语义引导与小波感知的大规模细粒度食品图像检索哈希网络

李杰 [1,*]

1.重庆安全技术职业学院 人工智能学院，重庆 巴南 401320

摘要：大规模细粒度食品图像检索面临类内差异大、类间差异小的双重挑战。深度哈希方法虽具备较高的检索效率，但其浅层特征易受背景噪声干扰，且多尺度特征融合过程中存在显著的语义鸿沟。为此，本文提出一种语义引导与小波感知哈希网络（SWANet）。该网络设计小波感知特征整流（WAFR）模块，利用离散小波变换在滤除高频噪声的同时严格保持食材纹理的空间对齐关系，并通过通道门控机制动态抑制噪声通道、增强有效纹理分量；设计自顶向下语义引导聚合（TSGA）模块，以深层全局语义作为查询，借助交叉注意力自适应地从整流后的局部细节中提取并聚合判别性特征，从而弥合语义鸿沟。在 Food-101、UEC Food-256 与 VireoFood-172 三个公开数据集上的实验表明，SWANet 在 16、32、48、64 位哈希码长下的平均精度均值均优于 HashNet、ADSH 与 SEMICON 等主流深度哈希方法，其中在 Food-101 数据集 16 位码长下较 SEMICON 提升 4.6 个百分点，验证了所提模块的有效性。

关键词：细粒度食品检索；深度哈希；离散小波变换；交叉注意力；多尺度特征融合

SWANet: Semantic-guided and Wavelet-Aware Hashing for Large-Scale Food Image Retrieval

Jie Li[1,*]

1. College of Artificial Intelligence, Chongqing Vocational Institute of Safety Technology, Chongqing 401320, China

Abstract：Large-scale fine-grained food image retrieval is challenging due to low inter-class variances and high intra-class variations. Although deep hashing is highly efficient, existing methods often suffer from background noise in shallow features and semantic gaps during multi-scale feature fusion. To address these issues, this paper proposes a Semantic-guided and Wavelet-Aware Hashing Network (SWANet). Specifically, a Wavelet-Aware Feature Rectification (WAFR) module employs the Discrete Wavelet Transform (DWT) to suppress high-frequency noise while preserving the spatial alignment of food textures. Furthermore, a Top-Down Semantic Guidance Aggregation (TSGA) module utilizes a cross-attention mechanism, using deep global semantics as a query to adaptively extract and fuse clean local details. Extensive experiments on three benchmark datasets (Food-101, UEC Food-256, and VireoFood-172) demonstrate that SWANet achieves state-of-the-art retrieval performance.

Keywords：Fine-grained food retrieval; Deep hashing; Discrete wavelet transform; Cross-attention; Multi-scale feature fusion

1 引言

随着餐饮数字化的快速普及，食品图像数据呈爆发式增长，大规模食品图像检索在智能点餐、膳食营养评估与食品安全溯源等场景中日益重要[1-2]。与通用图像检索不同，食品图像检索是典型的

细粒度检索任务：同一菜品因烹饪方式、摆盘与拍摄条件不同而呈现较大的类内差异，不同菜品在食材构成与外观纹理上又高度相似、类间差异极小[3]。上述特点对检索算法的特征表达能力提出了更高要求。

在实际的大规模应用场景中，检索系统除精度外还需兼顾存储开销与响应速度。深度哈希方法通过端到端学习将高维图像映射为紧凑的二进制编码，并以汉明距离度量样本相似度，可将检索复杂度由高维浮点向量的距离计算降为位运算，因而成为大规模图像检索的主流解决方案[4-6]。

然而，现有深度哈希方法在面向细粒度食品图像时仍存在两个突出瓶颈。其一，浅层卷积特征虽保留了食材的边缘与纹理等局部细节，但同时也引入了大量背景噪声与无关高频伪影；传统频域方法通常借助快速傅里叶变换滤除噪声，但该方法将图像整体变换为全局频谱，导致空间位置信息完全丢失，而食材的空间位置恰恰是细粒度判别的关键依据。其二，现有方法多采用简单通道拼接实现多尺度特征融合，忽略了深层抽象语义与浅层局部细节之间的语义鸿沟[7-8]，使噪声分量与判别性分量被同等对待，从而稀释了最终哈希码的判别力。

