DOI:10.13364/j.issn.1672-6510.20250004

基于跨模态对齐的食谱–图像检索研究综述

张贤坤，蒲 臻，夏志鸿
（天津科技大学人工智能学院，天津 300457）

摘 要：随着全球肥胖问题的日益严重，食物计算作为提升人类健康的重要研究方向，已成为多领域研究的热点。跨模态食谱检索作为食物计算与跨模态检索领域的交叉前沿，具有独特的研究价值。然而，由于食谱与图像之间存在显著语义鸿沟以及在食材种类、烹饪方法和文本描述等方面的复杂性，给跨模态食谱检索任务带来挑战。随着数据集规模的扩大和技术的发展，基于双编码器、生成对抗网络（GAN）、视觉语言预训练模型（VLP）的方法逐渐成为食谱检索领域的主流技术。本文综述了基于跨模态对齐的食谱–图像检索技术的最新进展，分析不同方法的优势与局限性，并对未来的发展方向进行展望。

关键词：食物计算；跨模态检索；食谱检索；视觉语言预训练
中图分类号：TP391.3 **文献标志码**：A **文章编号**：1672-6510（2025）05-0001-12

Review of Research on Recipe-Image Retrieval Based on Cross-Modal Alignment

ZHANG Xiankun，PU Zhen，XIA Zhihong
（College of Artificial Intelligence，Tianjin University of Science and Technology，Tianjin 300457，China）

Abstract：With the increasingly serious global obesity problem，food computing，as a pivotal research domain to improve human health，has gained prominence in many fields. As the intersection of food computation and cross-modal retrieval，cross-modal recipe retrieval has demonstrated significant research potential. However，cross-modal recipe retrieval faces challenges due to the significant semantic gap between recipes and images，along with the complexities of varied ingredients，cooking methods and textual descriptions. With the expansion of dataset scale and the development of technology，methods based on dual encoders，generative adversarial network（GAN）and visual language pre-trained model（VLP）have gradually become the mainstream techniques in recipe retrieval research. This article reviews the latest development of recipe image retrieval techniques based on cross-modal alignment，analyzes the strengths and limitations of different methods，and discusses future development directions.

Key words：food computing；cross-modal retrieval；recipe retrieval；visual language pre-training

引文格式：
张贤坤，蒲臻，夏志鸿. 基于跨模态对齐的食谱–图像检索研究综述[J]. 天津科技大学学报，2025，40（5）：1–12.
ZHANG X K，PU Z，XIA Z H. Review of research on recipe-image retrieval based on cross-modal alignment[J]. Journal of Tianjin university of science and technology，2025，40（5）：1–12.

近年来，随着全球肥胖问题的日益严重，健康饮食和营养管理成为社会关注的焦点[1]。食物计算（food computing）作为一个多学科交叉的研究领域，以计算机视觉、自然语言处理和机器学习为核心技术，为健康饮食和个性化营养推荐提供解决方案。跨模态食谱检索是食物计算和跨模态检索的重要交叉

收稿日期：2025–01–06；**修回日期**：2025–05–19
基金项目：国家自然科学基金资助项目（62377036）
作者简介：张贤坤（1970—），男，安徽人，教授，zhxkun@tust.edu.cn

领域，通过对食物图像与食谱文本的语义对齐和检索，实现食材识别、营养分析以及精准食谱推荐。

尽管跨模态检索技术在深度学习的推动下取得了长足进展，但是食谱–图像的语义鸿沟仍是该领域的一大难点。图像呈现的视觉信息与文本包含的语义信息存在显著差异，尤其在食材种类、烹饪方式和语言描述等方面存在多样性，这增加了语义对齐的复杂性。随着数据集规模的扩大和技术的演进，跨模态食谱检索方法逐渐从早期的双编码器架构，发展到基于生成对抗网络（GAN）的架构和基于视觉语言预训练模型（VLP）的方法，为图像与文本之间的深度语义关联提供了新的解决思路。然而，这些方法在细粒度对齐、模型效率以及大规模检索任务中仍然存在鲁棒性差等问题。

本文围绕跨模态食谱检索的研究进展，系统梳理现有方法的核心思路与技术特点，分析双编码器、GAN 和 VLP 等方法的优势与不足，结合任务的实际需求，探讨未来的研究方向与可能的技术突破，旨在为更高效、更精确的跨模态食谱检索提供参考。

1　食物计算与跨模态检索

据世界卫生组织统计，全球有超过 19 亿 18 岁及以上的人超重，其中超过 6.5 亿人肥胖，肥胖是多种慢性疾病的主要危险因素之一[1]。随着网络技术的发展和智能手机的普及，越来越多与食物相关的网站开始涌现，用户在这些网站上传、记录、分享或者收集与食物相关的数据，平台通过分析检索，为用户提供更合理的饮食规划，与食品相关的研究[2-3]受到广泛关注。随着人工智能技术的不断发展，2015 年，Harper 等[4]提出“食物计算”这一术语，2019 年正式定义[5]，并给出此研究领域的一般框架和任务。食物计算在人类健康和疾病预防方面具有重要的研究和应用价值，已成为多媒体、计算机视觉、医学和健康信息学等多个领域的研究热点。

食物计算[5]主要是指通过计算机视觉、自然语言处理、机器学习、数据采样以及其他先进技术，获取并分析不同模态的食物数据。大部分研究集中在食物图像的分类[6]、识别[7]、分割[8]、检索[9]和生成[10]，也有研究关注营养成分或热量预测[11]以及食物推荐系统[12]。此外，还有一些更贴近日常生活的课题，例如健康饮食计划[13]和食谱推荐[14]等。

跨模态检索是多模态研究中的重要课题，旨在从不同模态中搜索语义相关的数据，大规模数据集的出现[15]，使跨模态检索在近几年受到越来越多的关注。截至 2024 年 12 月，在 Web of Science 上以“ross-modal retrieval”为关键词的论文发文量在近 15 年总体呈逐年上升的趋势（图 1）。跨模态检索被应用的学科领域广泛，跨模态食谱检索作为食物计算与跨模态检索的交叉课题也具有重要的研究价值。食物图像有自己独特的属性，它没有任何独特的空间布局，且具有可变形的食物外观。此外，烹饪方法也会影响食品的外观，结合食谱文本的跨模态检索可以识别图像中未显示的特征，在此基础上可以进一步进行原材料识别、营养成分分析等相关工作。

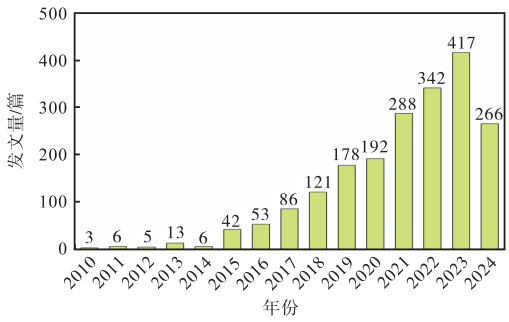

图 1　跨模态检索在 Web of Science 上论文数量的变化趋势

Fig. 1　Publication trend of cross-modal retrieval research in Web of Science

2　跨模态食谱检索面临的挑战

2.1　模态间的语义鸿沟

食谱文本和食物图像是常见的模态，而它们之间的语义对齐问题较为复杂。文本通常是对食谱的描述，包括食材、步骤、时间、烹饪技巧等；图像则呈现的是食物的外观，通常不包含详细的步骤和食材信息。因此，将图像中的视觉信息与文本中的语义信息对应，成为模态间对齐的基础。

