SAHNet: A Spectral-Aware Hashing Network for Large-Scale Fine-Grained Food Image Retrieval

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

Yihan Wang
Ludong University, School of Computer Science and Artificial Intelligence
Yantai, Shandong, China
yihan.wang@m.ldu.edu.cn

Bolin Yang
Ludong University, School of Computer Science and Artificial Intelligence
Yantai, Shandong, China
yyybl@m.ldu.edu.cn

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

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

Abstract

Fine-grained food image retrieval is vital for applications such as dietary monitoring and personalized nutrition. While hashing-based methods are popular for their storage and computational efficiency, existing approaches often fail to exploit category-specific spectral-spatial features and tend to preserve redundant representations, thereby limiting their discriminative ability. To address these issues, we propose SAHNet, a hierarchical network with a multi-scale backbone and two key modules: Spectral-Spatial Information Mining (SSIM) for dual-branch spectral/spatial feature extraction, and Selective Feature Filtering (SFF) for enhancing critical cues while suppressing noise. SAHNet generates compact, discriminative hash codes and achieves state-of-the-art performance on four benchmark fine-grained food datasets.

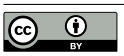
CCS Concepts

• **Information systems** → **Similarity measures**.

Keywords

fine-grained image retrieval, hash learning, spectral-spatial feature learning, food computing.

ACM Reference Format:
Junsong Wang, Yihan Wang, Bolin Yang, Guorui Sheng, and Lili Wang. 2025. SAHNet: A Spectral-Aware Hashing Network for Large-Scale Fine-Grained Food Image Retrieval. In *Proceedings of the 18th International Symposium on Visual Information Communication and Interaction (VINCI 2025), December 01–03, 2025, Linz, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3769534.3769598



VINCI 2025, Linz, Austria

ACM ISBN 979-8-4007-1845-8/25/12
https://doi.org/10.1145/3769534.3769598

1 Introduction

Food computing, situated at the intersection of artificial intelligence and nutritional science, has attracted growing attention due to its potential in dietary monitoring and personalized nutrition [11, 19, 20]. Within this field, food image retrieval is a key technique enabling efficient management of food-related visual data.

Compared to coarse-grained retrieval, fine-grained food image retrieval (FGIR) is particularly challenging because of large intra-class variations and subtle inter-class differences [26], e.g., visually similar dishes such as strawberry and raspberry shortcakes. To address scalability, hashing methods have been widely adopted for their efficiency on large datasets [21, 22].

Prior work in fine-grained retrieval can be categorized as content-based (FG-CBIR) and sketch-based (FG-SBIR). FG-SBIR retrieves real images using hand-drawn sketches, emphasizing subtle visual differences and requiring cross-modal alignment. Early works relied on Siamese or Triplet networks [12, 13, 18], while more recent methods adopt attention and structure-aware features [3]. FG-CBIR instead uses real images as queries to capture fine distinctions such as species or styles. It has been widely applied in product search and food recognition. Early CNN-based approaches utilized attention and localization to highlight discriminative regions [23], while recent methods incorporate multi-level fusion and hybrid CNN–Transformer frameworks [27].

Fine-grained hashing has emerged as an efficient solution to large-scale retrieval, aiming to learn compact binary codes that preserve subtle category differences. Earlier works localized discriminative regions for hashing [4], followed by attribute-guided and ensemble-based approaches [8, 10, 24]. In the context of food computing, most prior studies have emphasized recipe retrieval and cross-modal matching, leaving image retrieval comparatively underexplored [11]. To improve representation learning, Song et al. [16] introduced a generalization-driven sampling method, while EVHNet [14] leveraged ViT-based hashing to enhance local and semantic cues, and the NoLo-Transformer [17] combined CNNs and Transformers for robust food image retrieval.

