基于增强 Vision Transformer 的哈希食品图像检索

曹品丹[1]，闵巍庆[2]，宋佳骏[3]，盛国瑞[1]，杨延村[1]，王丽丽[1,*]，蒋树强[2]
(1.鲁东大学信息与电气工程学院，山东 烟台 264025；2.中国科学院计算技术研究所，北京 100190；
3.中国人民大学农业与农村发展学院，北京 100872)

摘 要：作为食品计算的一个主要任务，食品图像检索近年来受到了广泛的关注。然而，食品图像检索面临着两个主要的挑战。首先，食品图像具有细粒度的特点，这意味着不同食品类别之间的视觉差异可能很小，这些差异只能在图像的局部区域中观察到。其次，食品图像包含丰富的语义信息，如食材、烹饪方式等，这些信息的提取和利用对于提高检索性能至关重要。为了解决这些问题，本文基于预训练的视觉 Transformer（Vision Transformer，ViT）模型提出了一种增强 ViT 的哈希网络（Enhanced ViT Hash Network, EVHNet）。针对食品图像的细粒度特点，EVHNet 中设计了一个基于卷积结构的局部特征增强模块，使网络能够学习到更具有代表性的特征。为了更好地利用食品图像的语义信息，EVHNet 中还设计了一个聚合语义特征模块，根据类令牌特征来聚合食品图像中的语义信息。本文提出的 EVHNet 模型在贪婪哈希（Greedy Hash，GreedyHash）、中心相似量化（Central Similarity Quantization，CSQ）和深度极化网络（Deep Polarized Network，DPN）三种流行的哈希图像检索框架下进行了评估，并与 AlexNet，ResNet50、ViT-B_32 和 ViT-B_16 四种主流的网络模型进行了比较，在 Food-101、Vireo Food-172、UEC Food-256 三个食品数据集上的实验结果表明 EVHNet 模型在检索精度上的综合性能优于其他模型。

关键词：食品图像检索；食品计算；哈希检索；Vision Transformer 网络；深度哈希学习

Hash Food Image Retrieval Based on Enhanced Vision Transformer

CAO Pindan[1], MIN Weiqing[2], SONG Jiajun[3], SHENG Guorui[1], YANG Yancun[1], WANG Lili[1,*], JIANG Shuqiang[2]
(1. School of Information and Electrical Engineering, Ludong University, Yantai, Shandong 264025, China;
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
3. School of Agricultural and Rural Development, Renmin University of China, Beijing 100872, China)

Abstract: Food image retrieval, as a major task in food computing, has garnered extensive attention in recent years. However, it faces two primary challenges. Firstly, food images exhibit fine-grained characteristics, implying that the visual differences between different food categories can be subtle and often only observable in local regions of the image. Secondly, food images contain abundant semantic information, such as ingredients and cooking methods, the extraction and utilization of which are crucial for enhancing retrieval performance. To address these issues, this paper proposes an Enhanced ViT Hash Network (EVHNet) based on the pre-trained Vision Transformer (ViT) model. To cater to the fine-grained nature of food images, a Local Feature Enhancement Module based on convolutional structure is designed in EVHNet, enabling the network to learn more representative features. To better leverage the

基金项目：国家自然科学基金(61705098；61872170)；山东省自然科学基金(ZR2023MF031)；
第一作者简介：曹品丹(1998—)（ORCID: 0009-0004-0662-4241），女，硕士研究生，研究方向为食品计算和计算机视觉。E-mail: caopindan@m.ldu.edu.cn
*通信作者简介：王丽丽(1978—)（ORCID: 0000-0002-1025-3955），女，教授，博士，研究方向为宽带通信和多媒体通信。E-mail: wanglili@ldu.edu.cn

semantic information in food images, an Aggregated Semantic Feature Module is designed in EVHNet, aggregating the semantic information in food images based on class token features. The proposed EVHNet model is evaluated under three popular hash image retrieval frameworks, namely Greedy Hash (GreedyHash), Central Similarity Quantization (CSQ), and Deep Polarized Network (DPN), and compared with four mainstream network models, AlexNet, ResNet50, ViT-B_32, and ViT-B_16. Experimental results on the Food-101, Vireo Food-172, and UEC Food-256 food datasets demonstrate that the EVHNet model outperforms other models in terms of comprehensive retrieval accuracy.

Keywords: food image retrieval; food computing; hash retrieval; vision transformer network; deep hash learning
中图分类号：S126　　文献标志码：A
DOI：10.7506/spkx1002-6630-20231231-270

食品在人类生活中占据了重要地位，与人类的健康和文化密切相关，因此食品相关研究成为热点。在数字化的社会背景下，食品科学正在发生巨大的变革，数字化的发展为食品科学带来了新的机遇[1]，张南等[2]分析了我国食品科学领域中具备交叉学科的研究方向，并展示了我国食品科学方向主要的交叉学科竞争优势。食品计算[3]作为食品科学和计算机科学交叉学科的主要研究之一，旨在利用人工智能，数据处理与分析等技术对食物本身的营养特性、原材料和制造过程中的营养特征变化等信息进行数据化和整合，通过分析这些数据，可以解决感知、分类、检索、推荐以及预测等问题。

食品图像检索，作为食品计算的一项基本任务，主要实现了“以图搜图”的功能，通过输入的查询图像从食品图像数据库中找到所有相似的图像。对于食品图像检索方法，梅舒欢等[4]提出了一种基于 Faster R-CNN 的食品图像检索和分类的方法，首先微调用于目标检测的 Faster R-CNN[5]模型，使其检测图像中的食品区域，之后利用卷积神经网络（Convolutional Neural Network，CNN）[6]提取食品区域的视觉特征，最后将提取的视觉特征应用到食品图像检索和分类任务中。Song 等[7]针对食品图像易受背景无关噪声影响和细粒度的特点，提出了一种噪声鲁棒性的局部 Transformer 网络。Song 等[8]针对食品图像中不同领域类别间的差异较大而导致的泛化问题，提出了一种面向泛化的食品图像检索分析方法。