图像虽然能够呈现食物的外观和某些特征，但是不能直接表达食材的具体成分、烹饪方法、步骤等信息。即使是相同的食物，图像也可能在不同的角度和光线下呈现不同的外观，这使从图像中提取的特征可能缺乏统一性。文本包含丰富的描述信息，文本中的某些信息可能在图像中没有明确呈现，甚至可能因为背景杂乱或视觉上过于复杂而被忽略。

2.2　数据的多样性与复杂性

食谱中涉及的食材种类非常丰富，且不同地区和不同文化背景下的食材名称和处理方式差异很大，这

使跨模态食谱检索在食材层面的语义理解和对齐变得复杂。同一种食材在不同的地区、语言甚至文化中可能有不同的名称，这种命名差异会影响食材的匹配。食材名称和烹饪术语中的同义词和多义词也使文本和图像对齐更加困难。

食谱中的烹饪方法和步骤往往会因文化、地域、饮食习惯等因素而有所不同。同一种食材可以使用不同的烹饪方式，甚至同一道菜肴的做法也会因家庭习惯或个人口味的不同而有所变化。“炒”“煎”“炸”“蒸”等虽然都是常见的烹饪方式，但是在具体实现上可能有细微差别，且不同地区或国家可能会有不同的名称或处理技巧。在文本中，烹饪步骤的描述往往是多样且有时是不精确的，即使是描述相同烹饪方法的步骤，表达方式也有很大差异。

3　跨模态食谱检索技术分析

深度学习的发展经历了多个重要阶段，其中卷积神经网络（CNN）是最早取得显著突破的模型之一，尤其在图像识别领域。随着时间的推移，循环神经网络（RNN）及其变体，如长短期记忆网络（LSTM）、门控循环单元（GRU）应运而生。2014 年，生成对抗网络（GAN）开启了深度生成模型的新时代。2017 年，Transformer 模型通过自注意力机制解决了 RNN 在处理长序列时有局限性的问题，显著提升了训练效果。此后，基于 Transformer 的预训练模型成为自然语言处理（NLP）任务的主流。跨模态食谱检索的研究[16]则是在深度学习技术兴起之后。

3.1　基于双编码器的方法

食谱的表示学习涉及捕捉食谱的文本和结构特征，通常由标题、成分和指令组成。许多工作将双编码器方法用于食谱图像检索任务，模型基本框架如图 2 所示，其中文本和图像由两个不同的编码器编码。大多数研究者只使用了成分和指令进行输入，早期使用 word2vec 进行词嵌入，使用 skip-thought 进行句子嵌入，再使用基于 LSTM、GRU 的架构获得最终的食谱嵌入。也有工作提出使用双向长短时记忆网络（Bi-LSTM）、双向门控循环单元（Bi-GRU）、层次 LSTM 和树状长短期记忆网络（Tree-LSTM）[17]可以更好地捕获成分和指令之间的依赖关系。图像处理部分大多采用 ResNet-50 模型[18]进行高效特征提取。此外，更多的改进是在损失函数部分，大多采用对比损失[9]或三元组损失[19]。

2017 年，Salvador 等[15]发布数据集 Recipe1M，为这项研究奠定了数据基础，同时提出一种跨模态食谱检索方法，使用配对的余弦损失函数和正则化损失函数学习不同模态数据的相似度。Adamine 模型[20]采用双重三元组学习方案，相比于余弦损失只关注匹配对的学习而忽视不匹配对的处理，双重三元组学习方案还可以同时关注不匹配对在高维公共空间的分布。MCEN 模型[21]在词级和句子级引入注意力机制，能够捕捉到食谱中不同部分的重要性，从而提取出更丰富、更具代表性的特征。HF-ICMA 模型[22]使用食谱内融合模块探索食材和指令之间的潜在关系，加强和丰富潜在信息的表达，在图像区域特征和食材特征之间引入多头交叉注意力机制，探索它们之间的潜在交互关系。

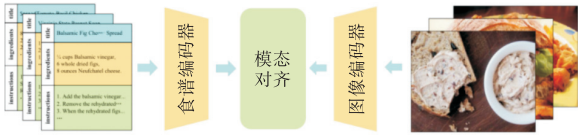

图 2　基于双编码器的模型基本框架

Fig. 2　Basic framework of model based on dual-encoder

MSJE 模型[23]将食物图像的类别语义作为食谱–图像关系语义的额外维度，利用食谱中主要成分的词频–逆文件频率（TF-IDF）语义[24]对学习到的特征进行增强。SEJE 模型[25]在食谱特征提取部分利用深度 NLP 模型、TextRank 和 TF-IDF 等技术识别和评分关键术语。Wang 等[17]通过 Recipe2tree 和 img2tree 模块学习结构化表示，在特定的树结构中对齐特征。SCAN 模型[26]使用 Kullback-Leibler（KL）散度衡量食物图像和食谱特征的语义概率分布之间的差异，通过对食谱文本的不同部分分配注意力权重，提升模态特征的区分性。

表 1 总结了基于双编码器跨模态食谱检索的代表性方法[15, 20-23, 25-28]。

3.2　基于生成对抗网络（GAN）的方法

生成对抗网络（GAN）是专门为了优化生成任务而提出的模型，其在食谱生成任务上也取得了良好表现。随着 GAN 突飞猛进的发展，图文检索工作也开始研究将 GAN 应用于该任务的可行性。

2019 年，Wang 等[29]提出 ACME 模型，将生成对抗网络引入跨模态食谱检索，使用食谱嵌入生成相应的食物图像，使用食物图像嵌入预测食谱成分进行模态对齐。不同于 ACME 模型使用 WGAN-GP 算法[30]

区分图像特征，Zhu 等[31]提出的 R2GAN 模型使用生成器和双判别器。判别器 D1 用于区分真实图像和生成的假图像，判别器 D2 用于区分从食谱嵌入生成的图像和从图像嵌入生成的图像，以确定模态的来源。该模型在生成图像时，同时考虑了在嵌入空间和图像空间中引入两层排序损失，有助于解释食谱的排名。CookGAN 食谱生成模型[32]利用 R2GAN 作为编码器提取食谱特征，通过图像生成中的因果链探索从文本到图像的合成问题。SGN 模型[33]考虑到食谱的层次结构，利用分层 ON-LSTM 网络[34]根据图像推断食谱树结构，利用 GAN 的思想增强模态一致性并生成食谱。X-MRS 模型[35]将食谱视为单个长的单词列表，使用单个 Transformer 编码器网络代替复杂的 RNN 组合，同时使用 StackGAN 网络架构[36]的简化版，去掉生成中间分辨率图像的步骤，而判别器不仅负责区分生成的图像和真实图像，还额外承担了食谱分类的任务。这增加了判别器的复杂性，但同时也提高了模型的实用性。RED-GAN 模型[37]对食物图像进行特征分解，在特征提取部分增加菜肴形状编码器，将图像的食谱信息和非食谱信息分开，生成器通过将这两类特征合并，与判别器一起进行对抗训练。李明等[38]提出在跨模态交互前利用对抗学习，从全局角度实现不同模态间的初步对齐，从而弥合因对不同模态独立编码导致的模态间差异。基于 GAN 的跨模态食谱检索的代表性方法见表 2。