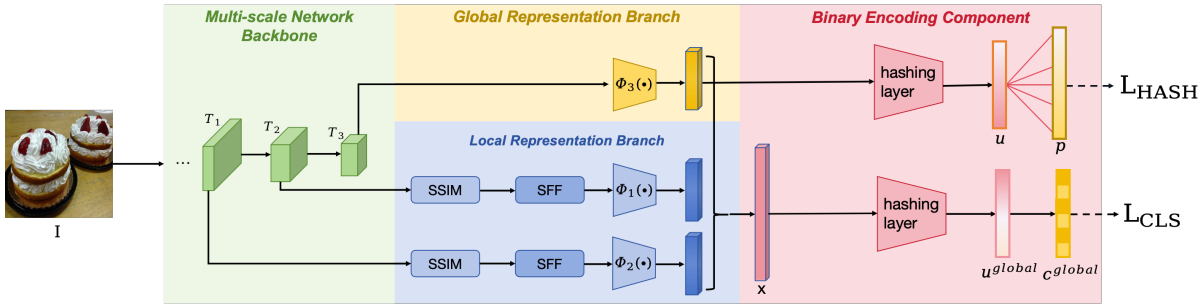


Figure 1: Overall architecture of SAHNet. The framework consists of four modules: multi-scale network backbone, global representation branch, local representation branch, and binary encoding component. Here, the parameter j is set to 3.

Despite notable progress, existing approaches still suffer from two critical limitations: (1) they often neglect the category-specific spectral–spatial characteristics that are essential for distinguishing subtle inter-class differences, and (2) they introduce redundancy through uniform channel transformations, which undermines discriminative power. In contrast to traditional methods, our work is the first to explicitly introduce spectral–spatial feature separation and selective filtering into food image retrieval. To this end, we propose SAHNet, a hierarchical framework that progressively integrates multi-scale representations while incorporating two novel modules: Spectral-Spatial Information Mining (SSIM), which disentangles and captures spectral and spatial cues, and Selective Feature Filtering (SFF), which adaptively enhances salient features while suppressing redundant or irrelevant information. This design yields compact yet highly discriminative hash codes, tailored for the fine-grained retrieval of food images.

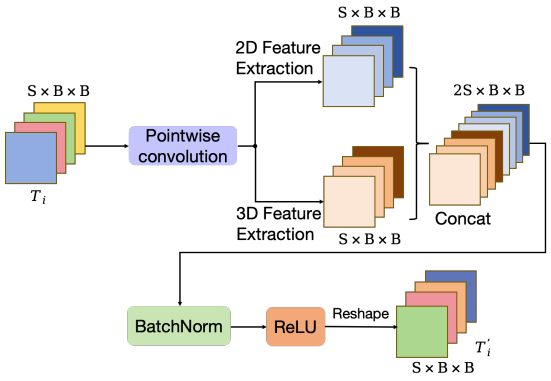


Figure 2: SSIM module: pointwise convolution reduces dimension, parallel 2D/3D branches capture spatial and spectral cues, and outputs are fused into spectral–spatial features.

2 Methodology

2.1 Overall Architecture

To exploit fine-grained food features, we propose SAHNet, which includes four components: a multi-scale backbone, global and local branches, and a binary encoding module, as shown in Figure 1.

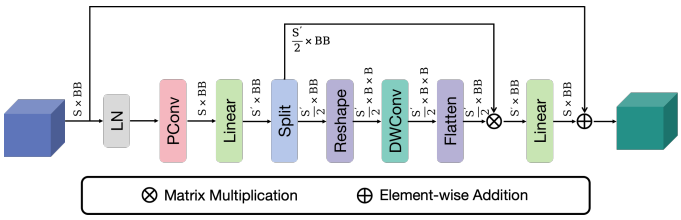


Figure 3: An Illustration of the SFF.

We adopt ResNet-50 as the backbone $N(\cdot;\Theta)$, where Θ denotes learnable parameters. Given an image I, the network outputs multi-scale feature maps $\mathbf{T} = \{T_1, \ldots, T_j\}$ from successive stages:

$$T_j = N_j(T_{j-1}, \Theta_j). \tag{1}$$

The final-layer feature goes to the global branch, while intermediate maps $\{T_1, \ldots, T_{j-1}\}$ feed the local branch. The local branch integrates two modules—Spectral-Spatial Information Mining (SSIM) and Selective Feature Filtering (SFF)—to enhance discriminative cues. Both global and local features are processed by stage-wise convolutional blocks $\Phi_i(\cdot)$, concatenated, and encoded into compact representations x. Hash codes are then generated via Tanh and sign functions. This design captures both global structures and local textures while supporting efficient end-to-end training with parameter sharing.