食品图像检索技术的发展，为消费者在海量食品图像中寻找所需食品提供了便利，满足了不同人群的个性化需求。因此，食品图像检索在推动食品计算和食品服务业[9]的发展等方面具有重大的研究价值和应用前景。尽管食品图像检索任务具有很大的研究潜力，但两个主要的挑战阻碍了这种潜力的充分发掘。首先，食品图像检索属于细粒度任务[10]，但又区别于传统的细粒度任务特点，传统的细粒度任务特点是类内差异和类间差异都小。而食品图像检索与传统细粒度任务的不同在于其类内差异大而类间差异小。例如，就寿司这一种食品来说，有许多不同的形式，包括 nigiri（握寿司），maki（卷寿司），sashimi（刺身）等，每种都有独特的外观，它们的颜色和形状可以根据其配料和制作方法而变化。这种多样性导致了类内的视觉差异性，对食品图像检索的精度提出了严格的要求。然而，两种完全不同的食品如水煮鱼和水煮肉片，可能因为烹饪方法和酱料的影响，在视觉上表现出高度的相似性，这增加了食品图像检索任务的难度。其次，食品图像包含更丰富的语义内容，而且在实际应用中常常会受到大量与食品无关的噪声信息的干扰，如餐具和配菜，这些与食品无关的信息会极大地影响食品图像特征的语义性。传统的检索策略主要针对具有明显几何形状的刚性建筑物或具有显著特征的实体（如车辆，鸟类等）进行结构设计，这些策略通常专注于检索目标的特点，并未充分考虑食品图像的独特性。如果采用的是通用检索方法而未对特定目标进行优化，这些方法在特定任务上的性能往往无法达到最优。因此，传统的检索框架在食品图像检索任务上的性能往往无法达到理想的效果。

随着食品数据的快速增长，如何提高食品图像检索的速度也是需要考虑的问题。由于 CNN 的快速发展，基于深度哈希学习[11]的 CNN 检索方法已经被深度关注。深度哈希检索充分利用了深度学习在特征提取方面的优势，以及哈希技术在大规模数据检索中的高效性，从而实现了高效且快速的检索。深度哈希检索首先利用 CNN 对输入图像进行特征提取，得到图像的特征表示。然后将特征表示输入到哈希层，通过哈希函数将特征映射为哈希码。最后将查询图像的哈希码与数据库图像的哈希码进行

相似性匹配，得到最相似的图像。深度监督哈希（Deep Supervised Hashing，DSH）[12]利用 CNN 将网络输出量化为二进制哈希码，在实值网络输出上使用正则化器产生离散的二进制值。HashNet[13]引入了基于 tanh 函数的拓展方法，实现了实值特征平滑过渡到二进制码。贪婪哈希（Greedy Hash，GreedyHash）[14]在哈希层中使用 sign 函数。改进的深度哈希网络（Improved Deep Hashing Network，IDHN）[15]利用交叉熵损失和均方误差损失实现多标签图像检索。中心相似度量化（Central Similarity Quantization，CSQ）[16]对数据点之间的中心相似度进行哈希中心优化。深度极化网络（Deep Polarized Network，DPN）[17]利用位铰链损失使不同的输出通道都远离零。这些流行的检索框架已经证明将哈希学习 [18]应用到通用的图像检索领域可以有效地提高效率。因此，本文的研究主要在于使用深度哈希学习的方法来提高食品图像检索效率。

尽管 CNN 在实现图像特征空间的全局表示方面已取得显著成果，但其精确度的提升仍依赖于网络深度的增加。近年来，Transformer[19]网络作为深度学习的新趋势，展现出了巨大的潜力。视觉 Transformer（Vision Transformer，ViT）[20]在图像识别 [21]、图像分割 [22]、行人再识别 [23]、多模态检索 [24]等视觉任务中表现出优异的性能。利用 ViT 进行哈希图像检索的研究也日渐增多。例如，TransHash[25]利用连体 ViT 进行特征学习，但 TransHash 无法在大规模数据集上进行图像哈希。Dubey 等 [26]提出了一种用于图像检索的视觉 Transformer 哈希（Vision Transformer Hashing，VTS），VTS 模型利用 ViT 作为通用特征提取器，对哈希检索工作进行了深入研究。本文在 VTS 的基础上结合食品图像的特性，对网络进行了优化和改进。通过与当前先进的网络模型 AlexNet，ResNet50，ViT-B_32 和 ViT-B_16 进行比较，并对改进的 EVHNet 模型进行了性能评估。实验结果表明，本文提出的方法提高了食品图像检索的性能。

1 材料与方法

1.1 数据集及划分

本研究采用了三个流行的食品数据集，为了避免数据分布不平衡的影响，根据检索协议 [13]对数据集中的每个类别进行了重新划分（食品数据集划分见表 1）。此外，本文还对每个数据集的类别标签进行了 one-hot 编码，用于将离散的分类标签转换为二进制向量，并分别进行了评估。

Food-101 数据集 [27]包含来自 101 个食品类别的图像数据集，主要用于图像分类，共有 101000 张食品图像，每个类别的测试图像有 250 张，训练图像 750 张。本文对其进行了重新划分，每个类别随机选择 100 张图像进行训练，50 张图像进行测试。