表 1　基于双编码器跨模态食谱检索的代表性方法

Tab. 1　Representative methods of cross-modal recipe retrieval based on dual-encoder

模型	来源	模型结构			特点
		图像	食材	烹饪步骤	
JE[15]	CVPR'17	VGG-16/ResNet50	Bi-LSTM	分层 LSTM	引入语义正则化，通过共享高级分类权重，使食谱和图像嵌入在语义层面更好对齐
AdaMine[20]	SIGIR'18	ResNet50	Bi-LSTM	分层 LSTM	在训练阶段引入自适应挖掘方案，调整随机梯度下降的过程
MCEN[21]	CVPR'20	ResNet50	Bi-GRU	Bi-GRU	在训练阶段用潜在变量显式捕捉图像和文本间的交互，在推理阶段独立计算不同模态的嵌入
HF-ICMA[22]	SIGIR'21	ResNet50	Bi-GRU	sentence2vec	采用内模态和跨模态的注意力机制
MSJE[23]	TSC'21	ResNet50	Bi-LSTM	分层 LSTM	从食谱中提取 TF-IDF 特征，同时结合了图像的类别语义
SEJE[25]	TOIS'21	ResNet50	Bi-LSTM	分层 LSTM	改进批次难样本三元组损失函数，结合软边界和双重负采样
SCAN[26]	TMM'21	ResNet50	Bi-LSTM	分层 LSTM	结合语义一致性损失和注意力机制
CHEF[27]	AAAI'21	ResNet50	Tree-LSTM		对食材的重要性隐式评分，支持食谱修改
M-SIA[28]	CIBM'21	ResNet50	BERT+多头自注意力		在子空间级别上隐式对齐图像嵌入和食谱嵌入

表 2　基于 GAN 的跨模态食谱检索的代表性方法

Tab. 2　Representative methods of cross−modal recipe retrieval based on GAN

模型	来源	特征提取			生成对抗网络结构
		图像	食材	烹饪步骤	
ACME[29]	CVPR'19	ResNet50	LSTM	LSTM	没有生成器重建图像，而是从图像或文本字幕中生成特征，供判别器猜测模态的来源
R2GAN[31]	CVPR'19	ResNet50	Bi-LSTM	分层 LSTM	采用了一个生成器(G)和两个判别器(D1 和 D2)
X-MRS[35]	ACMMM'21	ResNet50	单个 Transformer 编码器		StackGAN 简化版，舍弃中间分辨率的图像，增加一个食谱分类器
RDE-GAN[37]	ACMMM'21	ResNet50	Bi-LSTM	分层 LSTM	生成器合并食谱特征和形状特征生成食物图像

3.3　基于视觉语言预训练(VLP)的方法

2017 年，谷歌公司提出 Transformer 模型[39]的基础框架，该模型处理长文本和图像的能力使其成为近年来非常受欢迎的架构。自 BERT(bidirectional encoder representations from Transformers)模型[40]在自然语言处理(NLP)领域问世以来，各种预训练模型如雨后春笋般出现在单模态领域，如计算机视觉(CV)领域的 Vision Transformer(ViT)[41]和语音领域的 Wav2Vec[42]，但多模态领域同时也存在高质量标注数据较少的问题。视觉语言预训练(VLP)主要通过对大规模数据进行预训练，学习不同模态之间的语义对应关系，然后在视觉问答(VQA)[43]、视觉语言推理(NLVR2)[44]、图像文本检索(TR)[45]和图像描述(image captioning)[46]等下游任务中应用。

3.3.1　基于预训练−微调的方法

先前的编码框架无法捕捉到成分和指令的相对重要性，Salvador 等[47]继发布 Recipe1M 数据集后，2021 年引入 Transformer 结构提出 H-T 模型，将这一研究领域带入新的发展阶段。采用 ViT 预训练模型[41]作为图像编码器，食谱文本部分采用层次化

Transformer 结构，同时提出将标题作为额外的输入，标题较为简短，只需经过一层 Transformer 处理，食材和烹饪方式以多个句子的列表形式呈现，则需经过两层 Transformer。这样模型能够更细致地理解和提取不同模态中的信息，提高检索性能。这一创新性想法的提出为之后的研究提供了新思路。层次化 Transformer 结构如图 3 所示，其中 T1、T2 表示两层 Transformer，FC 表示全连接层。T-Food 模型[48]同样使用该方法进行特征提取，提出多模态正则化的 Transformer 解码器，使用 CLIP-ViT 预训练模型[49]以增加模型鲁棒性。Yang 等[50]受 ACME 模型和 H-T 模型启发进行跨模态食谱检索研究，在实验中发现，将 CLIP-ViT 作为图像编码器主干网络时，模型的性能更佳，大批量训练可以显著提高模型在跨模态食谱检索任务中的性能。Papadopoulos 等[51]将食谱视为一个“程序”，烹饪步骤被视为一系列操作或变换，模型试图学习如何通过这些操作将原料转化为成品。这种方法为食谱的自动生成和理解提供了新的视角。VLPCook 模型[52]将现有的图像–文本对数据集转换为图像–结构化文本对数据集，使模型可以利用大规模的图像–结构化文本数据集进行预训练，在下游任务上对其进行微调。

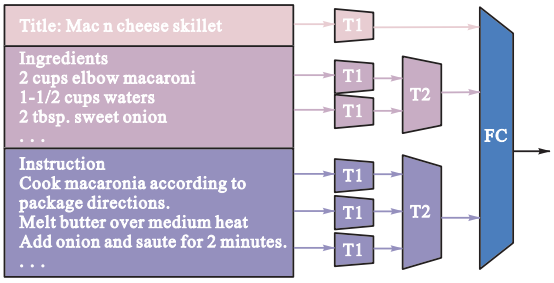

图 3　层次化 Transformer 结构

Fig. 3　Structure of hierarchical Transformer

随着检索技术的不断发展，研究者聚焦于图文细粒度匹配以提高检索精度。CREAMY 模型[53]从负样本中捕获互补信息，并提取所有样本对中的匹配信息，以减少匹配错误。FARM 模型[54]引入双曲线嵌入损失，双曲空间的负曲率特性非常适合表示食谱层次结构，可以使具有相似成分和技术的食谱靠近，具有不同元素的食谱远离，提高检索性能。MMACMR 模型[55]改进了层次化 Transformer 结构的食谱编码器，在两层 Transformer 之间加入了交叉注意力模块，提出多模态消歧和对齐策略加强细粒度语义对齐。

