RFHNet: Relational and Frequency-Aware Hashing Network for Large-Scale Fine-Grained Food Image Retrieval

Junsong Wang
School of Computer Science and Artificial Intelligence, Ludong University
Yantai, China
wangjunsong@m.ldu.edu.cn

Weiqing Min
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
minweiqing@ict.ac.cn

Guorui Sheng
School of Computer Science and Artificial Intelligence, Ludong University
Yantai, China
shengguorui@ldu.edu.cn

Tao Yao
School of Computer Science and Artificial Intelligence, Ludong University
Yantai, China
yaotao@ldu.edu.cn

Lili Wang*
School of Computer Science and Artificial Intelligence, Ludong University
Yantai, China
wanglili@ldu.edu.cn

Shuqiang Jiang
University of Chinese Academy of Sciences
Beijing, China
sqjiang@ict.ac.cn

Abstract

Fine-grained food image retrieval is a key task in computational gastronomy, with applications in food traceability, dietary monitoring, and smart catering systems. Although hashing-based retrieval is attractive for large-scale search due to its storage efficiency and fast Hamming-distance computation, existing methods often perform poorly in fine-grained food scenarios, where subtle local semantics and frequency-sensitive visual cues are essential. To address this challenge, we propose RFHNet, a cascaded hierarchical hashing network that captures both global structure and fine-grained local details through multi-level representations. RFHNet includes three components: (1) Fine-grained Relation Modeling (FRM) to capture subtle visual differences among similar food components; (2) Multi-Frequency Modulated Fusion (MFMF) to extract informative multi-frequency features; and (3) Hierarchical Semantic Synergy (HSS) to adaptively integrate multi-level representations and generate discriminative hash codes. Experiments on six food-specific benchmarks show that RFHNet consistently outperforms state-of-the-art hashing methods, with mAP gains of 4.44% to 17.20% at 12 bits. These results validate the effectiveness of RFHNet for large-scale visual food retrieval and smart catering applications. The source code will be released upon publication.

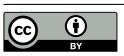
CCS Concepts

• **Information systems** → **Image search**; • **Computing methodologies** → **Visual content-based indexing and retrieval**.

Keywords

Fine-grained image retrieval, hash learning, multi-frequency information, food computing.

*Corresponding author



ICMR '26, Amsterdam, Netherlands

ACM ISBN 979-8-4007-2617-0/26/06
https://doi.org/10.1145/3805622.3810838

ACM Reference Format:
Junsong Wang, Weiqing Min, Guorui Sheng, Tao Yao, Lili Wang, and Shuqiang Jiang. 2026. RFHNet: Relational and Frequency-Aware Hashing Network for Large-Scale Fine-Grained Food Image Retrieval. In *International Conference on Multimedia Retrieval (ICMR '26), June 16–19, 2026, Amsterdam, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3805622.3810838

1 Introduction

Food image retrieval is a fundamental task in food computing, supporting applications such as intelligent menu recommendation, dietary monitoring, and nutrition management [16, 32]. Compared with generic image retrieval, food-related scenarios pose unique challenges due to fine-grained visual variations caused by ingredient composition, cooking styles, and presentation conditions. These challenges are further amplified in large-scale settings, where efficient and discriminative retrieval becomes critical.

With the increasing scale of food image datasets, Fine-Grained Image Retrieval (FGIR) [48] has attracted growing attention. Unlike coarse-grained retrieval [35], FGIR aims to distinguish visually similar instances within the same subcategory, where large intra-class variance and small inter-class variance coexist [46]. This problem is particularly severe for food images [4, 7, 19, 24, 33, 34], as visually similar dishes (e.g., beef versus bacon cheeseburgers) and appearance changes caused by preparation styles or viewpoints (Fig. 1) make discrimination highly challenging. To meet efficiency requirements in large-scale FGIR, hashing-based methods [30] have been widely adopted. Recent fine-grained hashing approaches [11, 23, 26, 47] improve retrieval performance by enhancing feature extraction and multi-level fusion. Representative methods, such as A^2-Net [45], A^2-Net^{++} [44], AGMH [29], and DAHNet [23], exploit attribute modeling, reconstruction, or multi-branch architectures to generate more discriminative hash codes.

Despite these advances, existing methods still face notable limitations in fine-grained food image retrieval. First, many approaches inadequately model subtle spatial relationships between local regions, limiting their ability to distinguish visually similar food categories. Second, most prior methods rely predominantly on spatial-domain

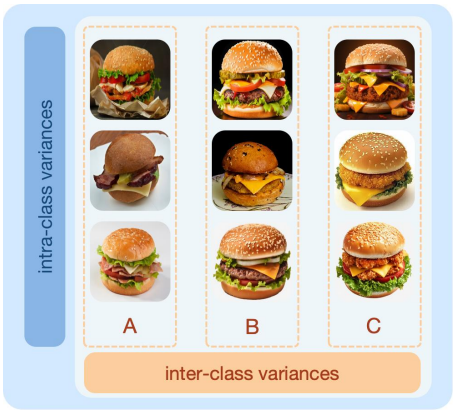


Figure 1: Illustration of inter-class and intra-class variations in fine-grained food image retrieval. Rows show inter-class differences among visually similar categories (A: Bacon Cheeseburger, B: Beef Cheeseburger, C: Chicken Cheeseburger), while columns depict intra-class variations within the same category.

feature modeling and are sensitive to both high-frequency noise and low-frequency artifacts, hindering the extraction of stable representations. Furthermore, existing deep hashing frameworks often resort to direct feature concatenation across different regions or layers, implicitly assuming statistical independence and equal contribution of all features. This rigid strategy overlooks the intrinsic semantic correlations and the imbalance of discriminative power among scales, inevitably leading to sub-optimal and less expressive hash codes.