Vireo Food-172 数据集 [28]将食品分为 172 个大类，每个大类中有 200-1000 张从百度和谷歌图像搜索中抓取的食品图片，基本覆盖日常生活中的绝大多数食品种类，共有 110241 张食品图像，本文随机对每个类别抽取了 60 张图片用于训练，30 张图片用于测试。

UEC Food-256 数据集 [29]包含 256 种美食的 31395 张图像。每个类别随机取 40 张图像进行训练，20 张图像进行测试。

表 1 食品数据集划分情况

Table 1 Division of food datasets

集合类别	Food-101	Vireo Food-172	UEC Food-256
Train	10100	10320	10240
Test	5050	5160	5120

1.2 EVHNet 模型的构建方法

1.2.1 模型设计动机

根据 CNN 和 Transformer 的相关研究 [30]，混合模型在较小的计算下略优于纯 CNN 或纯 Transformer

结构。最常见的方法是将卷积结构和归纳偏置引入到 Transformer 模型本身，其输入仍然是原始块[31]。这种混合模型的设计理念适用于食品图像检索模型的构建。

在食品图像检索任务中，一个高效的模型应该关注目标对象，而非背景或其他噪声干扰。因此，最近的方法无论是关注全局还是局部表示，在架构中通常包含一个全局分支和一个局部分支[32]。局部分支的目标是改善模型的定位特性。卷积网络在这方面表现出色，因为它不仅具有良好的定位特性，还具有局部化的交互机制。全局分支的目标是收集更多的全局表示。Transformer 通过依赖全局自注意力机制，在建模长距离依赖关系方面表现出优越的特性。此外，在一些密集预测任务中，通常使用来自不同网络层的不同尺度的特征，从而形成特征金字塔[33]。在结构中引入跳跃连接，稀疏或密集的跨层连接[34]也很常见。这些方法在食品图像检索任务上也具有显著的影响。

因此，本文提出了一种用于食品图像检索的模型，该模型结合了卷积网络的局部化交互方式和 Transformer 的全局自注意力机制，实现了食品图像特征的增强表示。针对食品图像细粒度任务的特点，引入了卷积结构来设计局部特征分支，使得网络能够捕获食品图像中更具代表性的细粒度特征。为了更好地利用食品图像中包含的丰富语义信息，本文研究了 Transformer 编码器的层次连接，聚合了不同尺度的特征，并且受 ViT 中类令牌嵌入具有全局语义信息的影响[35]，构建了基于类令牌的全局特征分支。局部与全局特征最终融合，用于生成哈希码的增强表示，使得模型在食品图像检索任务上的性能得到了有效提升。

1.2.2 ViT 预训练模型

ViT 框架的工作流程如图 1 所示。首先，输入的食品图像$I \in \Re^{m,m,c}$被划分为 N 个不重叠且固定大小的块$I_i \in \Re^{k,k,c}$，其中$i = 1,2,\ldots,N$，m是输入食品图像的大小，k是划分块的大小，满足$Nk^2 = m^2$，$c = 3$代表 RGB 三种颜色的通道数。然后，输入块I_i被展平为向量$V_i \in \Re^{1,d}$，其中$d = c \times k^2$，$i = 1,2,\ldots,N$。每个展平的向量经过线性投影，生成块投影嵌入（Patch Embedding , PE）：

$$PE_i = V_i \times W_{PE} \tag{1}$$

其中，$V_i \in \Re^{1,d}$为第i个块对应的展平向量，$W_{PE} \in \Re^{d,de}$为参数矩阵，$PE_i \in \Re^{1,de}$为第i个块对应的投影嵌入。这里，$de = 768$是投影的嵌入维数。因此，与整个输入的食品图像相对应的投影嵌入可以表示为$PE \in \Re^{N,de}$。

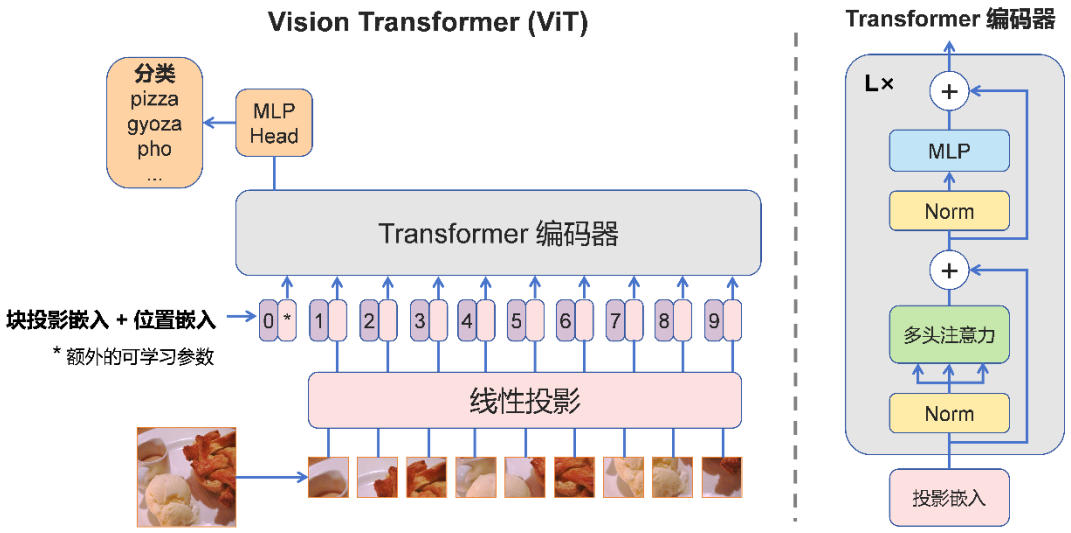

图 1 ViT 模型结构示意图

Fig.1 ViT model structure diagram

ViT 中引入了一个类令牌嵌入$CLS \in \Re^{1,de}$，作为用零初始化的可学习参数，并与投影嵌入拼接生成第一维的扩展嵌入$EE \in \Re^{N+1,de}$。为了融合空间位置信息，将位置嵌入$PoE \in \Re^{N+1,de}$作为可学习参