3.3.2　基于预训练–提示的方法

预训练–微调方法已经成为经典的预训练范式。随着预训练语言模型规模的不断增大，对其进行微调的硬件要求、数据需求和实际代价也在不断增加。此外，丰富多样的下游任务也使预训练–微调阶段的设计变得繁琐复杂，预训练–提示方法应运而生[56]。在多模态领域也出现了诸如 CLIP 模型[49]、CPT 模型[57]等代表性工作。

Sun 等[58]提出一个新的融合框架 PBLF 模型，将可迁移的视觉模型 CLIP 模型引入食谱检索任务中，针对食材和说明设计了两个不同的提示模板，有效缩小了预训练模型和下游任务之间的差距。Huang 等[59]为了使 VLP 模型能够更好地适应食谱检索任务，在 CLIP 预训练模型基础上提出 CIP 模型。CIP 模型针对图像特征构建细粒度提示，生成提示嵌入，并将其与食谱的组件嵌入进行对齐，从而实现跨模态的食谱检索，该模型能够学习结构化的食谱信息，在无需微调的情况下对视觉–文本表示进行对齐。该团队在此基础上引入基于证据深度学习的思想[60]，通过学习全局图像和食谱嵌入的证据，构建分类概率的狄利克雷分布，从而在全局视角下保持跨模态语义一致性。

3.4　大模型和轻量化模型在食谱检索中的探索

近年来，各种大语言模型（LLM）[61]层出不穷，许多研究团队将其引入各自研究领域，Song 等[62]利用大语言基础模型 Llama2 模型[63]将食谱扩充为视觉想象描述，聚焦视觉线索，利用视觉基础模型 SAM 模型[64]将食物图像扩充为捕捉关键成分的片段，并提出数据增强检索框架 DAR 协议。在实验阶段，采用 3 种评估协议衡量模型性能：DAR 协议基于原始的图像–食谱对检索评估；DAR+协议增加了由 Llama2 模型生成的视觉想象描述，通过结合图像与描述的距离提升检索的相关性；DAR++协议进一步整合了 SAM 模型产生的图像分割，这些分割对应于食谱中的关键食材，以提供更丰富的视觉信息。遗憾的是，查阅大量文献发现，目前尚缺乏通用多模态大模型在食谱检索中的系统性适配研究，但在食谱生成、食品图像识别等研究领域，部分研究者引入 LLaVA 模型[65]、OPT 模型[66]等大模型，取得了满意的效果。大模型强大的跨模态语义建模能力也为后续构建更通用、更精细的图文对齐框架提供了新思路。

与此同时，面对大模型计算开销大的问题，轻量化策略也成为跨模态检索模型落地的重要方向。IMHF 模型[67]使用基于 Transformer 模型的编码方式

统一处理视觉和文本特征，对二维图像进行切片并展平成序列，以简洁的方式处理二维图像。这种设计显著减少了模型的参数量和训练时间，同时在图像–食谱检索任务上取得了良好的性能。轻量级食谱检索框架 RecipeSnap 模型[68]使用 MobileNet-V2 作为图像编码器，参数量仅为 3.34 M，远低于 ResNet-50 模型的 23.5 M 和 ViT-B/16 模型的 86 M。RecipeSnap 模型通过优化设计，将内存和计算成本减少 90% 以上，在检索效率方面接近最先进模型的性能。部分研究已尝试通过蒸馏 CLIP 模型的图文嵌入小型双编码器网络[69]，在保持 70% 以上准确率的同时，减少了 70% 的计算量。模型结合 LoRA 等参数高效微调技术，也有助于降低大模型的训练与部署门槛，使其更适用于智能厨房终端或移动端环境。目前，这些轻量化技术在食谱检索领域的应用还较少，具有广阔的研究空间。

表 3 总结了基于 VLP 模型的跨模态食谱检索技术的代表方法。

表 3　基于 VLP 的跨模态食谱检索的代表方法

Tab. 3　Representative methods of cross-modal recipe retrieval based on VLP

模型名称	来源	模型结构				特点
		图像	标题	食材	烹饪步骤	
H-T[47]	CVPR'21	ResNet50/ ResNeXt101/ViT-B	一层 Transformer	两层 Transformer	两层 Transformer	基于分层 Transformer 的食谱编码器
LPR[51]	CVPR'22	ViT-B	Transformer			将食谱视为一个“程序”，模型试图学习如何通过操作将原料转化为成品
T-Food[48]	CVPRW'22	ViT-B/ CLIP-ViT	一层 Transformer	两层 Transformer	两层 Transformer	多模态正则化的 Transformer 解码器
VLPCook[52]	CVIU'23	ViT-B/ CLIP-ViT	一层 Transformer	两层 Transformer	两层 Transformer	将现有的图像–文本对数据集转换为图像–结构化文本对数据集进行预训练
TNLBT[50]	MMM'23	ViT-B/ CLIP-ViT	一层 Transformer	两层 Transformer	两层 Transformer	证明大批量训练可以显著提高模型在跨模态食谱检索任务中的性能
PFA[60]	TMM'24	ResNet50/ CLIP ViT/ViT-B	一层 Transformer	两层 Transformer	两层 Transformer	提示学习和基于证据深度学习
FARM[54]	WACV'24	CLIP ViT	两层 Transformer	三层 Transformer	三层 Transformer	双曲线嵌入损失表示食谱的层次结构
MMACMR[55]	FOOD'24	ViT-B	分层 Transformer 结构+交叉注意力模块			多模态消歧和对齐(MDA)

4　常用数据集和评价指标

4.1　数据集

4.1.1　Recipe1M 数据集

麻省理工学院于 2017 年公开发布 Recipe1M 数据集[15]，该数据集包含来自 24 个烹饪网站的 100 多万条烹饪食谱及 80 万张食物图像。Recipe1M 数据集的内容在逻辑上可以分为两层：第一层以自由文本的形式提供基本信息，包括标题、配料清单和准备菜肴的指令序列；第二层建立在第一层的基础上，包括任何与食谱相关的图像，这些图像以 JPEG 格式的 RGB 形式提供。食谱平均由 9 种食材和 10 个烹饪步骤组成，大约有一半的食谱包含完整制作后的食物图像。Recipe1M 数据集信息(表 4)包含大量图像–食谱对，在图像–食谱检索任务及模型泛化能力比较等实验中都展现出良好性能，为跨模态食谱检索研究提供了重要支撑。

4.1.2　Recipe1M+数据集

2019 年，经过进一步扩充和清洗 Recipe1M 数据集，最终得到数据集 Recipe1M+。该数据集包含 100 万份烹饪食谱和 1 300 万张食物图像[70-71]，数据结构更完整，在营养信息处理上更完善。Recipe1M 与 Recipe1M+数量差异见表 5。

表 4　Recipe1M 数据集信息

Tab. 4　Dataset specifications of Recipe1M

数据集	食谱/份	图像/张	图像–食谱/对
训练集	720 639	619 508	238 999
测试集	154 045	134 338	51 303
验证集	155 036	133 860	51 119