To address these challenges, we propose RFHNet, a hierarchical cascaded hashing network for large-scale fine-grained food image retrieval. RFHNet progressively extracts and fuses multi-level features to jointly model local fine-grained details and global structural information, balancing discriminative power and semantic consistency. Specifically, the proposed framework consists of three key components: (1) Fine-Grained Relational Modeling (FRM) for capturing subtle spatial distinctions between local regions; (2) Multi-Scale Frequency Modulated Fusion (MFMF) for decomposing features into multiple frequency components and suppressing irrelevant noise; and (3) Hierarchical Semantic Synergy (HSS) for adaptively integrating hierarchical features across network stages.

Extensive experiments on six food-specific benchmarks (*Food-101* [4], *Vireo Food-172* [7], *UEC Food-256* [24], *VegFru* [19], *ISIA Food-500* [33], and *Food2K* [34]) demonstrate the effectiveness and robustness of RFHNet.

The main contributions of this work are summarized as follows:

- We propose a hierarchical cascaded hashing framework that enhances fine-grained food image representation by progressively fusing multi-level features, balancing local discriminability and global consistency.
- We introduce FRM and MFMF to capture fine-grained spatial and frequency-based distinctions, and devise HSS to adaptively bridge the semantic gap between hierarchical layers for robust cross-scale fusion.
- Extensive experiments on six food-specific benchmarks demonstrate that RFHNet significantly outperforms state-of-the-art hashing methods, achieving substantial mAP improvements and showing strong robustness across diverse food scenarios.

2 Related Work

2.1 Fine-Grained Image Retrieval (FGIR)

FGIR aims to distinguish visually similar instances within the same semantic category, and can be broadly divided into fine-grained sketch-based image retrieval (FG-SBIR) [2] and fine-grained content-based image retrieval (FG-CBIR) [15] according to the query modality. FG-SBIR focuses on cross-modal alignment between sketches and images, with an emphasis on local structure and texture modeling. Early methods mainly relied on Siamese or Triplet networks [46, 49], whereas recent approaches introduce attention mechanisms and structure-aware representations to improve fine-grained discrimination [3, 9, 41].

FG-CBIR retrieves visually similar images with subtle appearance differences and has been widely studied in product search, species recognition, and food analysis. Traditional methods based on handcrafted features and encoding schemes [13, 20, 28, 39] are limited in handling complex intra-class variations. Recent deep methods employ multi-branch architectures and attention mechanisms to mine discriminative regions, such as RA-CNN [17] and MA-CNN [50], but often incur high computational and storage costs on large-scale datasets.

To improve retrieval efficiency, fine-grained hashing learns compact binary codes while preserving subtle visual distinctions. Representative methods, including ExchNet [12], FISH [10], A^2-Net^{++} [44], and DAHNet [23], enhance fine-grained hashing through part alignment, filtering, reconstruction, and multi-branch modeling. However, most existing methods generate hash codes by directly concatenating features, overlooking semantic dependencies and the imbalance of feature importance. In contrast, our method adopts attention-guided fusion to model inter-feature relationships before hashing.

2.2 Hash Learning

Hash learning accelerates approximate nearest neighbor search by mapping high-dimensional data into compact binary codes. Existing methods are generally divided into data-independent approaches, such as locality-sensitive hashing (LSH) [14], and data-dependent learning-based methods [42]. LSH and its variants [1, 31] rely on random projections and geometric properties, but often show limited discriminability.

Learning-to-hash methods improve retrieval by optimizing hash functions with supervised or unsupervised data. Early studies mainly focused on distance preservation or label consistency [43], whereas deep hashing methods learn feature representations and hash codes end-to-end, such as DPSH [25], HashNet [6], and DSH [27]. More recent methods introduce attention mechanisms, multi-branch architectures, and transformer-based designs to strengthen semantic representation, including cross-modal hashing [21] and TransHash [8]. Fine-grained hashing is more challenging due to subtle intra-class variations, requiring more expressive feature modeling and fusion strategies, as pursued in this work.

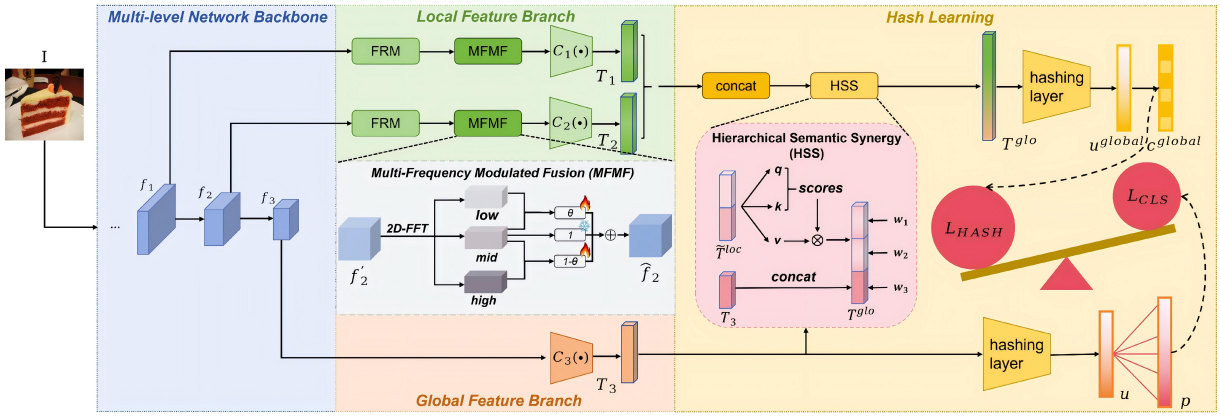


Figure 2: Overall framework of the proposed Relational and Frequency-Aware Hashing Network (RFHNet): Multi-level Backbone, Global Feature Branch, Local Feature Branch, and Hashing Learning.

