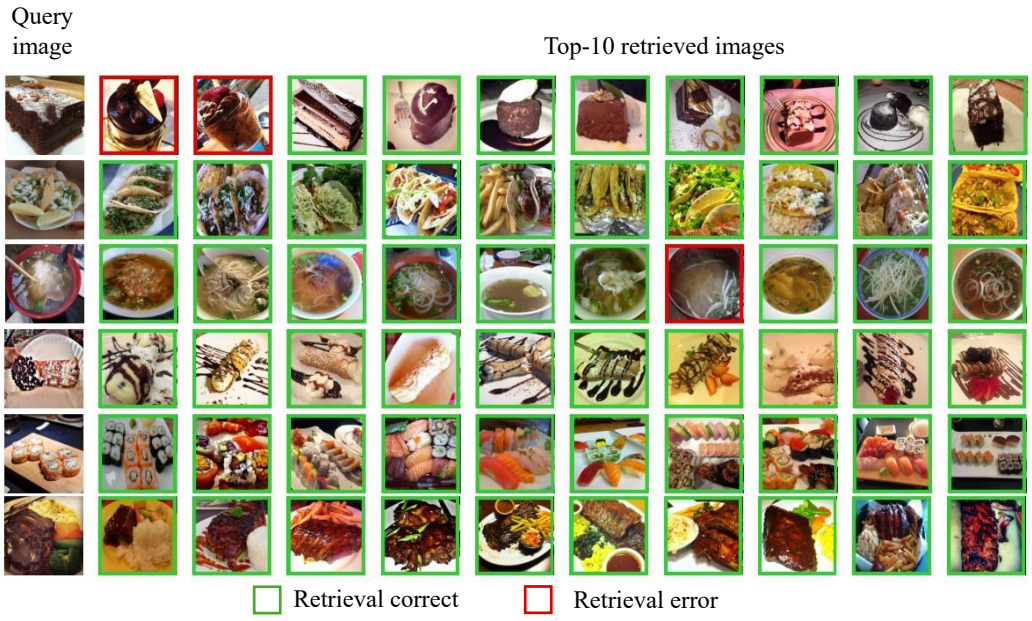
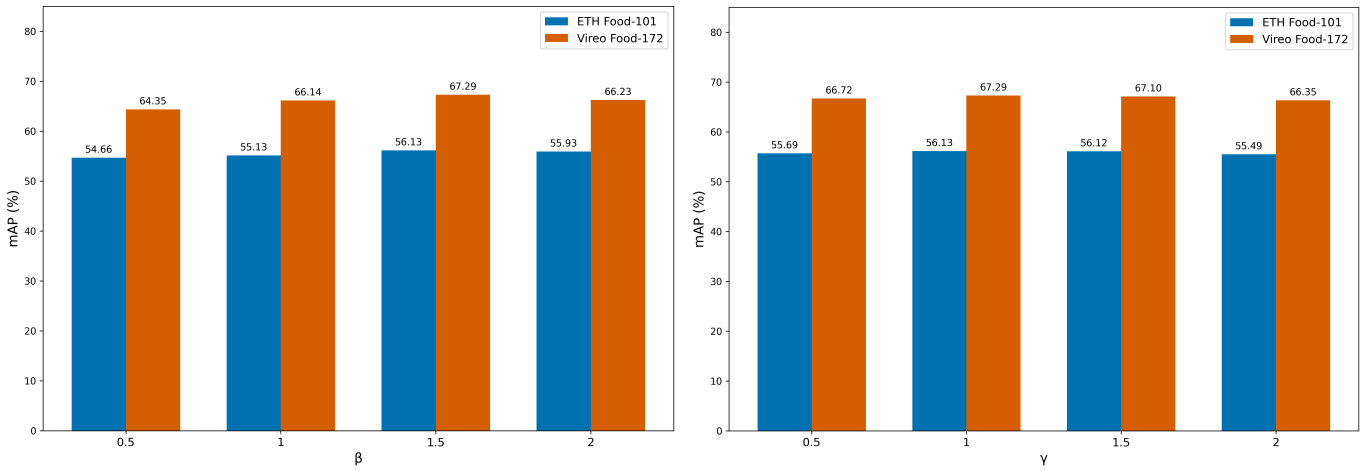


Fig. 5. Top-10 retrieval results on the ETH Food-101 dataset.