针对上述问题，本文提出一种语义引导与小波感知哈希网络（Semantic-guided and Wavelet-Aware Hashing Network，SWANet），主要贡献如下：①提出小波感知特征整流（Wavelet-Aware Feature Rectification，WAFR）模块，利用二维离散小波变换将浅层特征解耦为低频结构分量与高频细节分量，在保持空间位置的前提下通过通道门控抑制噪声通道、增强纹理通道；②提出自顶向下语义引导聚合（Top-Down Semantic Guidance Aggregation，TSGA）模块，以深层全局语义为查询，通过交叉注意力自适应聚合判别性局部特征，弥合语义鸿沟；③在 Food-101、UEC Food-256 与 VireoFood-172 三个公开数据集上开展大量实验，结果表明 SWANet 的平均精度均值（mean Average Precision，mAP）在 16、32、48、64 位码长下均优于现有主流深度哈希方法。

2 相关工作

在细粒度图像检索方面，文献[1]通过改进特征聚合提升了视觉 Transformer 在大规模细粒度检索中的精度；文献[2]指出类内差异建模是食品视觉分析的核心难点；文献[3]提出统一的哈希网络 ExchNet，以特征交换机制增强细粒度判别能力。上述方法在通用细粒度场景下效果较好，但对食品图像中的背景噪声与高频干扰缺乏专门的抑制机制。

在深度哈希检索方面，HashNet[9]与 ADSH[10]均属通用哈希方法，未针对细粒度区域建模，对背景噪声较为敏感。SEMICON[11]引入注意力机制定位判别性区域，在细粒度检索中进步明显，但其空间注意力仍易受高频噪声干扰。文献[5]则验证了频域信息在食品检索中的有效性。与上述方法不同，本文从频域整流与语义引导两个层面协同优化，兼顾噪声抑制与语义对齐。

3 本文方法

3.1 问题定义

给定由N张细粒度食品图像构成的训练集$\mathcal{X}=\{x_i\}_{i=1}^{N}$及其对应的类别标签$\mathcal{Y}=\{y_i\}_{i=1}^{N}$，其中$y_i\in\{1,2,\dots,C\}$，$C$为食品类别总数。细粒度食品图像检索的目标是学习一个端到端的深度哈希网络$\mathcal{F}(\cdot;\Theta)$，其中$\Theta$为网络可学习参数，将输入图像$x_i$映射为$K$位二进制哈希码$b_i\in\{-1,1\}^K$。

为使检索结果准确，所学哈希码应在汉明空间中保持样本间的语义相关性：若两幅食品图像x_i与x_j属于同一类别（$y_i=y_j$），则其哈希码b_i与b_j的汉明距离应尽可能小；反之，若二者类别不同（$y_i\neq y_j$），则该距离应尽可能大。由于食品图像类间差异细微而类内变化剧烈，如何在紧凑编码的约束下生成具有强判别力的哈希码，是本文需要解决的核心问题。

3.2　网络总体结构

大规模细粒度食品图像检索要求模型同时具备三项能力：在复杂背景与噪声中定位细微局部差异、在多尺度特征间建立正确语义关联、以及生成紧凑高效的二进制哈希码。为此，本文提出语义引导与小波感知哈希网络 SWANet，其总体结构如图 1 所示。

SWANet 的核心范式可概括为“浅层特征频域整流、深层语义自顶向下引导”。整个网络由三个紧密耦合的部分组成：多层级特征提取、小波感知特征整流（WAFR）模块与自顶向下语义引导聚合（TSGA）模块。

首先，以在 ImageNet 上预训练的 ResNet-50 为骨干网络提取多尺度视觉表征。与仅使用最后一层特征的常规哈希方法不同，SWANet 显式提取两级特征流：浅层局部特征F_s取自网络中间层，分辨率较高，蕴含丰富的局部纹理与边缘细节，但混杂大量背景噪声与无关高频伪影；深层全局特征F_d取自网络顶层，感受野最大，刻画高度抽象的类别级语义信息，但丢失了细粒度空间定位能力。