表 5　Recipe1M 与 Recipe1M+数量差异

Tab. 5　Different in quantitaty comparison between Recipe1M and Recipe1M+

数据集	Recipe1M & Recipe1M+	图像/张		
	食谱数量/份	Recipe1M	Recipe1M+	共享
训练集	720 639	619 508	9 727 961	493 339
验证集	155 036	133 860	1 918 890	107 708
测试集	154 045	134 338	2 088 828	115 373
总计	1 029 720	887 706	13 735 679	716 480

Recipe1M+数据集扩展了数据集的规模，提供更多种类的食谱及对应的复杂图像样本，有助于提高模

型的泛化能力和鲁棒性。除了基本的食谱描述和食品图像外，Recipe1M+数据集还包含更加细化的注释内容，有助于挖掘语义分析和模型理解的深度。Recipe1M+数据集在扩展数据集规模的同时，注重提高数据的质量，通过清理重复及低质量样本，确保数据的可用性和可靠性。

4.1.3　其他食谱数据集

在 Recipe1M 数据集出现前后，也有一些规模较小的数据集，均有其各自的特点。

Food-101 数据集[72]：流行的食物图像数据集，包含 101 个类别，共 101 000 张图像。每个类别包含 750 张训练图像和 250 张经过人工审核的测试图像。食谱数据部分是以原始的 HTML 格式呈现，需进行额外处理，主要应用于图像识别和分类任务。

VireoFood-172 数据集[73]：包含 110 241 张图像，标注了 353 个配料标签和 65 284 个配方，每个配方都有简要介绍、配料列表和制作步骤说明。该数据集只包含中国菜的食谱，为中文食谱检索提供可能。

Yummly-28K 数据集[74]：从 Yummly 网站爬取食谱数据，共得到 63 492 个食谱条目，经过预处理和划分后，最终包含 27 638 个食谱，每个食谱都包含 1 张食谱图像、配料、菜肴和课程信息，涵盖 16 种不同菜系以及 13 种食谱课程，适用于图像分类和推荐系统。

Go-Cooking 数据集[75]：来自“下厨房”网站的中文数据集，包含 61 139 对图像–食谱对。其中，54 139 对用于训练，2 000 对用于交叉验证，5 000 对用于测试。数据集涵盖不同种类的食物，每个食谱都包括配料清单和烹饪程序。

4.2　评价指标和性能比较

4.2.1　评价指标

在图像–文本跨模态检索任务中，召回率@K（R@K）是常用的性能衡量指标，用于评估检索系统从大规模数据集中检索与查询最相关项的能力。

R@K 的定义是在前 K 个检索结果中，与查询项正确匹配的项占所有正确匹配项的比例。具体到图像–文本跨模态检索中，这意味着检索系统在给定文本查询返回的前 K 张图像中，或在给定图像查询返回的前 K 个文本描述中准确的比例，K 的常见取值为 1、5、10，R@K 值越大，表明模型的检索效果越好。在实验阶段，通常采用 R@1、R@5 和 R@10 评估模型性能。此外，还可以将各项召回率相加，得到整体评价指标。

另一个常用指标是中位数排序（median rank，MedR），是指对于一批查询请求，在模型返回的检索结果中正样本所排位置的中位数。该指标数值越低，表明模型的检索准确率越高。

4.2.2　现有模型性能比较

对不同算法在 Recipe1M 数据集中食谱图像对的表现进行全面比较，其主要性能评价指标为 R@K 和 MedR。在测试集中分别取出 1 000 个（1 k）和 10 000 个（10 k）来自不同子集的食谱–食物样本对，将模态中的每一项视为一个查询，并根据查询嵌入和候选嵌入之间的欧氏距离对另一模态中的实例进行排序。通过欧氏距离进行检索，用标准度量评估跨模态检索任务的性能，对于之前采样的 1 k 和 10 k 测试子集计算 MedR 和 R@K，结果见表 6 和表 7。

由表 6、表 7 可以看出，图像–文本检索的召回率普遍比文本–图像检索的召回率高，这是因为图像所携带的信息更加丰富且多样，而文本则较为简洁、抽象，导致从文本到图像的检索面临更多挑战。在 10 k 测试样本上的总体表现不如在 1 k 测试样本上的表现，这是大规模检索任务中的噪声、样本多样性、标签不一致性等因素导致模型在面对更大、更复杂的检索空间时，无法像在小数据集上稳定地保持良好的性能。

基于双编码器的 HF-ICMA 模型和 SEJE 模型相较于其他模型表现良好，这说明交叉注意力机制的引入确实可以提高图文关联度。加入对抗训练机制后，在 1 k 和 10 k 的测试子集上 R@1 分别提升约 10% 和 30%，显著提升了检索精度。在引入预训练模型之后，模型在 1 k 测试子集上的 R@5 和 R@10 达 90% 以上，R@1 的提升不明显，但都显著优于普通方法的检索结果，验证了基于大规模数据预训练方法的有效性和重要性。大语言模型兴起后提出的 DAR++网络结构 1 k 上的 R@10 更是达到 97.9%，已接近最好性能，这说明用大语言模型对图像和文本特征进行增强后能更好地进行模态对齐。

综上所述，随着技术的不断更新，各种模型的效果有了显著提升，但在大规模检索任务中性能不稳定，图像–文本检索任务效果不优等问题还亟待解决。除了模型效果比较，模型的大小和复杂度也是研究者们关心的问题，在跨模态匹配任务中，文本编码器和图像编码器是参数量的主要来源，其参数量远远大于后续的对齐模块或损失函数部分。

表 6　在 Recipe1M 数据集上进行 1k 对比较

Tab. 6　1 k pairs comparison on Recipe1M datasets

方法类型	方法名称	来源	图像–食谱				食谱–图像			
			MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
基于双编码器	JE[15]	CVPR'17	5.2	24.0	51	65.0	5.1	25.0	52.0	65.0
	AdaMine[20]	SIGIR'18	2.0	40.2	68.1	78.7	2.0	39.8	69.0	77.4
	MCEN[21]	CVPR'19	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7
	HF-ICMA[22]	SIGIR'21	**1.0**	55.1	**86.7**	**92.4**	**1.0**	56.8	**87.5**	**93.0**
	SEJE[25]	TOIS'21	**1.0**	**58.1**	85.8	92.2	**1.0**	58.5	86.2	92.3
	SCAN[26]	TMM'22	**1.0**	54.0	81.7	88.8	**1.0**	54.9	81.9	89.0
基于 GAN 模型	ACME[29]	CVPR'19	**1.0**	51.8	80.2	87.5	**1.0**	52.8	80.2	87.6
	R2GAN[31]	CVPR'19	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3
	X-MRS[35]	ACMMM'21	**1.0**	64.0	**88.3**	**92.6**	**1.0**	**63.9**	**87.6**	**92.6**
	RDE-GAN[37]	ACMMM'21	**1.0**	**59.4**	81.0	87.4	**1.0**	61.2	81.0	87.2
基于 VLP 模型	H-T[47]	CVPR'21	**1.0**	60.0	87.6	92.9	**1.0**	60.3	87.6	93.2
	LPR[51]	CVPR'22	**1.0**	66.9	90.9	95.1	**1.0**	66.8	89.8	94.6
	T-Food(CLIP-ViT)[48]	CVPRW'22	**1.0**	72.3	90.7	93.4	**1.0**	72.6	90.6	93.4
	VLPCook(ViT)[52]	CVIU'23	**1.0**	73.6	90.5	93.3	**1.0**	74.7	90.7	93.2
	CIP[59]	ACMMM'23	**1.0**	77.1	94.2	97.2	**1.0**	77.3	94.4	97.0
	PFA[60]	TMM'24	**1.0**	72.9	94.0	96.9	**1.0**	72.6	93.9	96.8
	DAR+[62]	ECCV'24	**1.0**	76.9	94.9	97.4	**1.0**	**77.7**	**95.4**	**97.9**
	DAR++[62]	ECCV'24	**1.0**	**77.3**	**95.3**	**97.7**	**1.0**	77.1	**95.4**	**97.9**