2.3 Food Image Retrieval

Food image retrieval is a fundamental task in food computing, supporting efficient visual search for dietary monitoring and nutrition management. Although existing surveys [32] have extensively reviewed food-related vision tasks, most prior studies focus on recipe retrieval [7, 36, 51]. By comparison, food image retrieval remains relatively underexplored despite its practical value. Early methods mainly relied on metric learning with spatial pyramid features [38], whereas more recent approaches combine transformers and convolutional networks to improve robustness and discriminability [40]. Building on these efforts, we propose a new framework for fine-grained food image retrieval that explicitly models spatial relationships and exploits multi-frequency visual features to enhance discriminability.

3 Method

3.1 Overall Architecture

We propose RFHNet, a hierarchical cascaded hashing framework for fine-grained food image retrieval, consisting of a multi-stage backbone, a global feature branch, a local feature branch, and a hash learning module (Fig. 2). A hierarchical feature-extraction backbone based on ResNet-50 is adopted to capture discriminative representations at multiple scales.

Formally, let the backbone be denoted as $\Phi(\cdot;\Theta)$, where Θ represents the learnable parameters. Given an input image I, multi-level feature maps $f = \{f_1, f_2, \ldots, f_J\}$ are extracted from progressively deeper layers:

$$f_j = \begin{cases} \Phi_1(I, \Theta_1), & j = 1, \\ \Phi_j(f_{j-1}, \Theta_j), & j = 2, \ldots, J. \end{cases} \tag{1}$$

The deepest feature f_J is directed to the global branch, while the intermediate features $\{f_1, \ldots, f_{J-1}\}$ are processed by the local branch. Within the local branch, for each level $j \in \{1, \ldots, J-1\}$, the FRM and MFMF modules jointly model spatial relationships and multi-frequency characteristics, yielding relational features f'_j and frequency-enhanced features $\hat{f}_j$. Finally, convolutional blocks $C_j(\cdot)$ transform these enhanced local features, together with the global feature f_J, to generate the final representations $T = \{T_1, \ldots, T_J\}$.

The transformed features are then fused by HSS, which adaptively models cross-scale semantic correlations and aggregates them into a unified representation T^{glo}. The entire network is trained end-to-end with shared parameters to control model complexity. Finally, hash codes are generated via a Tanh activation followed by a sign function, encoding both global structure and fine-grained local details.

3.2 Fine-Grained Relational Modeling (FRM)

To capture subtle spatial differences between local regions, we introduce Fine-Grained Relational Modeling (FRM) in the local feature branch (Fig. 3). FRM consists of a spatial relation branch and a feature mapping branch, jointly modeling inter-region relations and preserving structural information.

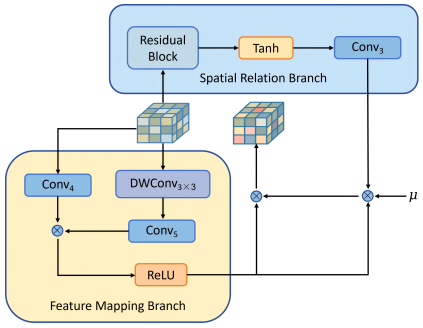


Figure 3: Illustration of Fine-Grained Relational Modeling. The residual block contains two 1×1 convolutional layers: Conv_1 and Conv_2.

Given multi-level features f_j, $j \in \{1, \ldots, J-1\}$, the spatial relation branch first reduces the channel dimension by a ratio $r = 8$ for efficiency. It then extracts relational features via a difference-based mechanism:

$$f_j^{\text{diff}} = \text{Conv}_3\big(\text{Tanh}(\text{Conv}_1(f_j) - \text{Conv}_2(f_j))\big), \tag{2}$$

where Conv_1 and Conv_2 perform the channel reduction, and Conv_3 restores the dimensions. Physically, this subtraction acts as a differential operator, highlighting the contrastive information between

different feature projections to capture subtle fine-grained details. Unlike global self-attention mechanisms that incur quadratic computational complexity, this operation efficiently models local relations with minimal overhead.

In parallel, the feature mapping branch preserves local structure using point-wise and depth-wise convolutions:

$$y_1 = \text{Conv}_4(f_j), \quad y_2 = \text{Conv}_5(\text{DWConv}_{3\times3}(f_j)), \tag{3}$$

where $\text{DWConv}_{3\times3}$ employs padding to maintain spatial alignment. This is followed by non-linear fusion: $y_3 = \text{ReLU}(y_1 + y_2)$. Finally, the outputs are fused with a weighted residual formulation:

$$f_j' = \mu(f_j^{\text{diff}} + y_3) + y_3, \tag{4}$$

where μ is a learnable fusion coefficient initialized to 1.0 to adaptively integrate the relational cues.

3.3 Multi-Frequency Modulated Fusion (MFMF)

To address the limitations of spatial-domain modeling and filter irrelevant noise, we propose the Multi-Frequency Modulated Fusion (MFMF) (Fig. 4). Leveraging 2D-FFT and a dynamic gating mechanism, the MFMF adaptively fuses frequency components to generate robust food feature representations.

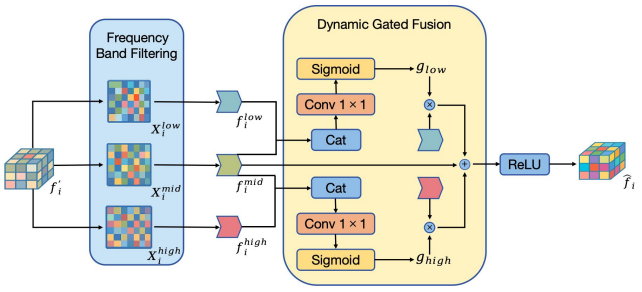


Figure 4: Overview of Multi-Frequency Modulated Fusion.