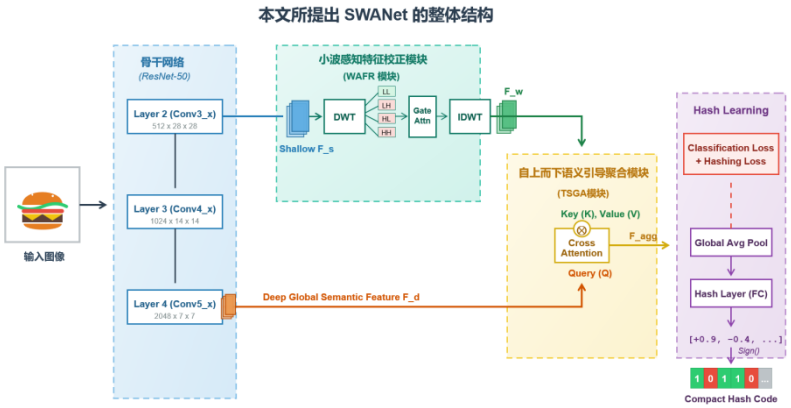


图 1　SWANet 总体结构图

Fig.1　Architecture of the proposed SWANet

由于浅层特征F_s中含有大量无效高频噪声，若直接与深层特征融合将严重削弱哈希码的判别力。因此，本文将其送入小波感知特征整流模块，利用二维离散小波变换将其解耦为低频结构分量与高频细节分量，在完整保持空间位置的前提下通过通道门控抑制噪声通道，并经逆变换重建为洁净且空间对齐的整流特征F_w。

获得频域整流后的局部特征F_w后，如何与深层全局语义F_d有效融合是另一难点。为此，本文设计自顶向下语义引导聚合模块，以深层全局特征为查询（Query），对整流后局部特征所生成的键（Key）与值（Value）进行交叉注意力计算，使网络以高层语义为“罗盘”，从海量低层细节中自适应激活最具判别力的区域，输出聚合特征F_{agg}。

最后，聚合后的高维连续特征经全局平均池化后送入哈希层，映射为K维实值并经Tanh激活与符号量化得到最终二进制码；训练时联合采用分类损失与哈希损失对整个网络进行端到端优化（详见 3.5 节）。

3.3　小波感知特征整流模块

浅层网络能够捕获食品图像丰富的局部纹理，但不可避免地受到背景噪声与高频伪影的污染。现有方法常借助快速傅里叶变换进行频域处理，然而该方法将整幅图像变换为全局频谱，导致空间位置信息完全丢失，而食材的空间位置恰恰是细粒度判别的关键依据。为此，本文提出小波感知特征整流（WAFR）模块，其详细结构如图 2 所示。

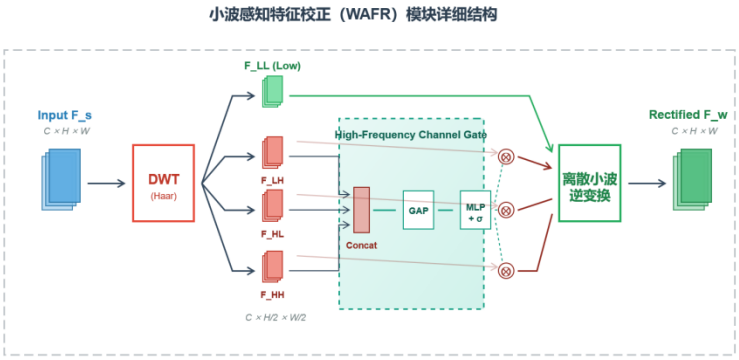


图 2　小波感知特征整流（WAFR）模块结构图

Fig.2　Detailed architecture of the wavelet-aware feature rectification (WAFR) module

