密级:_____

硕士学位论文

大规模食品图像数据集构建及识别方法研究

作者姓名：罗正东

指导教师：蒋树强 研究员 中国科学院计算技术研究所

闵巍庆 副研究员 中国科学院计算技术研究所

学位类别：工程硕士

学科专业：计算机技术

培养单位：中国科学院大学人工智能学院

2020 年 6 月

Research on large-scale food image dataset construction and recognition

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in Computer Technology

By

Zhengdong Luo

Supervisor: **Professor Jiang Shuqiang**

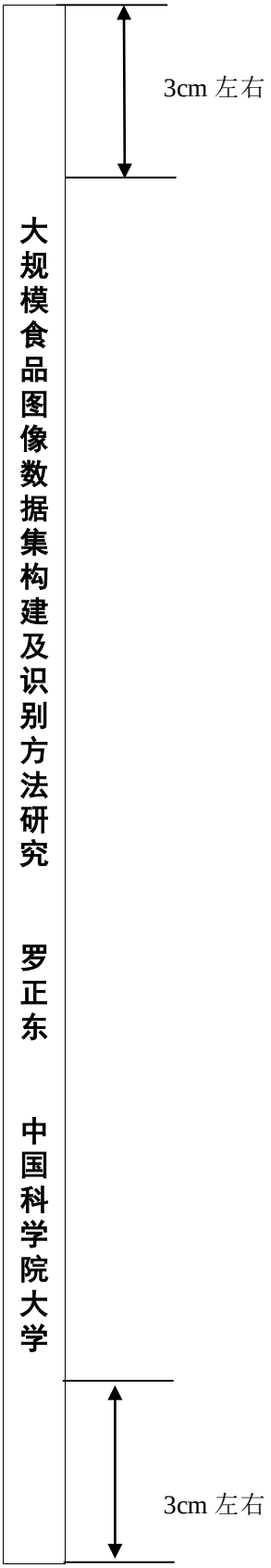
Associate Professor Min Weiqing

School of Artificial Intelligence

University of Chinese Academy of Sciences

June 2020

书脊（此页仅用于制作书脊，不用单独打印放入论文）



中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

随着生活质量提高，越来越多的人开始注重健康饮食，这就需要对食品有更加深入的了解，而食品识别技术可以在这方面更好地帮助人们。食品识别技术已开始在实际生活中应用，比如智慧餐厅、智能冰箱和膳食推荐等。然而，人工智能热潮背景下的食品识别技术仍然存在一些问题，比如食品图像识别准确率不高、识别的食品种类较少等。因此，如何改进大规模食品图像识别性能成为食品识别领域最关注的课题之一。

尽管很多研究人员已经在食品识别方面做了大量工作，但与实际应用的期望还有一定差距。造成差距的原因有很多，两个主要原因是：目前食品图像数据集规模较小；尚无完全有效适合食品识别的方法。针对以上两个问题，本文从数据集构建和识别方法两个方面进行了研究和探讨，主要研究内容和贡献如下：

（1）构建了一个大规模食品图像数据集：数据集是各种技术方法的重要基础，一个高质量数据集能够为方法有效性验证提供关键保障。考虑到现有食品图像数据集规模较小，本工作从食品种类概念体系构建、数据收集、数据清洗到扩充和验证分析数据，系统化地构思和构建了一个大规模食品图像数据集 ISIA Food-500，这个数据集包含来自 500 类食品的 405,776 张图片，食品种类覆盖各大洲共 52 个国家。现有基准食品图像数据集包含食品种类和图片量较少，相比之下，ISIA Food-500 数据集在图片量、种类数和可拓展性等方面具有较大优势。

（2）提出了一种基于堆叠多尺度多注意力网络（Stacked Multi-Scale Multi-Attention Network，SMSMANet）的识别方法：考虑到现实很多情况下，人们偏向于视觉的食品图像识别，而且食品图像类间差异小、类内差异大等属性也需要一个既关注整体全局特征又注重细节局部特征的食品识别方法。本工作提出一种堆叠多尺度多注意力网络来进行食品识别，该方法以图像模态为输入，利用空间和通道注意力机制，结合多尺度策略，共同学习图像的全局特征和局部特征进行食品识别。作者通过大量实验证明了堆叠多尺度多注意力网络的有效性。

关键词：食品图像数据集，食品识别，多尺度，多注意力

Abstract

With the improved quality of life, more and more people are beginning to focus on healthy eating, this requires a deeper understanding of food, and food recognition technology can better help people in this regard. Food recognition technology has begun to be applied in real life, such as smart restaurant, smart refrigerator, and diet recommendation. However, under the background of artificial intelligence upsurge, there are some problems in food recognition technology, such as the accuracy of food recognition and the limited diversity of food categories. Therefore, improving large-scale food image recognition performance are desperately needed nowadays in the field of food recognition.