Given an input feature map $f_j' \in \mathbb{R}^{B\times C\times S\times S}$, we apply 2D-FFT to obtain its frequency representation $X_j^f = \text{FFT2D}(f_j')$. To precisely partition the spectrum, we define $d(u,v)$ as the Euclidean distance from the spectrum center to position (u,v), and R as the maximum radius of the spectrum. By setting a frequency threshold ratio τ (empirically set to 0.3), the binary masks M^* are generated as follows:

$$M^*_{(u,v)} = \begin{cases} 1, & \text{if } * = \text{low and } d(u,v) \le \tau R, \\ 1, & \text{if } * = \text{mid and } \tau R < d(u,v) < (1-\tau)R, \\ 1, & \text{if } * = \text{high and } d(u,v) \ge (1-\tau)R, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Then, the frequency bands are obtained by $X_j^* = X_j^f \odot M^*$, where $* \in \{\text{low, mid, high}\}$. These bands capture global shapes, textures, and fine details (or noise), respectively. Each band is transformed back to the spatial domain:

$$f_j^* = \text{Real}(\text{IFFT2D}(X_j^*)) \in \mathbb{R}^{B\times C\times S\times S}. \tag{6}$$

To adaptively fuse these components, we employ a gating mechanism using mid-frequency features as the contextual anchor to guide the enhancement:

$$\begin{aligned} g_{\text{low}} &= \sigma(\text{Conv}_{1\times1}([f_j^{\text{mid}}; f_j^{\text{low}}])), \\ g_{\text{high}} &= \sigma(\text{Conv}_{1\times1}([f_j^{\text{mid}}; f_j^{\text{high}}])), \end{aligned} \tag{7}$$

where $\sigma(\cdot)$ is the Sigmoid function. Finally, the fused feature is obtained by enhancing low- and high-frequency bands while preserving the structural consistency from the mid-frequency band:

$$\widehat{f}_j = g_{\text{low}} \odot f_j^{\text{low}} + f_j^{\text{mid}} + g_{\text{high}} \odot f_j^{\text{high}}. \tag{8}$$

A ReLU activation is subsequently applied.

3.4 Hierarchical Semantic Synergy (HSS)

To synergize local details with global context while preventing semantic dilution, we introduce the Hierarchical Semantic Synergy (HSS). By decoupling local feature interaction from global integration, HSS explicitly models intrinsic cross-scale correlations (or fine-grained layer-wise dependencies) before unifying them with the holistic representation.

Given the feature set $T = \{T_1, \ldots, T_J\}$, where T_J denotes the global feature and the subset $\{T_1, \ldots, T_{J-1}\}$ represents local features, we first isolate the local branch to model fine-grained correlations. The local features are stacked to form a sequence:

$$T^{\text{loc}} = \text{Stack}(T_1, \ldots, T_{J-1}) \in \mathbb{R}^{B\times(J-1)\times D}. \tag{9}$$

To capture region-to-region dependencies, self-attention is performed exclusively on the local sequence T^{loc}. The query, key, and value projections are computed as:

$$Q = T^{\text{loc}}W_Q, \quad K = T^{\text{loc}}W_K, \quad V = T^{\text{loc}}W_V, \tag{10}$$

followed by the scaled dot-product attention to generate the enhanced local features $\widetilde{T}^{\text{loc}}$:

$$\widetilde{T}^{\text{loc}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{D}}\right)V \in \mathbb{R}^{B\times(J-1)\times D}. \tag{11}$$

Subsequently, to reintegrate holistic semantic information, the enhanced local features are concatenated with the preserved global feature T_J, yielding the complete feature set:

$$\widetilde{T} = \text{Concat}(\widetilde{T}^{\text{loc}}, T_J) \in \mathbb{R}^{B\times J\times D}. \tag{12}$$

Finally, a learnable weight vector $\mathbf{W} \in \mathbb{R}^J$ is applied to adaptively balance the contribution of the refined local details and the global context. The fused representation is obtained by weighting and flattening the features:

$$T^{\text{glo}} = \text{Reshape}\left(\widetilde{T} \odot \text{Softmax}(\mathbf{W})^\top\right) \in \mathbb{R}^{B\times JD}, \tag{13}$$

where $\odot$ denotes channel-wise broadcasting multiplication.

3.5 Loss Function

As shown in Fig. 2, RFHNet is trained using a multi-task objective that jointly optimizes classification and hashing, enabling collaborative learning of semantic discrimination and compact hash codes.

Table 1: Comparison of mAP (%) with state-of-the-art methods on six fine-grained food datasets.