数，初始化为零，添加到扩展嵌入中。因此最终的投影嵌入 FE 为：

$$FE = dropout(EE + PoE, 0.1) \tag{2}$$

其中，$dropout$是一个因子为 0.1 的$dropout$层，代表每个神经元都有 10%的概率被临时从网络中移除，这是一种正则化技术，用于防止网络过于复杂而导致过拟合。$FE \in \Re^{N+1,de}$是最终的投影嵌入输出，也是 Transformer 编码器的输入。

Transformer 编码器由 L 个 Transformer 块的堆叠组成。在自注意力机制的帮助下，Transformer 块将给定的隐藏层输入转换为相同维度的输出。每个块都包含两个重要的部分，第一个是多头注意力（Multi-Headed Self Attention, MSA），它对不同的输入令牌执行自注意力操作。另一个是前向传播网络（Feed Forward Network, FFN），用来更新权重，优化模型。公式表达为：

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{3}$$

$$FFN(x) = \sigma\left((xW_1 + b_1)W_2 + b_2\right) \tag{4}$$

其中，Q，K和V分别为查询（Query）、键（Key）和值（Value）。在多头注意力中，先将输入序列中的令牌线性变换到QKV空间中的查询、键和值，如果只有一个头，QKV维度都是$(N+1)\times d$，如果有 12 个头，则QKV的维度是$(N+1)\times a$，满足$12\times a = d$。最后进行一个拼接操作，输出维度是$(N+1)\times d$。这里的W_1和W_2是两个线性变换的权重，$\sigma(\cdot)$是 ViT 中采用的非线性激活函数$GELU$ [36]。最后，通过多层感知器头（MLP Head）输出最后的分类结果。

1.2.3 增强 ViT 网络的建立

如图 2 所示，本文提出的 EVHNet 框架对 ViT 进行了一系列的修改。首先，EVHNet 移除 ViT 中的 MLP Head。然后，对 Transformer 编码器进行了多次迭代，并在局部和全局层面上引入了两个分支模块，即局部特征增强模块（Local Feature Enhancement Module，LFEM）和聚合语义特征模块（Aggregated Semantic Feature Module，ASFM）。最后，将两个分支上的特征信息进行融合，输入到哈希层，哈希层将生成的高维特征映射为低维的哈希码，从而实现了食品图像的快速检索。

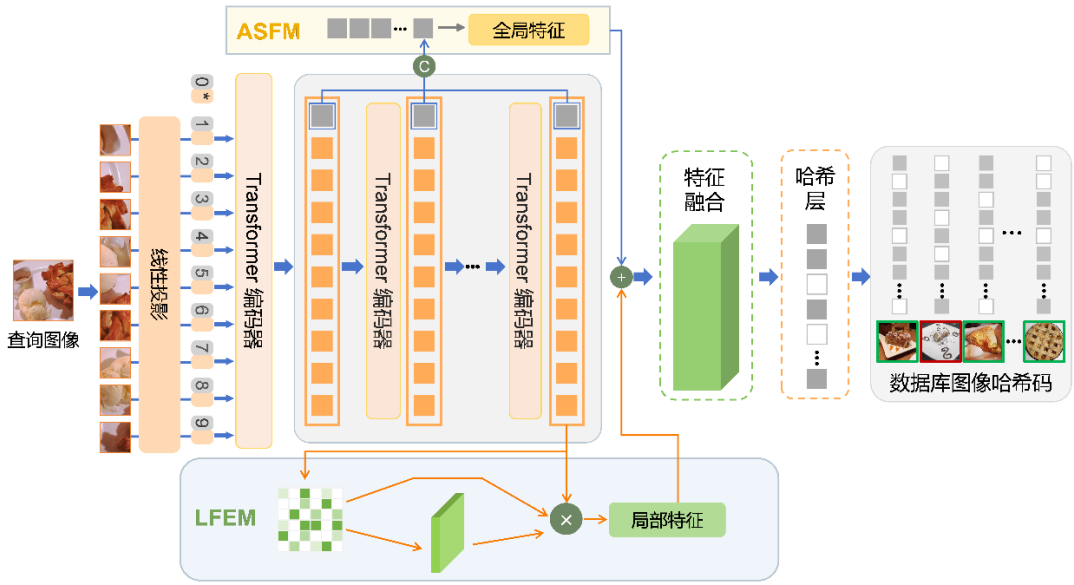

图 2 EVHNet 模型结构示意图

Fig.2 EVHNet model structure diagram

在经过 K 次迭代之后，Transformer 编码器的最终输出为X_f。X_f经过局部特征增强模块和聚合语义特征模块，分别获取全局特征F_g和局部特征F_l。将F_g和F_l融合，得到融合后的特征F_f。F_f经过哈希层映射为指定位数的哈希码H_F：

$$H_F = H(F_f) \tag{5}$$

其中，$H(\cdot)$是哈希映射层，将 Transformer 编码器的最终输出X_f重构为$(N+1)\times de$，并通过一个$dropout$层，将特征线性投影为 1024 维，之后使用$ReLU$激活函数以及一个线性投影来生成最终的哈希特征H_F，其值的数量与哈希长度相同。最后将生成的哈希码与数据库哈希码进行相似度比较，返回与输入图像最相似的图像。