注:数据加粗表示最优结果。

表 7　在 Recipe1M 数据集上进行 10 k 对比较

Tab. 7　10 k pairs comparison on Recipe1M datasets

方法类型	方法名称	来源	图像–食谱				食谱–图像			
			MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10
基于双编码器	JE[15]	CVPR'17	41.9	—	—	—	39.2	—	—	—
	AdaMine[20]	SIGIR'18	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
	MCEN[21]	CVPR'19	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
	HF-ICMA[22]	SIGIR'21	5.0	24.0	51.6	65.4	4.2	25.6	**54.8**	**67.3**
	SEJE[25]	TOIS'21	**4.2**	**26.9**	**54.0**	**65.6**	**4.0**	**27.2**	54.4	66.1
	SCAN[26]	TMM'22	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
基于 GAN 模型	ACME[29]	CVPR'19	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
	R2GAN[31]	CVPR'19	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
	X-MRS[35]	ACMMM'21	**3.0**	32.9	**60.6**	**71.2**	**3.0**	33.0	**60.4**	**70.7**
	RDE-GAN[37]	ACMMM'21	3.5	**36.0**	56.1	64.4	**3.0**	**38.2**	57.7	65.8
基于 VLP 模型	H-T[47]	CVPR'21	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
	T-Food(CLIP-ViT)[48]	CVPRW'22	**2.0**	43.4	70.7	79.7	**2.0**	44.6	71.2	79.7
	VLPCook(ViT)[52]	CVIU'23	**2.0**	45.3	72.4	80.8	**2.0**	46.4	73.1	80.9
	CIP[59]	ACMMM'23	**2.0**	44.9	72.8	82.0	**2.0**	45.2	73.0	81.8
	PFA[60]	TMM'24	**2.0**	41.9	71.2	80.9	**2.0**	41.7	70.9	80.7
	DAR+[62]	ECCV'24	**2.0**	47.4	75.3	83.8	**2.0**	**48.3**	**75.9**	**84.4**
	DAR++[62]	ECCV'24	**2.0**	**47.8**	**75.9**	**84.3**	**2.0**	47.4	75.5	84.1

注:数据加粗表示最优结果,—表示未提供结果。

表 8 进一步从模型参数量与计算效率两个维度对基于双编码器、基于 GAN 模型和基于 VLP 模型的不同类型模型进行比较。ViT-B/16 模型的参数量约为 86 M,是 ResNet-50 模型的 3 倍以上;而层次化 Transformer 文本编码器相比 Bi-LSTM 模型,也增加了近 10 倍的参数量。VLP 模型虽然性能最强,但训练与推理耗时显著,适用于资源充足的服务器。因此,从计算效率角度看,双编码器方法更适合对资源

敏感的部署需求，而 VLP 模型则适合性能优先的　　系统。

表 8　不同模型类型的参数量与计算效率对比

Tab. 8　Comparison of parameter size and computational efficiency across different model types

方法类型	典型模型	图像编码器	文本编码器	参数量估计	特点
基于双编码器	JE、AdaMine	ResNet-50（25 M）	Bi-LSTM（3 M）	约 30 M	结构简单，计算速度快，适合中小规模检索
基于 GAN 模型	ACME/R2GAN	ResNet-50（25 M）	Bi-LSTM+GAN 模块（10 M）	约 40 M	引入对抗训练，提升表示质量但训练较慢
基于 VLP 模型	H-T/VLPCook	ViT-B/CLIP-ViT（86 M）	层次化 Transformer（＞20 M）	＞100 M	表现最优但参数量大，计算成本高，不适合边缘部署

5　总结与展望

本文对基于深度学习的跨模态食谱检索技术进行综述，针对食谱–图像样本对从数据、模型、网络结构等方面总结跨模态对齐问题的解决方法。视觉语言预训练技术的提出以及 Transformer 架构的广泛应用，推动了该领域的进步，但仍面临诸多挑战。

由于食谱和图像的特殊性，食材在经过一系列加工后变成各种不同的形态，在设计模型时更应该考虑变换后食材和图像的细粒度对齐，现有的方法往往依赖手工设计的特征或预定义的对齐机制，可能无法充分捕获不同模态之间的复杂关系。后续研究需要考虑多模态数据固有的异质性和多变性，有效对齐和整合跨模态语义信息。

面对多模态大模型的高资源门槛，应探索参数高效微调方法，在不改变模型主体结构的前提下，实现任务定制化与部署优化。当前跨模态食谱检索研究多集中于性能提升，缺乏对模型体积、计算复杂度的考虑。未来可结合知识蒸馏、剪枝、量化等技术，构建具备端到端训练能力的轻量化模型，推动模型在移动设备、智能厨房等场景中的实际应用。

大语言模型掀起人工智能热潮，现有模型经过大量图像和文本数据的训练，能够捕捉通用的视觉和语言特征。为了在食谱检索任务中获得更好的效果，需要对模型进一步微调以适应食谱领域的研究需求。

当前跨模态食谱检索模型多侧重于图像与文本的模态对齐与语义表示，尚未充分引入营养学、食品科学等专业知识。在实际健康饮食推荐场景中，营养成分、食物功能性、过敏原风险等信息对用户决策尤为关键。未来研究可尝试将营养成分标签（如热量和脂肪、蛋白质、维生素的含量）作为辅助监督信息，融入跨模态对齐过程，增强模型对健康属性的感知能力。

参考文献：

［1］ 王佳仪. 2024 年《世界粮食安全和营养状况》报告发布［N］. 中国食品报，2024–07–31（003）.

［2］ OUYANG R，HUANG H，OU W，et al. Multimodal recipe recommendation with heterogeneous graph neural networks［J］. Electronics，2024，13（16）：3283.

［3］ ZHANG B，KYUTOBU H，DOMAN K，et al. Cross-modal recipe retrieval based on unified text encoder with fine-grained contrastive learning［J］. Knowledge-based systems，2024，305：112641.