2.2 Spectral-Spatial Information Mining (SSIM)

In fine-grained food image retrieval, to effectively capture both spatial (texture) and spectral (color) information, we propose the SSIM module, as shown in Figure 2. Each input block of size $P \times B \times B$ is first passed through a 1×1 convolution to reduce dimensionality and computational cost, producing a feature map $f \in \mathbb{R}^{S \times B \times B}$.

SSIM then splits f into two parallel branches. The spectral branch applies a lightweight 3D convolution ($1 \times 1 \times 3$) along the channel axis, effectively learning spectral dependencies while minimizing overhead. The spatial branch employs a 3×3 2D convolution to capture local texture and spatial structures.

The outputs of both branches are concatenated and refined, generating fused spectral–spatial feature maps that enhance the discriminative power of the network. The computation in the 3D

convolution block can be formally described as follows:

$$f_{3d} = \Psi(f\Gamma\omega_{3d} + \lambda_{2d}), \tag{2}$$

where f_{3d} denotes the resulting 3D feature map, w_{3d} and λ_{3d} are the learnable weight and bias parameters, respectively. Γ represents the 3D convolution operation, and Ψ denotes the corresponding activation function.

In the spatial branch, a standard 2D convolution is applied to capture local texture information from the input feature map f. The operation can be expressed as:

$$f_{2d} = \Psi(f \odot \omega_{2d} + \lambda_{2d}), \tag{3}$$

where $\odot$ denotes 2D convolution, ω_{2d} and λ_{2d} are learnable parameters, and Ψ is the activation function.

The outputs from the spectral (3D) and spatial (2D) branches are then concatenated to exploit their complementary strengths:

$$T^{'} = \text{Concat}(f_{3d}, f_{2d}), \tag{4}$$

followed by a 1×1 convolution to reduce channels, yielding compact spectral–spatial features for subsequent processing.

2.3 Selective Feature Filtering (SFF)

Conventional Feed-Forward Networks (FFNs) process pixels independently, which may lead to redundant channel information even after spectral–spatial enhancement with SSIM. To address this, we propose the SFF module (Figure 3), which combines partial convolution (PConv) and a gating mechanism to emphasize informative cues and suppress irrelevant background. The computation is as follows:

$$\hat{T}^{'} = \text{GELU}(W_1\text{PConv}(T^{'})), [\hat{T}_1^{'}, \hat{T}_2^{'}] = \hat{T}^{'} \tag{5}$$

$$\hat{T}_r^{'} = \hat{T}_1^{'} \otimes F(\text{DWConv}(R(\hat{T}_2^{'}))) \tag{6}$$

$$\hat{T}_{out}^{'} = \text{GELU}(W_2\hat{T}_r^{'}), \tag{7}$$

where W_1 and W_2 are projection matrices, PConv and DWConv denote partial and depth-wise convolutions, and $\otimes$ represents matrix multiplication. This design selectively filters features, reducing redundancy while enhancing discriminative capacity.

Overall, SFF strengthens feature learning by selectively retaining informative cues while suppressing redundant channels. This mechanism yields more compact and discriminative representations, enhancing the network's retrieval performance.

2.4 Loss Function and Hash Code Generation

We adopt a joint objective combining classification loss and hashing loss. The classification loss employs cross-entropy with label smoothing to improve generalization:

$$L_{\text{CLS}} = -\sum_{i=1}^{n} y_i^{\text{LS}} \cdot \log \frac{\exp(\bar{y}_i)}{\sum_{k=1}^{l} \exp(\bar{y}_{ik})}, \tag{8}$$

where $y_i^{\text{LS}} = y_i(1-\mu) + \mu/l$ is the smoothed label distribution.