（1）聚合语义特征模块—ASFM

类令牌在 Transformer 编码器的 K 次迭代中不断更新，能够编码整个数据集的统计特性。该令牌对所有其他令牌上的信息进行汇聚（全局特征聚合），捕获输入数据的全局信息，生成一系列的全局特征表示。

在给定参数 L 的情况下，CL收集多层 Transformer 编码器的CLS嵌入：

$$CL = \{CLS_1, CLS_2, \ldots, CLS_K\} \tag{6}$$

然后将CL收集的嵌入进行拼接，得到全局特征表示F_g：

$$F_g = concat\left[CL_{[0]}, CL_{[1]}, \ldots, CL_{[K-1]}\right] \tag{7}$$

（2）局部特征增强模块—LFEM

最后一层 Transformer 编码器输出的特征X_f包含丰富的空间信息，因此，LFEM 将特征X_f作为输入，如图 3 所示。首先，通道注意力[37]模块（CAM）在空间维度上对特征X_f进行压缩。在压缩过程中，不仅使用了平均池化，还引入了最大池化作为补充。通过这两个池化函数，得到了两个一维向量。压缩后，分别对它们进行多层感知器（MLP）连接。CAM 对特征进行加权：

$$M_C(X_f) = \sigma\left(MLP\left(AvgPool(X_f)\right) + MLP\left(MaxPool(X_f)\right)\right) \tag{8}$$

其中$\sigma(\cdot)$是 sigmoid 激活函数。

然后，将 CAM 的输出经过单个卷积核的隐藏层进行卷积和 sigmoid 操作，并与 CAM 的输出和 Transformer 编码器的输出X_f进行逐元素乘法操作，得到局部特征的输出F_l：

$$F_l = \sigma\left(Conv\left(M_C(X_f)\right)\right) \times M_C(X_f) \times X_f \tag{9}$$

增强的局部模块能够自适应地学习和关注更重要的特征，并抑制不重要的特征，提高了模型的可解释性。

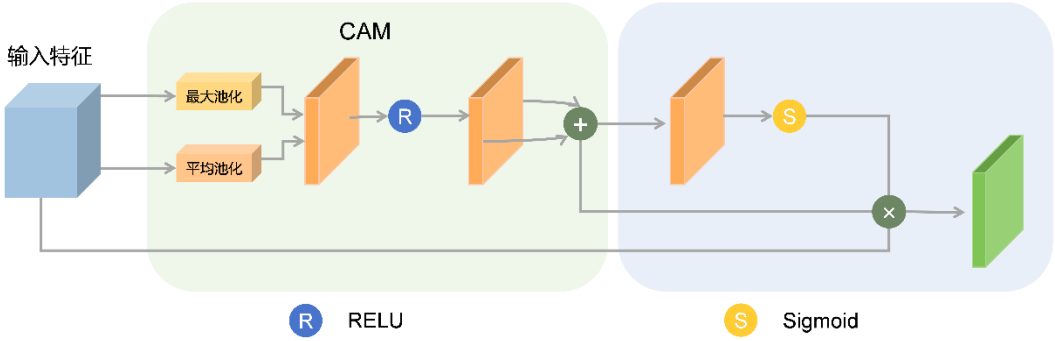

图 3 LFEM 模块结构示意图

Fig.3 LFEM module structure diagram

1.3 模型评估指标

在本研究中，采用平均精度均值（mean Average Precision，mAP）作为评估所提方法性能的主要指标。本文中的 mAP 均指 mAP@1000，它反映了前 1000 个查询结果的性能。具体而言，mAP@1000 是前 1000 个返回结果中，每个正样本所在位置对应的精确率（Precision）的平均值。

为了计算 Average Precision（AP），首先需要理解混淆矩阵（Confusion Matrix）的概念，混淆矩阵如表 2 所示：

表 2　混淆矩阵

Table 2　Confusion matrix

Confusion		预测值	
		负例	正例
真实值	负例	TN（真负例）	FP（假正例）
	正例	FN（假负例）	TP（真正例）

混淆矩阵中四个值的具体含义如下：

TP（True Positive）：被正确预测的正例。即该数据的真实值和预测值都为正例；

TN（True Negative）：被正确预测的负例。即该数据的真实值和预测值都为负例；

FP（False Positive）：被错误预测的正例。即该数据的真实值为负例，但被错误预测为正例；

FN（False Negative）：被错误预测的负例。即该数据的真实值为正例，但被错误预测为负例。

Precision 也称为查准率，是预测正确的正例数据占预测为正例数据的比例。其计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

Recall 也称召回率，是预测为正例的数据占实际为正例数据的比例。其计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Precision/Recall 曲线也叫 PR 曲线，是所有 Precision-Recall 点连成的曲线，用于计算 AP 值。AP 是 PR 曲线下的面积，它是精确率和召回率的函数，取值范围在 0 到 1 之间，AP 越高，代表模型性能越好。

1.4　实验环境

（1）NVIDIA A800 80G GPU：NVIDIA A800 是一款专为高性能计算和人工智能任务设计的 GPU 显卡，具有 80GB 的图形内存。这款显卡在科学研究领域得到了广泛的应用，其强大的计算能力可以有效地加速模拟、数据分析和可视化等过程，从而显著提高研究效率。

（2）深度学习框架 PyTorch：PyTorch 是 Facebook 开发的一款开源机器学习框架，旨在为研究人员提供一个灵活且直观的环境，以便于编写和训练自定义神经网络模型。该框架支持动态图和静态图计算模式，具有丰富的功能和强大的扩展性。在计算机视觉领域，PyTorch 被广泛应用于图像识别、目标检测和图像检索等任务。