V. Visualization

To provide a more intuitive and in-depth understanding of our model's capabilities, we conducted a series of qualitative experiments on the ETH Food-101 dataset. We employed Grad-CAM for visualizing model attention, t-SNE for analyzing the feature embedding space, and presented qualitative retrieval results to assess real-world performance. As illustrated in Fig. 3, the Grad-CAM visualizations offer a clear comparison of the attention mechanisms between our method and competing approaches. It is evident that competing approaches often produce diffuse heatmaps or incorrectly focus on irrelevant background elements. In contrast, our MLSD framework, by synergistically integrating multi-layer semantic guidance with knowledge from unambiguous samples, learns to precisely localize the key discriminative regions of the

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3718659



(a) The influence of β value on mAP (%) performance. (b) The influence of γ value on mAP (%) performance.

Fig. 6. Analysis of hyperparameter influence. (a) The effect of varying the β parameter. (b) The effect of varying the γ parameter.

target food item. This focused attention allows the model to effectively filter out background noise and concentrate on the features most salient for fine-grained distinction, which directly contributes to its superior retrieval accuracy.

Furthermore, the t-SNE visualization in Fig. 4 reveals the structural quality of the feature space learned by different methods. The feature embeddings produced by our model exhibit significantly more desirable properties: tighter, more compact intra-class clusters and greater, more distinct inter-class separability. This demonstrates that our method successfully learns a more robust and discriminative embedding space, where variations within the same category are minimized while differences between separate categories are maximized. This well-structured feature space is fundamental to achieving high performance in metric learning-based retrieval tasks.

Finally, a qualitative assessment of retrieval performance is presented in Fig. 5. We randomly selected six query images from different categories to showcase the Top-10 retrieved results. In the figure, the first column contains the query images, while the subsequent columns display the ranked results from the gallery. Correct matches are indicated with a green checkmark, and incorrect ones with a red cross. The results highlight our model's strong generalization capabilities, as it consistently retrieves correct same-class images despite significant variations in viewpoint, lighting, and preparation style. Nevertheless, the model is not without its limitations, as evidenced by occasional failure cases. For instance, in the first row, the query for chocolate_cake returns chocolate_mousse in its top results. This confusion underscores the inherent difficulty of ultra-fine-grained retrieval, where certain categories share a very high degree of visual similarity, posing a challenge even for advanced models.

A. Hyper-parameter Analysis

We conduct a sensitivity analysis on the key hyperparameters β and γ to determine their optimal values. As shown in Fig. 6(a) and Fig. 6(b), the experimental results demonstrate

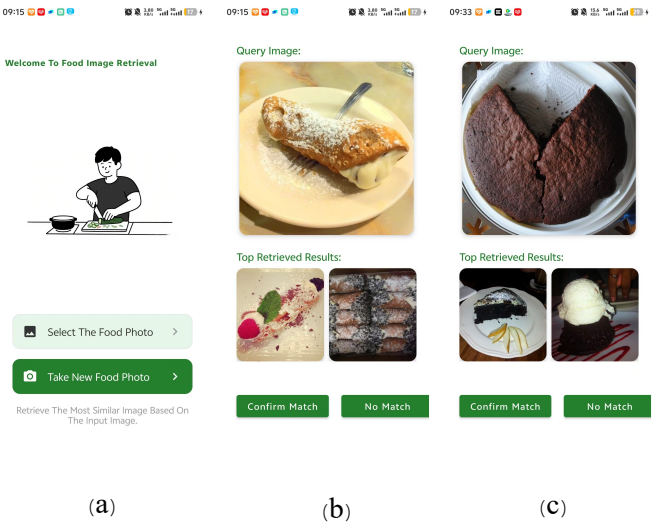


Fig. 7. Examples of the user interface for the food image retrieval application.

that the model achieves the best performance on both the ETH Food-101 [6] and Vireo Food-172 [7] datasets when β is set to 5. Similarly, the model's performance peaks when γ is set to 1. Therefore, in our final configuration, we set β and γ to 5 and 1, respectively.