Datasets	# bits	HashNet	ADSH	A^2-Net	SEMICON	A^2-Net^{++}	AMGH	DAHNet	SPBH	FoodHash	**Ours**
Food-101	12	24.42	35.64	46.44	50.00	54.51	62.59	67.67	80.70	77.26	**85.14**
	24	34.48	40.93	66.87	76.57	81.46	80.94	79.38	86.07	83.01	**86.17**
	32	35.90	42.89	74.27	80.19	82.91	82.31	82.05	**86.57**	82.62	86.44
	48	39.65	48.81	82.13	82.44	83.66	83.21	83.11	**87.53**	83.44	86.42
Vireo Food-172	12	1.65	20.80	37.95	32.03	38.25	36.66	36.73	68.13	75.55	**83.14**
	24	2.34	38.86	64.45	59.89	68.49	68.70	60.43	78.79	84.73	**86.44**
	32	2.96	43.99	71.40	65.25	75.73	73.95	68.24	79.83	84.96	**86.78**
	48	4.22	61.15	76.68	69.87	79.42	76.56	75.39	81.22	85.11	**87.35**
UEC Food-256	12	3.23	11.58	24.65	14.40	20.00	24.97	41.75	59.09	48.52	**65.45**
	24	10.35	19.32	47.83	35.66	41.43	50.21	63.89	**72.05**	63.88	69.23
	32	15.64	21.88	58.52	39.67	47.01	59.97	68.05	**71.93**	64.57	70.22
	48	16.33	30.48	67.60	55.99	58.40	69.57	**72.85**	72.64	66.12	71.42
VegFru	12	3.70	8.24	25.52	30.32	30.54	43.99	56.11	63.14	65.83	**83.03**
	24	6.24	24.90	44.73	58.45	60.56	68.05	78.07	80.60	80.33	**86.28**
	32	7.83	36.53	52.75	69.92	73.38	76.73	82.19	83.02	81.83	**86.60**
	48	10.29	55.15	69.77	79.77	82.80	84.49	85.56	84.32	82.76	**87.14**
ISIA Food-500	12	0.77	4.93	10.09	5.36	10.29	11.29	5.93	15.40	32.03	**48.01**
	24	4.11	10.40	17.37	7.93	17.78	20.26	17.43	27.18	47.28	**53.85**
	32	6.26	12.64	21.27	8.94	21.53	23.62	31.23	28.42	48.91	**54.58**
	48	8.08	15.87	26.61	22.16	26.41	29.02	38.12	29.47	50.09	**55.54**
Food2K	12	0.18	0.21	3.94	1.25	3.94	5.42	0.14	1.72	26.73	**43.66**
	24	0.83	1.03	9.05	1.96	9.22	8.64	0.19	7.70	**67.01**	63.45
	32	3.17	2.50	12.00	2.76	12.22	9.22	0.36	11.21	**67.03**	65.15
	48	4.42	4.85	15.95	3.50	15.66	12.64	1.94	15.32	**71.15**	66.47

3.5.1 Classification Loss. We adopt the cross-entropy loss with label smoothing to improve generalization. Let $\mathbf{y}_i \in \{0,1\}^K$ denote the one-hot ground-truth vector for the i-th image, where K represents the number of classes. The smoothed label distribution $\tilde{y}_{i,k}$ for the k-th class is defined as:

$$\tilde{y}_{i,k} = (1-\lambda)y_{i,k} + \frac{\lambda}{K}, \tag{14}$$

where λ is the smoothing parameter (set to 0.1) and $y_{i,k}$ is the k-th element of $\mathbf{y}_i$ (1 for the target class, 0 otherwise). Consequently, the classification loss is formulated as:

$$L_{\text{CLS}} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \tilde{y}_{i,k}\log(p_{i,k}), \tag{15}$$

where $p_{i,k}$ is the predicted probability.

3.5.2 Hash Loss. Following FISH [10], we adopt a proxy-based hashing loss for stable optimization in fine-grained hashing. Let $c_i = H'(T_i^{\text{glo}})$ denote the hash code of image i, and d_u be the proxy of class u. The relaxed proxy objective is given by:

$$\begin{aligned} L'_{\text{hash}} &= \|Y - DC\|_F^2 + \sum_{i=1}^{n}\|c_i - W^h T_i^{\text{glo}}\|_F^2, \\ &\text{s.t.} \quad C \in \{-1,1\}^{k\times n}. \end{aligned} \tag{16}$$

where D denotes the class prototypes, C is the binary hash code matrix, W^h is the projection weight of the hashing layer, and c_i is the target code. The optimization is decoupled into two steps:

$$L_{\text{HASH}} = \sum_{i=1}^{n}\|c_i - W^h T_i^{\text{glo}}\|_F^2. \tag{17}$$

The offline hash code learning is formulated as:

$$\min_{C,D}\|Y - DC\|_F^2, \quad \text{s.t.} \quad C \in \{-1,1\}^{k\times n}, \tag{18}$$

which is solved via relaxation and orthogonal rotation following FISH.

3.5.3 Total Loss and Hash Code Generation. To balance classification and hashing tasks automatically without manual tuning, we employ a multi-task loss with learnable uncertainties:

$$\begin{aligned} L_{\text{TOTAL}} =& \frac{1}{\alpha^2}L_{\text{HASH}} + \frac{1}{\beta^2}L_{\text{CLS}} \\ &+ \log(\alpha+1) + \log(\beta+1), \end{aligned} \tag{19}$$

where α and β are learnable parameters optimized jointly with the network.

After training, the final binary hash code for a query image is obtained as:

$$C_q = \text{sign}(W^h T_q^{\text{glo}}), \tag{20}$$

which compactly encodes semantic and fine-grained information for efficient retrieval.

Table 2: Food retrieval accuracy (% mAP) with incremental components of the proposed RFHNet model.

Datasets	# bits	Vanilla Backbone (ResNet-50)	+FRM (Sec. 3.2)	+MFMF (Sec. 3.3)	+HSS (Sec. 3.4)
Food-101	12	83.30	84.12	85.03	85.14
	24	85.11	85.31	85.89	86.17
	32	86.15	86.21	86.38	86.44
	48	85.54	85.77	86.30	86.42
Vireo Food-172	12	80.48	81.22	82.73	83.14
	24	84.66	85.83	86.28	86.44
	32	84.59	85.23	86.45	86.78
	48	85.39	86.25	87.23	87.35
UEC Food-256	12	62.77	63.17	64.53	65.45
	24	67.55	68.13	68.62	69.23
	32	67.84	68.03	69.83	70.22
	48	68.63	70.02	70.85	71.42
VegFru	12	81.33	82.04	82.86	83.03
	24	84.61	85.11	86.03	86.28
	32	85.01	85.67	86.25	86.60
	48	86.69	86.60	86.97	87.14
ISIA Food-500	12	42.70	44.05	46.72	48.01
	24	51.17	51.99	53.32	53.85
	32	52.31	53.08	53.94	54.58
	48	52.56	53.93	55.49	55.54
Food2K	12	38.58	40.36	43.17	43.66
	24	61.33	62.38	63.29	63.45
	32	63.77	63.91	64.82	65.15
	48	64.02	64.98	66.27	66.47

4 Experiments

4.1 Datasets

We evaluate RFHNet on six large-scale benchmarks, strictly adhering to official protocols: *Food-101* [4] (101k images, 101 classes, standard 750/250 split); *Vireo Food-172* [7] (110k images, 172 classes, random 80%/20% split); *UEC Food-256* [24] (~32k images, 256 classes, using valid bounding boxes); *VegFru* [19] (>160k images, 292 categories, official train/val/test split); *ISIA Food-500* [33] (399k images, 500 classes, fixed 800/200 split); and *Food2K* [34] (>1M images, 2,000 categories, standard split).