2　结果与分析

2.1　实验设置

本实验使用预训练的 ViT 模型，所有的输入图像都调整为 224*224。实验中设置了 EVHNet 网络的两种变体，即 EVHNet32 和 EVHNet16，它们的块大小分别为 32 和 16。生成的哈希码长度分别为 16 位、32 位和 64 位。使用 Adam 优化器对所有模型进行 150 个 epoch 的训练，批处理大小为 32。测试每 30 个 epoch 报告一次，且报告最佳结果。

2.2　性能分析

表 3 总结了在 Food-101、Vireo Food-172、UEC Food-256 三个数据集上的实验结果。本文所提出的 EVHNet32 和 EVHNet16 模型与 AlexNet、ResNet50、ViT-B_32 和 ViT-B_16 模型进行了比较。结果是在三个检索框架（即 GreedyHash、CSQ 和 DPN）下使用 16 位、32 位、64 位哈希码计算的。实

验结果表明，EVHNet16 在三个数据集上的三种检索框架下均表现出良好的性能。尤其是在 16 位的低哈希码位数下，其效果更为显著。这主要是因为在特征学习过程中，模型更加侧重于主要特征的学习，即使在映射到低哈希码位数时会损失一部分特征信息，但映射后的低哈希码所包含的特征更具代表性，从而提升了检索性能。表 3 中还可以观察到，在三种检索框架下，64 位哈希码的检索性能均优于 16 位和 32 位，这主要归因于深度哈希在将高维特征向量映射为低维的二进制哈希码时，一些干扰信息被丢弃，只保留了重要的特征信息。相比于 16 位和 32 位哈希码，64 位哈希码提供了更大的输出空间，这意味着它能够映射到更多的唯一输出，从而更好的保留深层特征信息，因此，64 位哈希码在提高检索精度方面具有显著优势。

表 3　食品数据集上的实验结果

Table 3　Experimental results on food datasets

Network	GreedyHash			CSQ			DPN		
	16b	32b	64b	16b	32b	64b	16b	32b	64b
Result on Food-101 dataset (mAP@1000 in %)									
AlexNet	27.5	41.2	47.9	32.3	46.1	50.2	31.8	44.3	50.7
ResNet50	50.3	59.9	63.4	54.1	63.0	65.9	53.5	60.7	64.2
ViT-B_32	64.4	68.4	70.5	65.3	70.3	70.0	63.6	70.0	71.0
ViT-B_16	72.6	76.4	77.9	73.5	76.8	78.0	73.8	76.9	78.0
EVHNet32	63.7	69.5	71.4	63.6	70.0	69.9	65.1	68.2	71.4
EVHNet16	**73.6**	**77.4**	**78.1**	**74.1**	**77.8**	**78.2**	**74.7**	**77.6**	**78.4**
Result on Vireo Food-172 dataset (mAP@1000 in %)									
AlexNet	35.1	50.9	57.7	41.8	56.5	59.2	40.6	54.5	59.9
ResNet50	56.7	66.6	69.9	60.1	68.7	70.4	60.3	67.9	70.3
ViT-B_32	63.3	68.5	70.6	60.5	70.1	70.7	61.2	69.4	70.4
ViT-B_16	73.3	**76.1**	78.2	71.1	**76.8**	**78.6**	71.5	77.2	**78.3**
EVHNet32	58.7	63.9	65.1	61.5	68.8	70.4	63.8	65.1	69.4
EVHNet16	**74.1**	**76.1**	**79.0**	**72.2**	75.7	78.1	**71.6**	**77.6**	76.8
Result on UEC Food-256 dataset (mAP@1000 in %)									
AlexNet	29.7	38.8	45.7	36.2	45.3	50.8	35.6	44.6	50.2
ResNet50	44.7	55.1	58.9	45.2	58.7	61.6	43.8	57.2	60.1
ViT-B_32	55.0	64.0	66.8	51.8	61.5	64.3	51.4	63.0	64.7
ViT-B_16	63.5	69.8	71.5	59.9	67.8	**70.7**	59.1	68.1	70.5
EVHNet32	53.3	61.2	66.2	53.3	64.4	65.4	53.4	64.2	62.6
EVHNet16	**63.8**	**69.9**	**72.1**	**61.0**	**68.7**	69.7	**60.3**	**68.3**	**70.7**

图 4 详细地展示了在 Food-101、Vireo Food-172、UEC Food-256 数据集上，各种骨干网络（包括 AlexNet、ResNet50、ViT-B_32、ViT-B_16、EVHNet32 和 EVHNet16）在 16 位哈希码下的精确召回 RoC 曲线。

从图中可以清晰地观察到，在 DPN 检索框架下，所提出的 EVHNet32 骨干网络在大多数情况下

的表现优于 AlexNet、ResNet50、ViT-B_32 骨干网络。而且，EVHNet16 骨干网络在三个数据集上的表现都优于其他网络，实现了最佳性能。EVHNet16 之所以能够实现更高的准确率，主要是因为它采用了更小的块（EVHNet16 将输入图像切割成 16*16 的块，而 EVHNet32 将输入图像切割成 32*32 的块），这使得 EVHNet16 能够捕获图像中的更多细节信息，从而更好地保留原始图像的信息，提高了模型的性能。