VI. Real-World Demonstration

To comprehensively evaluate the real-world retrieval performance of our distilled model, we deployed its lightweight version (equipped with ResNet-18) onto a Pixel 4 mobile device for field testing. A demonstration of the application is shown in Fig. 5. The testing procedure simulates a real user scenario: first, we used a powerful ResNet-101 model to perform offline feature extraction on the ETH Food-101 [6] dataset to construct the retrieval gallery. Subsequently, we executed online retrieval on the Pixel 4 device (Processor: Qualcomm Snapdragon 855, RAM: 6GB) with the deployed

distilled model, using 300 randomly sampled images as queries to examine the model's on-device deployment performance. The query method supports both real-time camera capture and selection from the local gallery. As shown in Fig. 7(b) and Fig. 7(c), the Top-1 and Top-2 results returned by the model both exhibit precise matching performance, which fully verifies our method's high accuracy and strong adaptability in real-world application scenarios. Notably, the model's average on-device inference time is only 220 millisecond per image, demonstrating excellent real-time processing capability and fully meeting the rapid response demands of mobile applications.

VII. Conclusion

In this paper, we introduced a novel framework for asymmetric image retrieval, termed Multi-layer and Selective Distillation (MLSD). This framework addresses the critical challenges in knowledge transfer from a powerful teacher network to a lightweight student network by proposing two key innovations. First, our multi-layer feature fusion mechanism effectively bridges the semantic gap arising from the structural differences between the two networks. By integrating intermediate features, the student model not only mimics the teacher's final output but also learns its deep-seated knowledge in feature aggregation. Second, we proposed a decoupled differential relation distillation method based on unambiguous samples. This method uses a binary mask m to identify and filter out ambiguous samples, focusing on distilling pairwise similarity differential relations from reliable samples. This approach maintains ranking consistency and mitigates the negative effects of noisy data, which is crucial for fine-grained retrieval tasks. Our extensive experiments on various benchmark datasets demonstrate that the proposed MLSD framework significantly surpasses state-of-the-art methods in retrieval accuracy while maintaining the lightweight nature of the student model. Furthermore, a real-world deployment on a mobile device validated the robustness and real-time capability of our method. For future work, we suggest exploring more advanced feature fusion techniques, refining the unambiguous sample selection strategy, and extending this framework to other asymmetric learning scenarios.

References

[1] X. Jiang, H. Tang, and Z. Li, "Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[2] X. Jiang, H. Tang, R. Yan, J. Tang, and Z. Li, "Dvf: Advancing robust and accurate fine-grained image retrieval with retrieval guidelines," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2379–2388.

[3] C. H. Song, J. Yoon, S. Choi, and Y. Avrithis, "Boosting vision transformers for image retrieval," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 107–117.

[4] M. Kweon and J. Park, "Ultron: Unifying local transformer and convolution for large-scale image retrieval," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 4000–4016.

[5] H. Huang, W. Zhan, G. Min, Z. Duan, and K. Peng, "Mobility-aware computation offloading with load balancing in smart city networks using mec federation," *IEEE Transactions on Mobile Computing*, vol. 23, no. 11, pp. 10 411–10 428, 2024.

[6] L. Bossard, M. Guillaumin, and L. Van Gool, "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*. Springer, 2014, pp. 446–461.

[7] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 32–41.

[8] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *European Conference on Computer Vision*. Springer, 2014, pp. 3–17.