4.2 Evaluation Metric

We evaluate retrieval performance using mean Average Precision (mAP). Specifically, we report mAP@1000, which computes the Average Precision (AP) over the top 1,000 retrieved results for each query and averages the AP scores across all queries. This metric is widely used to assess effectiveness in large-scale retrieval scenarios.

4.3 Comparison Methods and Experimental Settings

4.3.1 Comparison Methods. We compare RFHNet with nine representative hashing methods, including HashNet [18] and ADSH [22] for coarse-grained retrieval, and seven state-of-the-art fine-grained hashing methods: A^2-Net [45], SEMICON [37], A^2-Net^{++} [44], AMGH [29], DAHNet [23], SPBH [26], and FoodHash [5].

4.3.2 Experimental Settings. For fair comparison, all CNN-based baselines adopt ResNet-50 or ResNet-18 backbones pre-trained on ImageNet. For the Transformer-based method FoodHash, we employ ViT-Small to ensure comparable model complexity. All backbones have their original pooling and classification layers removed, and only image-level category labels are used for supervision. During training, images are randomly cropped and resized to 224×224 with horizontal flipping; during testing, images are resized to 256×256 and center-cropped to 224×224. Models are trained using SGD with momentum 0.9 and weight decay 3×10^{-4}. Unless otherwise specified, the learning rate is 0.008, batch size is 128, and training lasts 300 epochs. For *ISIA Food-500*, training is reduced to 200 epochs; for *VegFru*, the learning rate is 0.005; and for *Food2K*, training is conducted for 150 epochs.

4.4 Performance Comparison

Table 1 compares RFHNet with state-of-the-art methods. RFHNet establishes a new benchmark for short-code hashing, achieving the best performance on *all six* datasets at 12 bits. Compared to the strongest competitors (SPBH and FoodHash), RFHNet yields substantial improvements in the 12-bit setting: 17.20% on *VegFru*, 16.93% on *Food2K*, and 15.98% on *ISIA Food-500*. Significant leads are

Table 3: Comparison of mAP (%) between proposed RFHNet (based on Resnet18 and Resnet50) and DAHNet (based on Resnet50).

Datasets	# bits	DAHNet	$\text{RFHNet}_{\text{ResNet18}}$	$\text{RFHNet}_{\text{ResNet50}}$
Food-101	12	67.67	80.02	85.14
	24	79.38	82.72	86.17
	32	82.05	82.76	86.44
	48	83.11	83.22	86.42
Vireo Food-172	12	36.73	78.25	83.14
	24	60.43	84.50	86.44
	32	68.24	84.62	86.78
	48	75.39	85.19	87.35
UEC Food-256	12	41.75	56.79	65.45
	24	63.89	64.19	69.23
	32	68.05	65.86	70.22
	48	72.85	67.67	71.42
VegFru	12	56.11	76.97	83.03
	24	78.07	81.58	86.28
	32	82.19	81.97	86.60
	48	85.56	82.98	87.14
ISIA Food-500	12	5.93	39.18	48.01
	24	17.43	46.59	53.85
	32	31.23	48.13	54.58
	48	38.12	49.68	55.54
Food2K	12	0.14	20.13	43.66
	24	0.19	28.77	63.45
	32	0.36	31.49	65.15
	48	1.94	33.08	66.47

Table 4: mAP (%) on the testing set with fixed hyperparameters of loss functions on *UEC Food-256*.

Method	Parameter		*UEC Food-256*			
	α	β	12bits	24bits	32bits	48bits
Fixed	0.8	1	62.08	66.46	67.24	69.42
	1	0.8	62.67	67.57	67.09	69.92
	1	1	62.94	67.40	68.67	70.14
	0.5	1	63.00	68.54	69.27	70.89
	1	0.5	62.37	68.67	68.93	71.11
Dynamic	-	-	**65.45**	**69.23**	**70.22**	**71.42**

Table 5: Inference efficiency comparison between RFHNet and FoodHash on four large-scale fine-grained food datasets. GPU Memory (MB) and Retrieval Time (ms/img) denote peak memory usage and average latency per image, respectively. ↓ indicates lower is better.

Dataset	GPU Memory (MB) ↓		Retrieval Time (ms/img) ↓	
	FoodHash	**RFHNet (Ours)**	FoodHash	**RFHNet (Ours)**
Food-101	11550.5	**1651.5**	0.726	**0.282**
Vireo Food-172	11183.3	**1659.9**	0.731	**0.283**
UEC Food-256	11536.6	**1648.0**	0.732	**0.286**
Food2K	25527.5	**11855.5**	0.748	**0.305**

also achieved on *Vireo Food-172* (+7.59%), *UEC Food-256* (+6.36%), and *Food-101* (+4.44%). Moreover, on datasets like *Vireo Food-172*, *VegFru*, and *ISIA Food-500*, our method consistently outperforms benchmarks across all bit lengths (12-48 bits), validating its effectiveness in capturing discriminative fine-grained features.