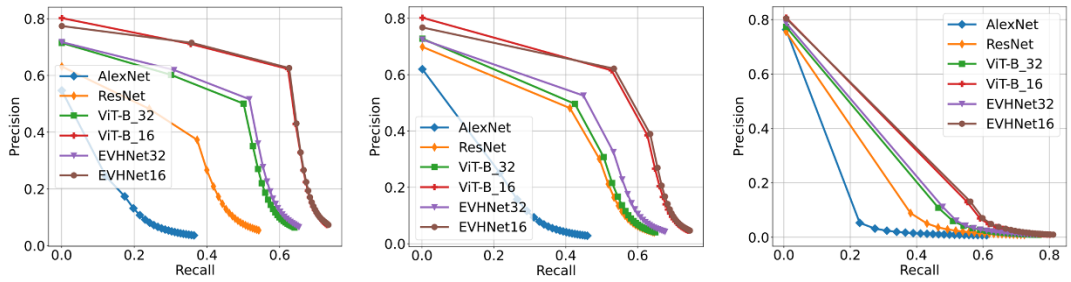

图 4　16 位哈希码下的精确-召回 RoC 曲线

Fig.4　Precision-Recall RoC curves on 16-bit hash code

2.3　消融实验

2.3.1　不同模块的消融实验

为了评估模型中各个模块对实验结果的影响，本实验中选择使用 ViT-B_16 模型作为基准。考虑到 Food-101 数据集的丰富性和多样性，因此选择在 Food-101 数据集上对两个分支模块进行了实验，以确保结果的普遍适用性。

如表 4 所示，实验结果表明，两个分支模块都对检索性能产生了积极的影响。ASFM 利用卷积结构，使网络捕捉食品图像中的细微差异，学习和提取更细粒度的特征，这使得最终映射的哈希码更具有代表性，从而提高了检索性能。LFEM 通过多层类令牌特征的聚合，学习到了食品图像中丰富的语义信息，这种丰富的语义表示有助于提高检索性能。然而，当两个模块同时存在时，全局类令牌特征和局部特征的融合改善了最终特征的表示，使得模型的综合性能达到了最优。这进一步证实了卷积局部化交互的重要性和全局类令牌的互补性。实验结果充分证明了本研究提出的模块在食品图像检索任务上的有效性。

表 4　ASFM 和 LFEM 在 Food101 数据集上的消融对比实验

Table 4　Ablation comparison experiments of ASFM and LFEM on Food101 dataset

Network	GreedyHash			CSQ			DPN		
	16b	32b	64b	16b	32b	64b	16b	32b	64b
ViT-B_16	72.6	76.4	77.9	73.5	76.8	78.0	72.3	76.6	**78.5**
+ ASFM	73.5	77.2	**79.4**	**75.0**	77.3	78.4	74.0	77.4	78.1
+ LFEM	72.8	76.2	77.8	73.8	76.5	77.7	74.2	76.8	78.2
EVHNet16	**73.6**	**77.4**	78.1	74.1	**77.8**	**78.2**	**74.7**	**77.6**	78.4

2.3.2　参数 K 的消融实验

本实验深入探讨了 Transformer 编码器的迭代次数 K 对实验结果的影响，并设计了一系列消融实验。值得注意的是，由于 Transformer 编码器本身的计算量较大，本文选择在 K=1，2，3，4，5 的范围内进行。

实验结果如表 5 所示。在这里，K 影响了不同网络层中不同尺度的特征，对网络性能产生影响。在综合考虑各哈希码长度的检索精度后，发现 K=4 时，实验结果达到了最佳。因此，在本文中选择 K=4 作为实验的最佳参数。

表 5　Food-101 数据集上参数 K 的消融实验结果

Table 5　Ablation experiment results of parameter K on Food-101 dataset

K	DPN- EVHNet16		
	16b	32b	64b
1	73.5	77.0	**78.4**
2	73.4	**77.8**	**78.4**
3	73.9	**77.8**	77.8
4	**74.7**	77.6	**78.4**
5	72.9	77.0	**78.4**

3　结论

本文针对食品图像细粒度和具有丰富语义信息的特点，提出了一种食品图像检索方法 EVHNet，该方法有效结合了卷积结构的局部特征提取能力和 Transformer 的全局表达能力，构建了基于增强 Vision Transformer 的哈希食品图像检索，在三个食品数据集上进行的相关研究验证了该方法的有效性。EVHNet 包含两个分支模块：聚合语义特征模块和局部特征增强模块。聚合语义特征模块从多层迭代的 Transformer 编码器中收集类令牌，收集的类令牌包含了食品图像中不同尺度的语义信息。局部特征增强模块对 Transformer 编码器的最后一层输出进行了局部特征的增强，使得网络能够学习食品图像中更具代表性的特征，并生成具有改进的局部特征表示。增强的局部模块使模型能够自动学习所关注的主要特征，同时抑制不重要的特征。在融合阶段，将局部特征和全局语义特征进行相互补充，从而增强最终的特征表示，增强的特征表示包含了食品图像中的细粒度特征以及更深层次的语义特征。研究发现，相比于纯 Transformer 结构或者纯 CNN 结构，混合架构兼顾局部特征和全局特征，在食品图像检索领域表现出良好的潜力。

参考文献：

[1] 李兆丰, 刘炎峻, 徐勇将, 等. 数字化食品在新时代下的发展与挑战[J]. 食品科学, 2022, 43(11): 1-8. DOI:10.7506//spkx1002-6630-20220324-292.

[2] 张南, 马春晖, 周晓丽, 等. 食品科学研究现状、热点与交叉学科竞争力的文献计量学分析[J]. 食品科学, 2017, 38(03): 310-315.

[3] MIN W, JIANG S, LIU L, et al. A Survey on Food Computing[J]. ACM Computing Surveys, 2019, 52(5): 1-36. DOI:10.1145/3329168.

[4] 梅舒欢, 闵巍庆, 刘林虎, 等. 基于 Faster R-CNN 的食品图像检索和分类[J]. 南京信息工程大学学报(自然科学版), 2017, 9(06): 635-641. DOI:10.13878/j.cnki.jnuist.2017.06.007.