[9] W. Min, Z. Wang, Y. Liu, M. Luo, L. Kang, X. Wei, X. Wei, and S. Jiang, "Large scale visual food recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9932–9949, 2023.

[10] M. Wen, J. Song, W. Min, W. Xiao, L. Han, and S. Jiang, "Multi-state ingredient recognition via adaptive multi-centric network," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 5692–5701, 2023.

[11] S. Knez and L. Šajn, "Food object recognition using a mobile device: evaluation of currently implemented systems," *Trends in Food Science & Technology*, vol. 99, pp. 460–471, 2020.

[12] M. Budnik and Y. Avrithis, "Asymmetric metric learning for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8228–8238.

[13] P. Suma and G. Tolias, "Large-to-small image resolution asymmetry in deep metric learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1451–1460.

[14] Y. Xie, Y. Lin, W. Cai, X. Xu, H. Zhang, Y. Du, and S. He, "D3still: Decoupled differential distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 181–17 190.

[15] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Transactions on circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 742–753, 2020.

[16] G. Ke, B. Wang, X. Wang, and S. He, "Rethinking multi-view representation learning via distilled disentangling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 774–26 783.

[17] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 731–15 740.

[18] Q. Wang, L. Liu, W. Yu, S. Chen, J. Gong, and P. Chen, "Bckd: block-correlation knowledge distillation," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 3225–3229.

[19] W. Zhang, D. Liu, W. Cai, and C. Ma, "Cross-view consistency regularisation for knowledge distillation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2011–2020.

[20] C. Wang, J. Zhong, Q. Dai, R. Li, Q. Yu, and B. Fang, "Local structure consistency and pixel-correlation distillation for compact semantic segmentation," *Applied Intelligence*, vol. 53, no. 6, pp. 6307–6323, 2023.

[21] J. Yuan, M. H. Phan, L. Liu, and Y. Liu, "Fakd: Feature augmented knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 595–605.

[22] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "Alp-kd: Attention-based layer projection for knowledge distillation," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 35, no. 15, 2021, pp. 13 657–13 665.

[23] H. Wu, M. Wang, W. Zhou, and H. Li, "A general rank preserving framework for asymmetric image retrieval," in *The Eleventh International Conference on Learning Representations*, 2023.

[24] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, "Contextual similarity distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9489–9498.

[25] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets. rxiv preprint," *arXiv preprint arXiv:1412.6550*, 2014.

[26] S. Li, M. Hu, X. Xiao, and Z. Tu, "Patch similarity self-knowledge distillation for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 5091–5103, 2023.

[27] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[28] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Transactions on*

Neural Networks and learning systems, vol. 32, no. 5, pp. 2030–2039, 2020.

[29] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5007–5016.

[30] Y. Song, P. Zhang, W. Huang, Y. Zha, and Y. Zhang, "Flexible temperature parallel distillation for dense object detection: Make response-based knowledge distillation great again," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 5, pp. 4963–4975, 2025.

[31] X. He, J. Wang, Y. Su, D. Liu, J. Zhao, and G. Lu, "Monotonic rank knowledge distillation via kendall correlation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[32] Z. Hao, J. Guo, K. Han, H. Hu, C. Xu, and Y. Wang, "Vanillakd: Revisit the power of vanilla knowledge distillation from small scale to large scale," *arXiv preprint arXiv:2305.15781*, 2023.

[33] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[34] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4004–4012.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] P. Wang, C. Ding, W. Tan, M. Gong, K. Jia, and D. Tao, "Uncertainty-aware clustering for unsupervised domain adaptive object re-identification," *IEEE Transactions on Multimedia*, vol. 25, pp. 2624–2635, 2022.

[37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[38] W. Cai, H. Zhang, X. Xu, S. He, K. Zhang, and J. Qin, "Contextual-assisted scratched photo restoration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5458–5469, 2023.