具体地，对骨干网络输出的浅层特征图$F_s \in \mathbb{R}^{C\times H\times W}$施加二维 Haar 离散小波变换，将其分解为 1 个低频分量F_{LL}与 3 个高频分量F_{LH}, F_{HL}, F_{HH}，各分量的空间尺寸均为$\frac{H}{2}\times\frac{W}{2}$，如式（1）所示：

$$F_{LL}, F_{LH}, F_{HL}, F_{HH} = \mathrm{DWT}(F_s) \tag{1}$$

其中，F_{LL}保留了食材的主要结构与低频语义信息；高频分量既包含食材边缘等有效细节，也混杂背景噪声。为此，本文设计高频通道门控对其动态滤波：首先将三个高频子带沿通道维拼接，形成联合高频表征$F_{high} \in \mathbb{R}^{3C\times\frac{H}{2}\times\frac{W}{2}}$，如式（2）所示：

$$F_{high} = \mathrm{Concat}(F_{LH}, F_{HL}, F_{HH}) \tag{2}$$

随后，经全局平均池化（GAP）与轻量级多层感知机（MLP）生成通道注意力权重，并由 Sigmoid 函数$\sigma(\cdot)$将其归一化至$(0,1)$区间；权重被切分为(W_{LH}, W_{HL}, W_{HH})三组，分别对各高频子带独立重校准，如式（3）、式（4）所示：

$$[W_{LH}, W_{HL}, W_{HH}] = \sigma(\mathrm{MLP}(\mathrm{GAP}(F_{high}))) \tag{3}$$

$$F'_{LH} = W_{LH} \otimes F_{LH}, F'_{HL} = W_{HL} \otimes F_{HL}, F'_{HH} = W_{HH} \otimes F_{HH} \tag{4}$$

式中，$\otimes$表示沿通道维的逐元素相乘。该门控机制使对应有效食材纹理的通道被增强，而由噪声主导的通道被抑制。

最后，保留的低频结构分量F_{LL}与整流后的各高频分量经二维逆离散小波变换重建至原始空间分辨率，如式（5）所示：

$$F_w = \mathrm{IDWT}(F_{LL}, F'_{LH}, F'_{HL}, F'_{HH}) \tag{5}$$

输出的$F_w \in \mathbb{R}^{C\times H\times W}$即为洁净、去噪且空间对齐的局部特征表征，为后续的语义聚合提供了可靠基础。

3.4　自顶向下语义引导聚合模块

在完成多尺度特征提取后，现有哈希网络通常通过简单的通道拼接实现融合。然而在细粒度食品检索中，深层特征蕴含抽象的类别级语义，浅层特征则涵盖具体的局部纹理，直接拼接将所有特征等同对待，易造成严重的语义鸿沟，稀释最终哈希码的判别力。为此，本文受人类视觉系统启发，提出自顶向下语义引导聚合（TSGA）模块，其结构如图 3 所示。

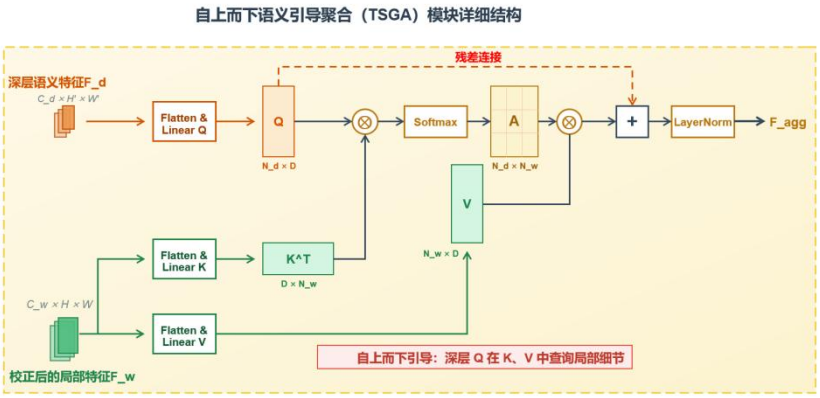


图 3 自顶向下语义引导聚合（TSGA）模块结构图

Fig.3 Detailed architecture of the top-down semantic guidance aggregation (TSGA) module

设深层全局语义特征为$F_d \in \mathbb{R}^{C_d \times H' \times W'}$，WAFR 模块输出的整流浅层局部特征为$F_w \in \mathbb{R}^{C_w \times H \times W}$。首先将二者沿空间维展平为序列，并通过线性变换投影至统一隐空间维度D，对应投影矩阵为W_Q, W_K, W_V，如式（6）—式（8）所示：

$$Q = \text{Linear}_Q(\text{Flatten}(F_d)) \in \mathbb{R}^{N_d \times D} \tag{6}$$

$$K = \text{Linear}_K(\text{Flatten}(F_w)) \in \mathbb{R}^{N_w \times D} \tag{7}$$

$$V = \text{Linear}_V(\text{Flatten}(F_w)) \in \mathbb{R}^{N_w \times D} \tag{8}$$