To enhance training stability, we use a proxy-based hashing loss. Relaxing the discrete constraint, the objective is:

$$L_{\text{HASH}} = \sum_{i=1}^{n} \left\| c_i - W^h x_i \right\|_F^2, \tag{9}$$

where c_i denotes the generated hash code and W^h is the projection matrix.

The overall training objective is:

$$L_{\text{TOTAL}} = \alpha L_{\text{HASH}} + \beta L_{\text{CLS}}, \tag{10}$$

with α and β balancing the two terms. After training, the binary hash code c_q is generated as:

$$c_q = \text{sign}(W^h x_q). \tag{11}$$

3 Experimental Analysis and Results

3.1 Dataset and Training Details

We evaluate SAHNet on four fine-grained food benchmarks: Food-101 [1] (101k images, 101 classes), Vireo Food-172 [2] (110k images, 172 classes), UEC Food-256 [9] (31k images, 256 classes), and VegFru [6] (161k images, 292 categories of vegetables and fruits). Each dataset follows the official train/test splits.

During training, images are randomly cropped to 224×224 with horizontal flipping for augmentation; for testing, they are resized to 256×256 and center-cropped to 224×224. Optimization uses SGD with momentum 0.9, weight decay 0.0003, an initial learning rate of 0.008, batch size of 128, and 200 epochs. Hyperparameters α and β in Equation 10 are set to 0.8 and 1, respectively.

3.2 Comparison with Other Models

We evaluate SAHNet against seven representative hashing methods—HashNet [5], ADSH [7], A^2-Net [25], SEMICON [15], A^2-Net^{++} [24], AMGH [10], and DAHNet [8]—on four fine-grained food benchmarks (Table 1). Among them, HashNet and ADSH are primarily designed for coarse-grained retrieval, whereas the remaining approaches specifically address fine-grained hashing. Across all datasets and code lengths (12, 24, 32, and 48 bits), SAHNet consistently achieves superior performance. For instance, on Vireo Food-172, it improves over DAHNet by 45.79%, 26.40%, 18.95%, and 12.17%, respectively. Moreover, with 12-bit codes, SAHNet achieves notable gains on Food-101 (17.25%), UEC Food-256 (23.06%), and VegFru (26.51%), underscoring its effectiveness in capturing fine-grained discriminative cues.

Figure 4 illustrates retrieval results on Food-101 and Vireo Food-172, showing that SAHNet effectively handles diverse backgrounds and achieves accurate subcategory retrieval. Further, the t-SNE visualizations in Figure 5 confirm the discriminative power of the learned 12-bit codes: while DAHNet produces overlapping clusters, SAHNet yields compact and well-separated distributions, highlighting its superior ability to capture subtle category differences and resist background interference.

3.3 Ablation Experiment

In this section, we evaluate the contributions of SSIM and SFF through ablation studies on Food-101. Starting from a cascaded ResNet-50 baseline, we progressively add each module. As shown in Table 2, both SSIM and SFF individually yield clear mAP gains across different code lengths, while their joint integration in SAHNet achieves the best performance. For instance, mAP increases from 83.65% to 84.92% at 12 bits and from 84.35% to 86.08% at 24 bits. These results not only confirm the complementary effectiveness