[39] Y. Xie, F. Shen, J. Zhu, and H. Zeng, "Viewpoint robust knowledge distillation for accelerating vehicle re-identification," *EURASIP Journal on Advances in Signal Processing*, vol. 2021, no. 1, p. 48, 2021.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[41] Y. Xie, H. Wu, J. Zhu, and H. Zeng, "Distillation embedded absorbable pruning for fast object re-identification," *Pattern Recognition*, vol. 152, p. 110437, 2024.

[42] W. Zheng, C. Xu, X. Xu, W. Liu, and S. He, "Ciri: curricular inactivation for residue-aware one-shot video inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 012–13 022.

[43] X. Liu, H. Zeng, Y. Shi, J. Zhu, and K.-K. Ma, "Deep rank cross-modal hashing with semantic consistent for image-text retrieval," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4828–4832.

[44] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[45] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

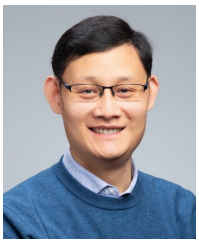
[46] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

Shijie Zhang received the B.S. degree in communication engineering from Ludong University, Yantai, China, in 2022. He is currently working towards the M.S. degree in computer science and technology at the School of Computer and Artificial Intelligence, Ludong University. His research interests include computer vision and image retrieval.

Weiqing Min (Senior Member, IEEE) is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored or coauthored more than 50 peer-referenced papers in relevant journals and conferences, including Patterns (Cell Press), ACM Computing Surveys, Trends in Food Science and Technology, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, Food Chemistry, ACM MM, AAAI, and IJCAI. His research interests include multimedia content analysis and food computing. He was a Senior Member of CCF. He was a recipient of the 2016 ACM Transactions on Multimedia Computing, Communications, and Applications, the Nicolas D. Georganas Best Paper Award, and the 2017 IEEE Multimedia Magazine Best Paper Award. He was the Guest Editor for the special issues on international journals, such as IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and Foods.

Guorui Sheng received the M.E. degree from Kunsan National University, South Korea in 2007, and received the Ph.D. degree from Nankai University, China, in 2017. He served as a research assistant to scholar Bruce Denby at the School of Computer Science and Technology, Tianjin University, from 2017 to 2018. He is currently a Lecturer at the Department of Information and Electrical Engineering, Ludong University, Yantai, China. He has authored or co-authored more than 20 peer-referenced papers in relevant journals and conferences, including ACM Transactions on Multimedia Computing, Communications, and Applications and Nutrients. His research interests include computer vision, deep learning and food computing.

Tao Yao received the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2017. He is currently an Associate Professor with the School of Information and Electrical Engineering, Ludong University, Yantai, China, and a Researcher with the Yantai Research Institute of New Generation Information Technology, Southwest Jiaotong University. He has authored or co-authored more than 30 papers in peer-reviewed journals and conferences, such as IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Cybernetics, and ACM Transactions on Multimedia Computing, Communications, and Applications. His research interests include multimedia retrieval, computer vision, and machine learning.



Shuqiang Jiang (Senior Member, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He has authored or coauthored more than 150 articles. He was supported by the National Science Fund for Distinguished Young Scholars in 2021, the NSFC Excellent Young Scientists Fund in 2013, and the Young Top-Notch Talent of Ten Thousand Talent Program in 2014. His research interests include multimedia analysis and multimodal intelligence. He is a Senior Member of CCF and a member of ACM. He has served as a TPC Member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM. He received the Lu Jiaxi Young Talent Award from CAS in 2012 and the CCF Award of Science and Technology in 2012. He is the Vice Chair of the IEEE CASS Beijing Chapter and the ACM SIGMM China Chapter. He was the General Chair of ICIMCS in 2015 and the Program Chair of the 2019 ACM Multimedia Asia and PCM in 2017. He is an Associate Editor of Multimedia Tools and Applications and ACM Transactions on Multimedia Computing, Communications, and Applications.