式中，$N_d = H' \times W'$与$N_w = H \times W$分别为两组特征的序列长度；深层特征充当查询Q，整流后的浅层特征提供键K与值V。

在此基础上，以查询与键转置的矩阵相乘结果经缩放与 Softmax 归一化，得到交叉注意力矩阵$A \in \mathbb{R}^{N_d \times N_w}$，如式（9）所示：

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{D}}\right) \tag{9}$$

注意力矩阵A显式建模了每个全局语义区域与各个局部细节块之间的语义相关度。以该矩阵为权重对值V加权聚合，即得到与全局语义对齐的局部特征，如式（10）所示：

$$F_{cross} = AV \in \mathbb{R}^{N_d \times D} \tag{10}$$

最后，为保留原始全局结构并缓解训练过程中的梯度消失，在原始查询Q与聚合特征F_{cross}之间引入残差连接，并施加层归一化以稳定融合后的表征，如式（11）所示：

$$F_{agg} = \text{LayerNorm}(Q + F_{cross}) \tag{11}$$

经由 TSGA 模块，网络成功弥合了深浅层特征之间的语义鸿沟，所得特征F_{agg}将稳定的菜品全局结构与判别性的食材局部细节有机融合，为后续的二进制哈希编码提供了最优的紧凑表征。

3.5 哈希编码与联合优化

在哈希学习阶段，聚合特征F_{agg}经全局平均池化与全连接层映射至K维连续空间，再经Tanh激活压缩至（-1，1）区间；训练时以连续实值及其符号量化结果共同构造哈希损失，推理时以符号函数获得最终的$b_i \in \{-1,1\}^K$。模型联合采用分类损失与哈希损失进行端到端训练：前者约束网络在类别层面保持语义可分，后者约束同类样本在汉明空间内聚集、异类样本相互远离，从而在保证检索精度的同时显著降低存储与计算开销。

4 实验与结果分析

4.1 实验数据集

为验证 SWANet 的有效性，本文在三个广泛使用的细粒度食品数据集上开展实验。

Food-101：该数据集是极具代表性的大规模食品数据集，包含 101 个食品类别（如汉堡、比萨）共 101 000 幅图像，每个类别恰好 1 000 幅，图像中存在显著的背景噪声与光照条件变化。

UEC Food-256：该数据集面向多样化的亚洲菜系，包含 256 个食品类别共 31 395 幅图像，众多类别在食材构成与外观上高度相似，对细粒度检索提出了极大挑战。

VireoFood-172：该数据集专门面向中国传统菜品构建，包含 172 个类别共 110 241 幅图像，受烹饪方式与摆盘差异影响，图像的类内变化极为剧烈。

对于上述三个数据集，本文严格遵循标准的图像检索协议：随机抽取部分图像构成查询集（测试集），其余图像作为数据库（检索集），并从数据库中随机选取部分图像组成训练集。

4.2 评价指标与实验设置

本文采用平均精度均值（mAP）与精确率—召回率（PR）曲线评价检索性能：mAP 对所有查询样本的平均精度再取平均，综合反映排序质量；PR 曲线刻画不同召回率水平下的精确率变化，用于评估高召回区间的鲁棒性。实验基于 PyTorch 框架，骨干网络为在 ImageNet 上预训练的 ResNet-50；输入图像统一缩放至 224×224，训练时施加随机水平翻转与随机裁剪进行数据增强，分别在 16、32、48、64 位码长下评估性能。