［4］ HARPER C，SILLER M. OpenAG：a globally distributed network of food computing［J］. IEEE Pervasive computing，2015，14（4）：24–27.

［5］ MIN W，JIANG S，LIU L，et al. A survey on food computing［J］. ACM Computing surveys（CSUR），2019，52（5）：1–36.

［6］ KAREEM A S R，TILFORD T，STOYANOV S. Fine-grained food image classification and recipe extraction using a customized deep neural network and NLP［J］. Computers in biology and medicine，2024，175：108528.

［7］ BOYD L，NNAMOKO N，LOPES R. Fine-grained food image recognition：a study on optimising convolutional neural networks for improved performance［J］. Journal of imaging，2024，10（6）：126.

［8］ PHIPPITPHATPHASIT S，SURINTA O. Multi-layer adaptive spatial-temporal feature fusion network for efficient food image recognition［J］. Expert systems with applications，2024，255：124834.

［9］ ZHANG C，YANG Y，GUO J，et al. Improving text-image cross-modal retrieval with contrastive loss［J］. Multimedia systems，2022，29（2）：569–575.

［10］ KULLBACK S，LEIBLER R A. On information and sufficiency［J］. Annals of mathematical statistics，1951，22：79–86.

［11］ KADAM A，SHRIVASTAVA A，PAWAR S K，et al. Calories burned prediction using machine learning［C］//IEEE. 2023 6th international conference on contemporary computing and informatics（IC3I）. New York：IEEE，2023：1712–1717.

[12] ZIOUTOS K, KONDYLAKIS H, STEFANIDIS K. Healthy personalized recipe recommendations for weekly meal planning[J]. Computers, 2023, 13(1): 1.

[13] ZAYED A, PARTHASARATHI P, MORDIDO G, et al. Deep learning on a healthy data diet: finding important examples for fairness[EB/OL]. (2023-02-07)[2024-12-10]. https://arxiv.org/abs/2211.11109.

[14] ANDREA M G, KAREL G B, MARIA J. Link prediction in food heterogeneous graphs for personalised recipe recommendation based on user interactions and dietary restrictions[J]. Computing, 2023, 106(7): 2133-2155.

[15] SALVADOR A, HYNES N, AYTAR Y, et al. Learning cross-modal embeddings for cooking recipes and food images[C]//IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2017: 3068-3076.

[16] 蒋国芝, 左劼, 孙频捷. 跨模态检索在食物计算里的研究进展[J]. 现代计算机, 2021(3): 3-7.

[17] WANG H, LIN G S, HOI S C H, et al. Learning structural representations for recipe generation and food retrieval[J]. IEEE Transactions on pattern analysis and machine intelligence, 2021, 45: 3363-3377.

[18] SENAPATI B, TALBURT J R, BIN NAEEM A, et al. Transfer learning based models for food detection using ResNet-50[C]//IEEE. 2023 IEEE International Conference on Electro Information Technology. New York: IEEE, 2023: 224-229.

[19] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: improving visual-semantic embeddings with hard negatives[EB/OL]. (2017-07-18)[2024-12-10]. https://arxiv.org/abs/1707.05612.

[20] CARVALHO M, CADÈNE R, PICARD D, et al. Cross-modal retrieval in the cooking context: learning semantic text-image embeddings[C]//ACM. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. New York: Association for Computing Machinery, 2018: 35-44.

[21] FU H, WU R, LIU C, et al. MCEN: bridging cross-modal gap between cooking recipes and dish images with latent variable model[C]//IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 14558-14568.

[22] LI J, XU X, YU W, et al. Hybrid fusion with intra- and cross-modality attention for image-recipe retrieval[C]// ACM. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2021: 244-254.

[23] XIE Z, LIU L, WU Y, et al. Learning TFIDF enhanced joint embedding for recipe-image cross-modal retrieval service[J]. IEEE Transactions on Services Computing, 2022, 15(6): 3304-3316.

[24] ZHANG Z, WU Z, SHI Z. An improved algorithm of TFIDF combined with Naive Bayes[C]//ACM. Proceedings of the 2022 7th International Conference on Multimedia and Image Processing. New York: Association for Computing Machinery, 2022: 167-171.

[25] XIE Z, LIU L, WU Y, et al. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering[J]. ACM Transactions on information systems, 2021, 40(4): 74.

[26] WANG H, SAHOO D, LIU C, et al. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism[J]. IEEE Transactions on multimedia, 2022, 24: 2515-2525.

[27] PHAM H X, GUERRERO R, PAVLOVIC V, et al. CHEF: cross-modal hierarchical embeddings for food domain retrieval[J]. Proceedings of the AAAI Conference on artificial intelligence, 2021, 35(3): 2423-2430.

[28] LI L, LI M, ZAN Z, et al. Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images[C]//ACM. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2021: 3211-3215.

[29] WANG H, SAHOO D, LIU C, et al. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images[C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2019: 11564-11573.

[30] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein GANs[C]//CAI. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 5769-5779.

[31] ZHU B, NGO C W, CHEN J, et al. R^2GAN: cross-modal recipe retrieval with generative adversarial network [C]//IEEE. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach:

IEEE，2019：11469-11478.

[32] ZHU B，NGO C W. CookGAN：causality based text-to-image synthesis[C]//IEEE. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Seattle：IEEE，2020：5518-5526.

[33] WANG H，LIN G，HOI S C H，et al. Structure-aware generation network for recipe generation from images [EB/OL]. （2020-09-02）[2024-12-10]. https://arxiv.org/abs/2009.00944.

[34] SHEN Y，TAN S，SORDONI A，et al. Ordered neurons：integrating tree structures into recurrent neural networks [EB/OL]. （2019-05-08）[2024-12-10]. https://arxiv.org/abs/1810.09536.

[35] GUERRERO R，PHAM H X，PAVLOVIC V. Cross-modal retrieval and synthesis（X-MRS）：closing the modality gap in shared subspace learning[C]//ACM. Proceedings of the 29th ACM International Conference on Multimedia. New York：Association for Computing Machinery，2021：3192-3201.

[36] ZHANG H，XU T，LI H，et al. StackGAN：text to photo-realistic image synthesis with stacked generative adversarial networks[EB/OL]. （2017-08-05）[2024-12-10]. https://arxiv.org/abs/1612.03242.

[37] SUGIYAMA Y，YANAI K. Cross-modal recipe embeddings by disentangling recipe contents and dish styles [C]//ACM. Proceedings of the 29th ACM International Conference on Multimedia. New York：Association for Computing Machinery，2021：2501-2509.

[38] 李明，周栋，雷芳，等. 基于模态语义增强的跨模态食谱检索方法[J]. 计算机应用研究，2024，41（4）：1131-1137.

[39] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//CAI. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook：Curran Associates Inc.，2017：6000-6010.

[40] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional Transformer for language understanding[EB/OL]. （2019-05-24）[2024-12-10]. https://arxiv.org/abs/1810.04805.

[41] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16 × 16 words：transformer for image recognition at scale[EB/OL]. （2021-06-03）[2024-12-10]. https://arxiv.org/abs/2010.11929.