4.5 Ablation Study

4.5.1 Ablation on Model Components. We progressively integrate FRM, MFMF, and HSS into a vanilla cascaded ResNet-50 backbone and evaluate their effects on six food datasets. As shown in Table 2, retrieval performance consistently improves with each module added, demonstrating the effectiveness and complementarity of all components.

Figure 5: Examples of the top-10 retrieved images using 48-bit hash codes generated by RFHNet on two representative fine-grained food image datasets (*Food-101* and *Vireo Food-172*).

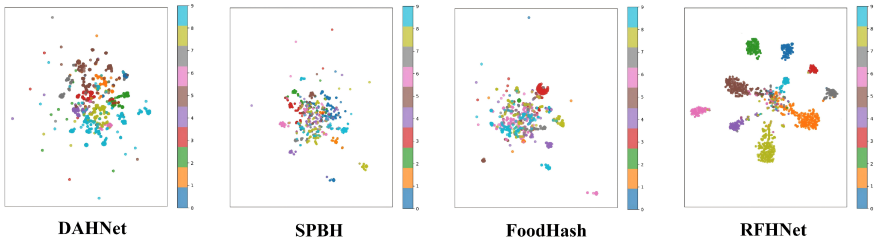


Figure 6: t-SNE visualization of 12-bit hash codes learned by DAHNet, SPBH, FoodHash, and the proposed RFHNet on the *VegFru* dataset.

4.5.2 Ablation on Backbone. To evaluate performance under lightweight settings, we replace ResNet-50 with ResNet-18. As reported in Table 3, RFHNet with ResNet-18 consistently outperforms DAHNet (ResNet-50) across all datasets, while stronger backbones further improve performance, indicating good scalability.

4.5.3 Ablation on Loss Function. We evaluate the impact of learnable loss weights on *UEC Food-256*. As shown in Table 4, models with learnable parameters consistently outperform those with fixed settings, confirming the benefit of adaptive loss balancing.

4.6 Qualitative Results

Fig. 5 presents representative retrieval results, demonstrating robust performance under diverse backgrounds and fine-grained variations. Fig. 6 visualizes the t-SNE embeddings of 12-bit hash codes on *VegFru*. Compared with DAHNet, SPBH, and FoodHash, the proposed RFHNet exhibits tighter intra-class clustering and larger inter-class margins. This qualitative result confirms that our method effectively captures subtle visual semantics, generating more discriminative representations for fine-grained retrieval.

4.7 Inference Efficiency Analysis

Inference efficiency is crucial for large-scale deployment. As shown in Table 5, RFHNet requires fewer resources than the Transformer-based FoodHash while preserving real-time retrieval performance. This benefit mainly stems from the Hierarchical Semantic Synergy (HSS) module, which avoids the quadratic complexity of Vision Transformers through lightweight feature interaction.

5 Conclusion

We proposed RFHNet, a hierarchical cascaded hashing network for large-scale fine-grained food image retrieval. By jointly modeling fine-grained spatial relations and multi-frequency characteristics, RFHNet learns compact and discriminative hash codes that preserve both local details and global structure. Extensive experiments on six fine-grained food benchmarks demonstrate consistent performance gains, particularly excelling in 12-bit short-code retrieval scenarios.

References

[1] Alexandr Andoni and Ilya Razenshteyn. 2015. Optimal data-dependent hashing for approximate near neighbors. In *Proc. Annu. ACM Symp. Theory Comput.* ACM,