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

Datasets	# bits	HashNet	ADSH	A^2-Net	SEMICON	A^2-Net^{++}	AMGH	DAHNet	**Ours**
Food-101	12	24.42	35.64	46.44	50.00	54.51	62.59	67.67	**84.92**
	24	34.48	40.93	66.87	76.57	81.46	80.94	79.38	**86.08**
	32	35.90	42.89	74.27	80.19	82.91	82.31	82.05	**86.23**
	48	39.65	48.81	82.13	82.44	83.66	83.21	83.11	**86.78**
Vireo Food-172	12	1.65	20.80	37.95	32.03	38.25	36.66	36.73	**82.52**
	24	2.34	38.86	64.45	59.89	68.49	68.70	60.43	**86.83**
	32	2.96	43.99	71.40	65.25	75.73	73.95	68.24	**87.19**
	48	4.22	61.15	76.68	69.87	79.42	76.56	75.39	**87.56**
UEC Food-256	12	3.23	11.58	24.65	14.40	20.00	24.97	41.75	**64.81**
	24	10.35	19.32	47.83	35.66	41.43	50.21	63.89	**68.91**
	32	15.64	21.88	58.52	39.67	47.01	59.97	68.05	**70.62**
	48	16.33	30.48	67.60	55.99	58.40	69.57	**72.85**	71.66
VegFru	12	3.70	8.24	25.52	30.32	30.54	43.99	56.11	**82.62**
	24	6.24	24.90	44.73	58.45	60.56	68.05	78.07	**85.91**
	32	7.83	36.53	52.75	69.92	73.38	76.73	82.19	**86.35**
	48	10.29	55.15	69.77	79.77	82.80	84.49	85.56	**87.07**

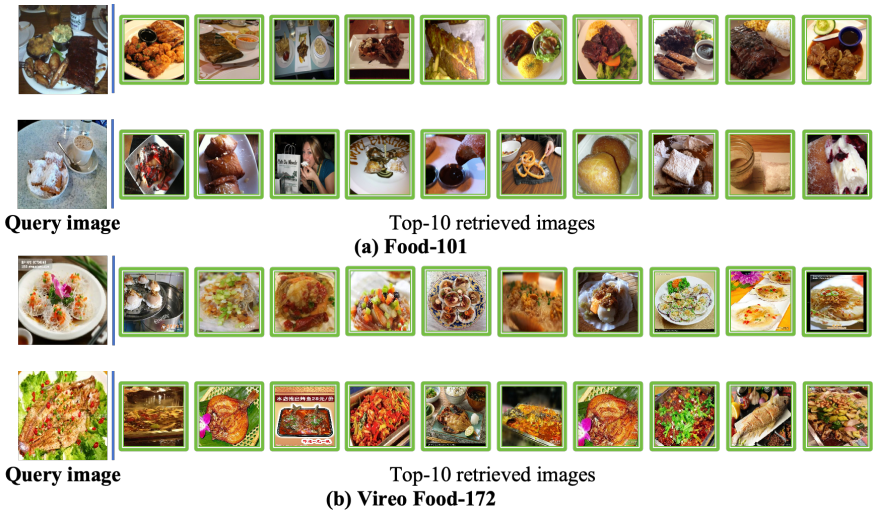


Figure 4: Illustration of the top 10 images retrieved by SAHNet using 48-bit hash codes across Food-101 and Vireo Food-172.

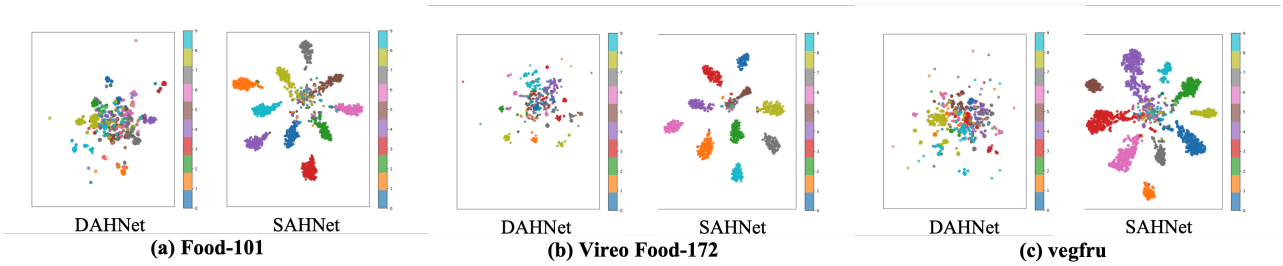


Figure 5: t-SNE visualizations of 12-bit hash codes generated by DAHNet and SAHNet on Food-101, Vireo Food-172, and VegFru. For each dataset, DAHNet results are shown on the left and SAHNet results on the right. We repeated the t-SNE experiments multiple times and obtained consistent results, confirming the stability of the visualizations.