[5] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149. DOI:10.1109/TPAMI.2016.2577031.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90. DOI:10.1145/3065386.

[7] SONG J, MIN W, LIU Y, et al. A Noise-robust Locality Transformer for Fine-grained Food Image Retrieval[C]//2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval. IEEE, 2022: 348-353. DOI:10.1109/MIPR54900.2022.00068.

[8] SONG J, LI Z, MIN W, et al. Towards Food Image Retrieval via Generalization-oriented Sampling and Loss Function Design[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(1): 13:1-13:19. DOI:10.1145/3600095.

[9] 张俊凯. 消费者对食品营养标签的使用行为及其影响因素[J]. 现代食品, 2017, (13): 64-66. DOI:10.16736/j.cnki.cn41-1434/ts.2017.13.024.

[10] ZHAO Q, WANG X, LYU S, et al. A feature consistency driven attention erasing network for fine-grained image retrieval[J]. Pattern Recognition, 2022, 128: 108618. DOI:10.1016/j.patcog.2022.108618.

[11] LUO X, CHEN C, ZHONG H, et al. Luo X, Wang H, Wu D, et al. A survey on deep hashing methods[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(1): 1-50. DOI:10.1145/3532624.

[12] LIU H, WANG R, SHAN S, et al. Deep Supervised Hashing for Fast Image Retrieval[C]//IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016: 2064--2072. DOI:10.1109/CVPR.2016.227.

[13] CAO Z, LONG M, WANG J, et al. Hashnet: Deep learning to hash by continuation[C]//Proceedings of the IEEE international conference on computer vision, 2017: 5608-5617. DOI:10.1109/ICCV.2017.598.

[14] SU S, ZHANG C, HAN K, et al. Greedy hash: towards fast optimization for accurate hash coding in CNN[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 806-815.

[15] ZHANG Z, ZOU Q, LIN Y, et al. Improved deep hashing with soft pairwise similarity for multi-label image retrieval[J]. IEEE Transactions on Multimedia, 2019, 22(2): 540-553. DOI:10.1109/TMM.2019.2929957.

[16] YUAN L, WANG T, ZHANG X, et al. Central similarity quantization for efficient image and video retrieval[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 3083-3092. DOI:10.1109/CVPR42600.2020.00315.

[17] FAN L, NG K, JU C, et al. Deep Polarized Network for Supervised Learning of Accurate Binary Hashing Codes[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020: 825-831. DOI:10.24963/IJCAI.2020/115.

[18] WANG J, ZHANG T, SONG J, et al. A survey on learning to hash[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 769-790. DOI:10.1109/TPAMI.2017.2699960.

[19] ZHUANG B, LIU J, PAN Z, et al. A survey on efficient training of transformers[C]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023: 6823-6831. DOI:10.24963/IJCAI.2023/764.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations, 2020.

[21] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778. DOI:10.1109/CVPR.2016.90.

[22] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 8759-8768. DOI:10.1109/CVPR.2018.00913.

[23] LI Y, HE J, ZHANG T, et al. Diverse part discovery: Occluded person re-identification with part-aware transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2898-2907. DOI:10.1109/CVPR46437.2021.00292.

[24] MIECH A, ALAYRAC J, LAPTEV I, et al. Thinking fast and slow: Efficient text-to-visual retrieval with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9826-9836. DOI:10.1109/CVPR46437.2021.00970.

[25] CHEN Y, ZHANG S, LIU F, et al. Transhash: Transformer-based hamming hashing for efficient image retrieval[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval, 2022: 127-136. DOI:10.1145/3512527.3531405.

[26] DUBEY S R, SINGH S K, CHU W T. Vision transformer hashing for image retrieval[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022: 1-6. DOI:10.1109/ICME52920.2022.9859900.

[27] BOSSARD L, GUILLAUMIN M, GOOL L V. Food-101–Mining Discriminative Components with Random Forests[J]. Springer International Publishing, 2014: 446-461. DOI:10.1007/978-3-319-10599-4_29.

[28] CHEN J, NGO C W. Deep-based ingredient recognition for cooking recipe retrieval[C]//Proceedings of the 24th ACM international conference on Multimedia, 2016: 32-41. DOI:10.1145/2964284.2964315.

[29] KAWANO Y, YANAI K. Kawano Y, Yanai K. Automatic expansion of a food image dataset leveraging existing categories with

domain adaptation[C]//Computer Vision-ECCV 2014 Workshops, 2015: 3-17. DOI:10.1007/978-3-319-16199-0_1.

[30] WU H, XIAO B, CODELLA N, et al. Cvt: Introducing convolutions to vision transformers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 22-31. DOI:10.1109/ICCV48922.2021.00009.

[31] GUO J, HAN K, WU H, et al. Cmt: Convolutional neural networks meet vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12165-12175. DOI:10.1109/CVPR52688.2022.01186.

[32] SONG C H, YOON J, CHOI S, et al. Boosting vision transformers for image retrieval[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 107-117. DOI:10.1109/WACV56688.2023.00019.

[33] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10778-10787. DOI:10.1109/CVPR42600.2020.01079.

[34] FANG J, SUN Y, ZHANG Q, et al. Densely connected search space for more flexible neural architecture search[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10625-10634. DOI:10.1109/CVPR42600.2020.01064.

[35] RU L, ZHENG H, ZHAN Y, et al. Token contrast for weakly-supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 3093-3102. DOI:10.1109/CVPR52729.2023.00302.

[36] HENDRYCKS D, GIMPEL K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units[J]. arXiv preprint arXiv:1606.08415, 2016.

[37] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision, 2018: 3-19. DOI:10.1007/978-3-030-01234-2_1.