New York, NY, USA, 793–801.
[2] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. 2021. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 4247–4256.
[3] Ayan Kumar Bhunia, Aneeshan Sain, Parth Hiren Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Adaptive fine-grained sketch-based image retrieval. In *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 163–181.
[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 446–461.
[5] Pindan Cao, Weiqing Min, Guorui Sheng, Yongqiang Song, Tao Yao, Lili Wang, and Shuqiang Jiang. 2026. FoodHash: Context-Aware Proxy Interaction and Fusion for Food Image Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* (2026).
[6] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. 2017. Hashnet: Deep learning to hash by continuation. In *Proc. IEEE Int. Conf. Comput. Vision*. IEEE, New York, NY, USA, 5608–5617.
[7] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proc. ACM Int. Conf. Multimed.* ACM, New York, NY, USA, 32–41.
[8] Yongbiao Chen, Sheng Zhang, Fangxin Liu, Zhigang Chang, Mang Ye, and Zhengwei Qi. 2022. Transhash: Transformer-based hamming hashing for efficient image retrieval. In *Proc. Int. Conf. Multimed. Retr.* ACM, New York, NY, USA, 127–136.
[9] Yangdong Chen, Zhaolong Zhang, Yanfei Wang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2022. AE-Net: Fine-grained sketch-based image retrieval via attention-enhanced network. *Pattern Recognit.* 122 (2022), 108291.
[10] Zhen-Duo Chen, Xin Luo, Yongxin Wang, Shanqing Guo, and Xin-Shun Xu. 2022. Fine-grained hashing with double filtering. *IEEE Trans. Image Process.* 31 (2022), 1671–1683.
[11] Zhen-Duo Chen, Li-Jun Zhao, Zi-Chao Zhang, Xin Luo, and Xin-Shun Xu. 2024. Characteristics matching based hash codes generation for efficient fine-grained image retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, Seattle, WA, USA, 17273–17281.
[12] Quan Cui, Qing-Yuan Jiang, Xiu-Shen Wei, Wu-Jun Li, and Osamu Yoshie. 2020. ExchNet: A unified hashing network for large-scale fine-grained image retrieval. In *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 189–205.
[13] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 1. IEEE, New York, NY, USA, 886–893.
[14] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions. In *Proc. Annu. Symp. Comput. Geom.* ACM, New York, NY, USA, 253–262.
[15] Shiv Ram Dubey. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Trans. Circuits Syst. Video Technol.* 32, 5 (2021), 2687–2704.
[16] Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, and Sebastiano Battiato. 2016. Retrieval and classification of food images. *Computers in biology and medicine* 77 (2016), 23–39.
[17] Jianlong Fu, Heliang Zheng, and Tao Mei. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 4438–4446.
[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 770–778.
[19] Saihui Hou, Yushan Feng, and Zilei Wang. 2017. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 541–549.
[20] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. 2010. Aggregating local descriptors into a compact image representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 3304–3311.
[21] Qing-Yuan Jiang and Wu-Jun Li. 2017. Deep cross-modal hashing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 3232–3240.
[22] Qing-Yuan Jiang and Wu-Jun Li. 2018. Asymmetric deep supervised hashing. In *Proc. AAAI Conf. Artif. Intell.*, Vol. 32. AAAI Press, Washington, DC, USA.
[23] Xin Jiang, Hao Tang, and Zechao Li. 2024. Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. *IEEE Trans. Knowl. Data Eng.* 36, 11 (2024), 6266–6279.
[24] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 3–17.
[25] Wu-Jun Li, Sheng Wang, and Wang-Cheng Kang. 2015. Feature learning based deep supervised hashing with pairwise labels. *arXiv:1511.03855* (2015).
[26] Xue Li, Jiong Yu, Junlong Cheng, Ziyang Li, and Chen Bian. 2025. Semantic Preservation-Based Hash Code Generation for fine-grained image retrieval. *Expert Sys. Appl.* 271 (2025), 126668.
[27] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2016. Deep supervised hashing for fast image retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 2064–2072.
[28] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *int. j. comput. vision* 60 (2004), 91–110.
[29] Xin Lu, Shikun Chen, Yichao Cao, Xin Zhou, and Xiaobo Lu. 2023. Attributes grouping and mining hashing for fine-grained image retrieval. In *Proc. ACM Int. Conf. Multimed.* ACM, New York, NY, USA, 6558–6566.
[30] Xiao Luo, Haixin Wang, Daqing Wu, Chong Chen, Minghua Deng, Jianqiang Huang, and Xian-Sheng Hua. 2023. A survey on deep hashing methods. *ACM Trans. Knowl. Discov. Data* 17, 1 (2023), 1–50.
[31] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. 2007. Multi-probe LSH: efficient indexing for high-dimensional similarity search. In *Int. Conf. Very Large Data Bases*. VLDB Endowment, White Plains, NY, USA, 950–961.
[32] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *ACM Comput. Surv.* 52, 5 (2019), 1–36.
[33] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2020. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proc. ACM Int. Conf. Multimed.* ACM, New York, NY, USA, 393–401.
[34] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. 2023. Large scale visual food recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 8 (2023), 9932–9949.
[35] AN Rajath, K Vidyalakshmi, and GN Keshava Murthy. 2023. A comprehensive analysis on deep learning based image retrieval. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*. IEEE, New York, NY, USA, 1–4.
[36] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. 2021. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 15475–15484.
[37] Yang Shen, Xuhao Sun, Xiu-Shen Wei, Qing-Yuan Jiang, and Jian Yang. 2022. SEMICON: A learning-to-hash solution for large-scale fine-grained image retrieval. In *Proc. Eur. Conf. Comput. Vis.* Springer, Cham, 531–548.
[38] Wataru Shimoda and Keiji Yanai. 2017. Learning food image similarity for food image retrieval. In *Proc. IEEE Int. Conf. Multimed. Big Data.* IEEE, New York, NY, USA, 165–168.
[39] Sivic and Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proc. IEEE Int. Conf. Comput. Vision.* IEEE, New York, NY, USA, 1470–1477.
[40] Jiajun Song, Weiqing Min, Yuxin Liu, Zhuo Li, Shuqiang Jiang, and Yong Rui. 2022. A noise-robust locality transformer for fine-grained food image retrieval. In *Proc. Int. Conf. Multimed. Inf. Process. Retr.* IEEE, New York, NY, USA, 348–353.
[41] Haifeng Sun, Jiaqing Xu, Jingyu Wang, Qi Qi, Ce Ge, and Jianxin Liao. 2022. DLI-Net: Dual local interaction network for fine-grained sketch-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* 32, 10 (2022), 7177–7189.
[42] Jun Wang, Wei Liu, Sanjiv Kumar, and Shih-Fu Chang. 2015. Learning to hash for indexing big data—A survey. *Proc. IEEE* 104, 1 (2015), 34–57.
[43] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. 2017. A survey on learning to hash. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2017), 769–790.
[44] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Peng Wang, and Yuxin Peng. 2023. Attribute-Aware Deep Hashing With Self-Consistency for Large-Scale Fine-Grained Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 11 (2023), 13904–13920. https://doi.org/10.1109/TPAMI.2023.3299563
[45] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. 2021. Aˆ2-Net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *Proc. Adv. neural inf. proces. syst.* 34 (2021), 5720–5730.
[46] Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. 2021. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 12 (2021), 8927–8948.
[47] Xinguang Xiang, Xinhao Ding, Lu Jin, Zechao Li, Jinhui Tang, and Ramesh Jain. 2024. Alleviating over-fitting in hashing-based fine-grained image retrieval: From causal feature learning to binary-injected hash learning. *IEEE Trans. Multimedia* 26 (2024), 10665–10677.
[48] Han Yu, Huibin Lu, Min Zhao, Zhuoyi Li, and Guanghua Gu. 2024. Gradient aggregation based fine-grained image retrieval: A unified viewpoint for CNN and Transformer. *Pattern Recognit.* 149 (2024), 110248.
[49] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. 2016. Sketch me that shoe. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 799–807.
[50] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 5209–5217.
[51] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, New York, NY, USA, 11477–11486.