Table 2: Food retrieval accuracy (% mAP) with incremental components of the proposed SAHNet model on the Food-101.

Methods	12-bit	24-bit	32-bit	48-bit
Base	83.65	84.35	85.89	86.12
Base+SSIM	84.33	85.90	86.20	86.54
Base+SFF	84.13	85.75	85.39	86.69
SAHNet	**84.92**	**86.08**	**86.23**	**86.78**

of SSIM and SFF but also validate the novelty of their synergistic integration, which enables SAHNet to learn more compact and discriminative hash codes for fine-grained food image retrieval.

4 Conclusion

In this work, we present SAHNet, a hierarchical hashing framework for fine-grained food image retrieval. By combining a multi-scale backbone with SSIM and SFF modules, SAHNet captures both global and local cues, enhancing discriminability while reducing redundancy. Experiments on four benchmark datasets show that SAHNet consistently surpasses state-of-the-art methods in retrieval accuracy and generalization, while ablation studies confirm the complementary benefits of SSIM and SFF. These results highlight the practical potential of SAHNet. Moreover, given its general design, the framework can be extended beyond food computing to broader fine-grained retrieval tasks across different domains. Future work will further explore multimodal retrieval, cross-domain applications, and efficient inference, while also addressing scalability and efficiency challenges in large-scale deployments.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. 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, 446–461.

[2] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*. 32–41.

[3] 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 recognition* 122 (2022), 108291.

[4] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 189–205.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Saihui Hou, Yushan Feng, and Zilei Wang. 2017. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE international conference on computer vision*. 541–549.

[7] Qing-Yuan Jiang and Wu-Jun Li. 2018. Asymmetric deep supervised hashing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[8] Xin Jiang, Hao Tang, and Zechao Li. 2024. Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[9] Yoshiyuki Kawano and Keiji Yanai. 2014. Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In *Proceedings of the 22nd ACM international conference on Multimedia*. 761–762.

[10] Xin Lu, Shikun Chen, Yichao Cao, Xin Zhou, and Xiaobo Lu. 2023. Attributes grouping and mining hashing for fine-grained image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6558–6566.

[11] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *Acm Computing Surveys (CSUR)* 52, 5 (2019), 1–36.

[12] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 2019. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 677–686.

[13] Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2017. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *The 28th British Machine Vision Conference*.

[14] Cao Pindan, Min Weiqing, Song Jiajun, YANG Yancun, WANG Lili, et al. 2024. Hash Food Image Retrieval Based on Enhanced Vision Transformer. *Food Science* 45, 10 (2024), 1–8.

[15] 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 *European conference on computer vision*. Springer, 531–548.

[16] Jiajun Song, Zhuo Li, Weiqing Min, and Shuqiang Jiang. 2023. Towards food image retrieval via generalization-oriented sampling and loss function design. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 1 (2023), 1–19.

[17] 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 *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 348–353.

[18] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*. 5551–5560.

[19] Gayathri Vaidyanathan. 2021. What humanity should eat to stay healthy and save the planet.

[20] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. 2022. Learning structural representations for recipe generation and food retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3363–3377.

[21] 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.

[22] Jingdong Wang, Ting Zhang, Nicu Sebe, Heng Tao Shen, et al. 2017. A survey on learning to hash. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 769–790.

[23] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE transactions on image processing* 26, 6 (2017), 2868–2881.

[24] 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 Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13904–13920.

[25] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. 2021. A2-Net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *Advances in Neural Information Processing Systems* 34 (2021), 5720–5730.

[26] Lingxi Xie, Jingdong Wang, Bo Zhang, and Qi Tian. 2015. Fine-grained image search. *IEEE Transactions on multimedia* 17, 5 (2015), 636–647.

[27] Ziyun Zeng, Jinpeng Wang, Bin Chen, Tao Dai, Shu-Tao Xia, and Zhi Wang. 2024. Pyramid hybrid pooling quantization for efficient fine-grained image retrieval. *Pattern Recognition Letters* 178 (2024), 106–114.