[42] SCHNEIDER S，BAEVSKI A，COLLOBERT R，et al. Wav2Vec：unsupervised pre-training for speech recognition[EB/OL]. （2019-09-11）[2024-12-10]. https://arxiv.org/abs/1904.05862.

[43] AGRAWAL A，LU J，ANTOL S，et al. VQA：visual question answering[J]. International journal of computer vision，2017，123（1）：4-31.

[44] SUHR A，ZHOU S，ZHANG A，et al. A corpus for reasoning about natural language grounded in photographs[C]//ACL. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Kerrville：ACL，2019：6418-6428.

[45] PLUMMER B A，WANG L，CERVANTES C M，et al. Flickr30k entities：collecting region-to-phrase correspondences for richer image-to-sentence models[J]. International journal of computer vision，2017，123：74-93.

[46] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft COCO：common objects in context[EB/OL]. （2015-02-21）[2024-12-10]. https://arxiv.org/abs/1405.0312.

[47] SALVADOR A，GUNDOGDU E，BAZZANI L，et al. Revamping cross-modal recipe retrieval with hierarchical Transformer and self-supervised learning[C]//IEEE. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. New York：IEEE，2021：15470-15479.

[48] SHUKOR M，COUAIRON G，GRECHK A，et al. Transformer decoders with multimodal regularization for cross-modal food retrieval[C]//IEEE. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops（CVPRW）. New York：IEEE，2022：4566-4577.

[49] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[EB/OL]. （2021-02-26）[2024-12-10]. https://arxiv.org/abs/2103.00020.

[50] YANG J，CHEN J，YANAI K. Transformer-based cross-modal recipe embeddings with large batch training [EB/OL]. （2022-12-16）[2024-12-10]. https://arxiv.org/abs/2205.04948.

[51] PAPADOPOULOS D P，MORA E，CHEPURKO N，et al. Learning program representations for food images and cooking recipes[EB/OL]. （2022-05-30）[2024-12-10]. https://arxiv.org/abs/2203.16071.

[52] SHUKOR M，THOME N，CORD M. Vision and structured-language pretraining for cross-modal food retrieval[J]. Computer vision and image understanding，

2024，247：104071.

[53] ZOU Z，ZHU X，ZHU Q，et al. CREAMY：cross-modal recipe retrieval by avoiding matching imperfectly[J]. IEEE Access，2024，12：33283–33295.

[54] WAHED M，ZHOU X，YU T，et al. Fine-grained alignment for cross-modal recipe retrieval[C]//IEEE. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. New York：IEEE，2024：5572–5581.

[55] ZOU Z，ZHU X，ZHU Q，et al. Disambiguity and alignment：an effective multi-modal alignment method for cross-modal recipe retrieval[J]. Foods，2024，13(11)：1628.

[56] 殷炯，张哲东，高宇涵，等. 视觉语言预训练综述[J]. 软件学报，2023，34(5)：2000–2023.

[57] YAO Y，ZHANG A，ZHANG Z，et al. CPT：colorful prompt tuning for pre-trained vision-language models[J]. AI Open，2024，5：30–38.

[58] SUN J，LI J. PBLF：prompt based learning framework for cross-modal recipe retrieval[C]//Artificial Intelligence and Robotics. Communications in Computer and Information Science. Singapore：Springer，2022：388–402.

[59] HUANG X，LIU J，ZHANG Z，et al. Improving cross-modal recipe retrieval with component-aware prompted CLIP embedding[C]//ACM. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa：Association for Computing Machinery，2023：529–537.

[60] HUANG X，LIU J，ZHANG Z，et al. Cross-modal recipe retrieval with fine-grained prompting alignment and evidential semantic consistency[J]. IEEE Transactions on Multimedia，2024，27：2783–2794.

[61] OPENAI，ACHIAM J，ADLER S，et al. GPT-4 technical report[EB/OL]. (2024–05–04)[2024–12–10]. https://arxiv.org/abs/2303.08774.

[62] SONG F，ZHU B，HAO Y，et al. Enhancing recipe retrieval with foundation models：a data augmentation perspective[EB/OL]. (2024–07–17)[2024–12–10]. https://arxiv.org/abs/2312.04763.

[63] TOUVRON H，MARTIN L，STONE K，et al. Llama 2：open foundation and fine-tuned chat models[EB/OL]. (2023–07–19)[2024–12–10]. https://arxiv.org/abs/2307.09288.

[64] MA J，HE Y，LI F，et al. Segment anything in medical images[J]. Nature communications，2024，15(1)：654.

[65] LIU G S，YIN H L，ZHU B，et al. Retrieval augmented recipe generation[EB/OL]. (2024–12–10)[2024–11–13]. https://arxiv.org/abs/2411.08715.

[66] LI P Y，HUANG X B，TIAN Y J，et al. ChefFusion：multimodal foundation model integrating recipe and food image generation[EB/OL]. (2024–12–10)[2024–09–18]. https://arxiv.org/abs/2409.12010.

[67] LI J，SUN J，XU X，et al. Cross-modal image-recipe retrieval via intra- and inter-modality hybrid fusion[C]//ACM. Proceedings of the 2021 International Conference on Multimedia Retrieval. New York：Association for Computing Machinery，2021：173–182.

[68] CHEN J F，YUE Y，XU Y F. RecipeSnap：a lightweight image to recipe model[EB/OL]. (2024–12–10)[2022–05–04]. https://arxiv.org/abs/2205.02141.

[69] WU K，PENG H W，ZHOU Z H，et al. TinyCLIP：CLIP distillation via affinity mimicking and weight inheritance[EB/OL]. (2023–09–21)[2024–12–10]. https://arxiv.org/abs/2309.12314.

[70] MARIN J，BISWAS A，OFLI F，et al. Recipe1M+：a dataset for learning cross-modal embeddings for cooking recipes and food images[J]. IEEE Transactions on pattern analysis and machine intelligence，2021，43(1)：187–203.

[71] 樵楠. 基于 Transformer 的食谱图文跨模态数据检索的研究[D]. 成都：电子科技大学，2024.

[72] WANG X，KUMAR D，THOME N，et al. Recipe recognition with large multimodal food dataset[C]//IEEE. 2015 IEEE International Conference on Multimedia & Expo Workshops. New York：IEEE，2015：1–6.

[73] CHEN J，NGO C. Deep-based ingredient recognition for cooking recipe retrieval[C]//ACM. Proceedings of the 24th ACM international conference on Multimedia. New York：ACM，2016：32–41.

[74] MIN W，JIANG S，SANG J，et al. Being a supercook：joint food attributes and multimodal content modeling for recipe retrieval and exploration[J]. IEEE Transactions on multimedia，2017，19(5)：1100–1113.

[75] CHEN J，PANG L，NGO C W. Cross-modal recipe retrieval：How to cook this dish?[C]//MultiMedia Modeling. International Conference on Multimedia Modeling. Cham：Springer International Publishing，2016：588–600.

责任编辑：郎婧