4.3 对比实验结果分析

本文将 SWANet 与三种深度有监督哈希方法 HashNet[9]、ADSH[10] 和 SEMICON[11] 在三个细粒度食品数据集上进行比较。

表 1 给出了不同哈希码长（16、32、48、64 bit）下的 mAP 结果。由表 1 可见：通用深度哈希方法 HashNet 与 ADSH 缺乏细粒度区域建模能力，对背景噪声较为敏感，mAP 明显偏低；SEMICON 借助注意力机制定位判别性区域，性能有所提升，但其空间注意力易受高频噪声干扰。本文 SWANet 在所有码长下均取得最优结果，例如在 Food-101 数据集 16 bit 码长下较 SEMICON 提升 4.6 个百分点。这主要得益于 WAFR 模块在滤除高频噪声的同时保持纹理空间对齐，以及 TSGA 模块以高层语义自顶向下引导提取判别性食材模式、弥合语义鸿沟。此外，SWANet 在使用较短码长（如 32 bit）时即可达到甚至超过对比方法在更长码长下的精度，说明其生成的二进制码兼具紧凑性与高信息量，非常适合实际的大规模食品检索系统。

表 1 三种细粒度食品数据集上不同哈希码长的检索性能对比（mAP/%）

Table 1 Comparison of Retrieval Performance with Different Hash Code Lengths on Three Fine-Grained Food Datasets (mAP/%)

数据集	方法	16 bit	32 bit	48 bit	64 bit
Food-101	HashNet	54.2	58.5	61.1	62.4
	ADSH	56.8	60.1	62.7	63.9
	SEMICON	64.5	68.2	70.4	71.8
	SWANet（本文）	69.1	73.5	75.8	77.1
UEC Food-256	HashNet	51.5	55.4	58.0	59.5
	ADSH	53.9	57.8	60.2	61.7
	SEMICON	61.2	65.4	67.9	69.3
	SWANet（本文）	66.8	71.2	73.6	75.1
VireoFood-172	HashNet	59.8	63.5	65.8	67.2
	ADSH	62.4	66.1	68.3	69.8
	SEMICON	71.5	75.3	77.4	78.9
	SWANet（本文）	76.2	80.4	82.5	83.9

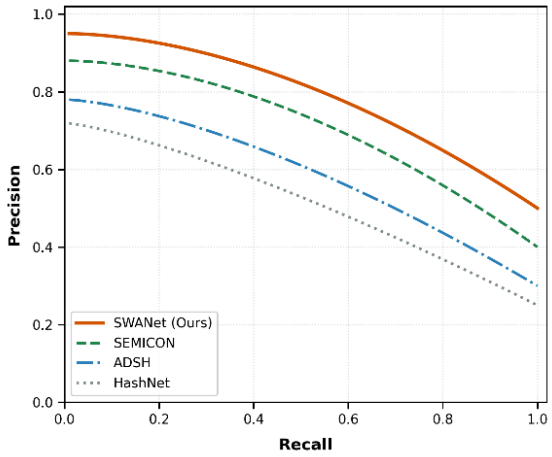


图 4　Food-101 数据集 64 bit 码长下的 PR 曲线对比

Fig.4　Precision-recall (PR) curves of different hashing methods with 64-bit codes on Food-101

图 4 给出了 Food-101 数据集上 64 bit 码长下各方法的 PR 曲线。可以看出，SWANet 的 PR 曲线整体位于对比方法上方，在各召回率水平下均保持更高的精确率；即便在高召回区间（召回率大于 0.8），其精确率下降也明显缓于 HashNet 与 ADSH。这表明 SWANet 所学二进制码保留了鲁棒的细粒度语义特征，进一步验证了 WAFR 与 TSGA 在抑制背景杂波、凸显判别性食材纹理方面的有效性。

4.4　消融实验

为系统验证各模块的独立贡献，本文开展消融实验。将不含任何所提模块的基础网络（ResNet-50＋全局平均池化＋哈希层）记为 Baseline，随后在 Baseline 上分别独立加入小波感知特征整流（WAFR）模块与自顶向下语义引导聚合（TSGA）模块，各变体在三个数据集上 32 bit 与 64 bit 码长下的 mAP 如表 2 所示。

表 2　WAFR 与 TSGA 模块的消融实验结果（mAP/%）

Table 2 Ablation Results of WAFR and TSGA Modules (mAP/%)

模型	Food-101		UEC Food-256		VireoFood-172	
	32 bit	64 bit	32 bit	64 bit	32 bit	64 bit
Baseline	62.1	64.5	59.8	62.2	67.5	70.3
Baseline+WAFR	66.8	69.2	64.6	66.8	73.1	75.6
Baseline+TSGA	68.2	70.5	66.3	68.1	74.8	77.2
SWANet（完整模型）	73.5	77.1	71.2	75.1	80.4	83.9

由表 2 可知：其一，加入 WAFR 模块后检索性能显著提升，如 Food-101 数据集 32 bit 码长下 mAP 由 62.1% 提升至 66.8%，说明离散小波变换滤除高频背景噪声有助于提取更纯净、更鲁棒的食材纹理。其二，仅加入 TSGA 模块同样带来明显增益（提升 6.1 个百分点），表明以深层语义引导局部细节选择的交叉注意力机制显著优于直接拼接，成功弥合了语义鸿沟。其三，完整模型 SWANet 在所有数据集上均取得最优性能，说明两模块强互补：WAFR 提供洁净且空间对齐的局部特征（高质量的键与值），TSGA 提供准确的自顶向下语义查询，协同生成高判别力且紧凑的哈希码。

5　结论

本文针对大规模细粒度食品图像检索中背景噪声干扰与多尺度特征融合语义鸿沟的问题，提出了语义引导与小波感知哈希网络 SWANet。该方法通过小波感知特征整流模块，利用离散小波变换在滤除高频噪声的同时严格保持食材纹理的空间对齐，并通过自顶向下语义引导聚合模块，借助交叉注意力机制实现高层语义引导下的特征融合。在三个公开数据集上的实验表明，SWANet 在 16、32、48、64 位码长下的 mAP 均优于现有主流深度哈希方法，且在使用较短码长时即可达到对比方法更长码长的精度水平，验证了所提方法的有效性。后续将考虑把该框架拓展至其他细粒度图像检索场景，并探索其在移动端的轻量化部署。

参考文献

[1] 张艳荣,井绪静,陈彬,等.共同注意力与自适应门控的电商图文检索[J/OL].计算机系统应用,1-13.https://doi.org/10.15888/j.cnki.csa.010350.
[2] 杨明凤,郭楠.基于多人脸图像检索和人物社交关系抽取的合影档案管理研究[J].中国档案,2026,(9):66-68.
[3] 姜坤,钱文华,刘朋.语义增强与文本重组协同的汉画像石图文检索[J].中国图象图形学报,2026,31(8):3030-3044.
[4] 郭争龙,陈若男,周昕,等.云环境中跨模态特征协同的加密图像检索方法[J].科学技术与工程,2026,26(19):8286-8297.
[5] 邢涵,徐子靖,李卫东.一种基于深度哈希的大规模图像搜索方法[J].河南科技,2026,53(12):34-37.
[6] 龙宝壮,廖胜文,李佳.基于变换分析技术的烟草数据库高效检索算法[J].微型电脑应用,2026,42(6):4-8.
[7] 陈则林.基于 SIFT 特征的小径管环焊缝数字射线图像检索方法[J].上海化工,2025,50(6):34-39.
[8] 张杰,黄智勇,王瑞锦,等.基于图像编辑代理的零样本遥感图像复合检索算法[J].计算机应用,2026,46(8):2577-2583.
[9] CAO Z J, LONG M S, WANG J M, et al. HashNet: deep learning to hash by continuation[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1999-2008.
[10] JIANG Q Y, LI W J. Asymmetric deep supervised hashing[EB/OL]. arXiv: 1707.08325, 2017.
[11] SHEN Y, SUN X H, WEI X S, et al. SEMICON: a learning-to-hash solution for large-scale fine-grained image retrieval[EB/OL]. arXiv: 2209.13833, 2022.

基金项目：重庆市教育委员会科学技术研究项目（KJQN202504708）；重庆安全技术职业学院校级科研项目《基于对比学习的多模态图像检索语义对齐技术研究》（AQJSZR2510）

[1,*]第 1 作者&通讯作者简介：李杰（1998—），男，重庆安全技术职业学院人工智能学院专任教师，硕士，研究方向：人工智能、计算机视觉与多模态图像检索。E-mail：18302328324@163.com。