Even though researchers have done a lot of works on food recognition, there is still some gap with the expectations of practical applications. The main reasons are: the scale of food image dataset is small, and there is no fully effective method for food recognition. Aiming to alleviate the above issues, this thesis has studied and discussed this topic from dataset construction and recognition methods. The main research contents and contributions are as follows:

(1) This thesis constructs a large-scale food image dataset. Datasets are an important basis for various methods. A high-quality dataset can provide key guarantees for the effective verification of methods. Considering the small-scale of the existing food image datasets, this work systematically conceives and constructs a large-scale food image dataset ISIA Food-500 from constructing food categories system, collecting data, cleaning data to expanding and verifying analysis data. This dataset contains 405,776 images and 500 food categories from 52 countries. Compared with existing popular benchmark food datasets, the ISIA Food-500 is a more comprehensive food dataset with larger category coverage, larger data volume and higher diversity.

(2) This thesis proposes a food recognition method based on Stacked Multi-Scale Multi-Attention Network (SMSMANet). In most existing works, people tend to use visual food image for food recognition. However, the attributes that have small inter-class differences and large intra-class differences could provide complementary information but still lack exploration such as discriminative global appearance and local details. This work proposes a Stacked Multi-Scale Multi-Attention Network to jointly learn image-oriented global and local features via combining hybrid spatial-channel

attention and multi-scale strategy for food recognition. The effectiveness of the Stacked Multi-Scale Multi-Attention Network is proved by sufficient experiments.

Key Words: Food image dataset, Food recognition, Multi-scale, Multi-attention

目 录

图目录

表目录

第 1 章 绪论

饮食是人类重要且必不可少的活动之一，现实中对食品识别技术的需求促使食品识别成为学术和产业界的一个新兴研究领域。本章将介绍食品识别研究背景与意义、国内外研究现状和本文主要研究内容，以便大家充分理解本工作意义，从整体上把握本工作概况。

1.1 研究背景与意义

1.1.1 研究背景

人们对饮食的研究由来已久，比如在营养学和食品学等领域就有专门系统的食品资料，但是那些专业性文本记录或其他形式存放的资料并不能广泛应用于民间生活。大多数人希望随时随地方便有效地进行食品识别，而不是经常查阅资料才能知道某种食品是否适合自己等，而且就算有丰富的资料，人们在对资料进行理解的时候也可能有偏差。再者，这些资料总结的是大多数人的规律，个性化需求需要一个智能化食品系统来服务每个人。除了个人服务需求，其他领域对食品识别技术也有广泛需求。在食品生产工厂里，食品识别技术可以用来分类和包装食品；在智慧餐厅里，食品识别技术帮助机器人进行智能化传菜；在食品安全方面，食品识别技术可以帮助质检人员甄别食品种类和质量；在智慧家居中，食品识别技术用于智能冰箱来监控食品新鲜度等等。由此可见，食品识别已成为实际生活中亟需发展的一项技术，具有重要潜在的实际应用价值。

食品计算[1]是一个新兴研究领域，它探讨与食品相关的检测、识别、检索和推荐等问题。在食品计算领域中，一个重要研究方向就是食品识别。食品识别是通过机器学习训练，让计算机识别出图片、视频和文本等所展示食品的类别。而本文中所说的食品识别，是当今比较流行的计算机视觉中图像识别概念，即以食品图像和类标签（食品类名）作为机器学习模型（比如深度学习模型）的主要输入，训练调试好模型之后，输入一张待测试食品图像，让模型预测出该图像的类标签。食品识别在计算机视觉领域中已经受到越来越多的关注[2][3][4][5]，这主要是因为它多种多样的应用和服务价值，比如可视化个人食品日记[3]、健康膳食推荐[6]和自助智慧餐厅[7]等。近些年，英国、日本和中国等多国科研人员在多个

顶级计算机会议（比如 ACM Multimedia 和 International Joint Conferences on Artificial Intelligence 等）和期刊（比如 ACM Computing Surveys 和 IEEE Trans. on Image Processing 等）上都发布了许多文献和食品图像数据集。由此可见，食品识别在科研学术界是一个热门研究方向。

尽管食品识别技术有很大应用潜力，但是目前从图片中识别出食品类别仍然是一个极具挑战的工作。这些挑战主要有以下两个方面：

（1）在数据集方面，目前缺乏一个大规模食品图像数据集。现有工作主要集中在利用小规模数据集进行食品识别，公认的两个基准食品数据集是 ETHZ Food-101[2]和 Vireo Food-172[8]。Lukas Bossard 等人[2] 构建了 ETHZ Food-101 食品图像数据集来进行食品识别，这是一个公开供免费使用的食品图像数据集，包含 101 类食品的 101,000 张图像，这个数据集的食品种类主要是西方菜品。Jingjing Chen 等人[8]构建了 Vireo Food-172 食品图像数据集，这个数据集包含来自中国 172 种食品的 110,241 张图像。然而，这两个主流的基准食品数据集包含的食品种类较少，图片总量只有十万左右，目前缺少一个图片量和食品种类都比较多的食品图像数据集。这两个基准数据集一个是西方食品为主（ETHZ Food-101），一个是中国菜品为主（Vireo Food-172），它们的食品种类覆盖范围有限，缺少一个同时包含中西方食品的高覆盖率图像数据集，现有基准食品图像数据集不能很好地满足构建复杂食品统计模型需求。提高食品图像识别性能最准确可靠的方法之一就是利用监督学习的方法来实现，监督学习训练需要包含足够多食品种类的数据集，而现今缺乏大规模食品图像数据集限制了监督学习下的食品识别技术发展，也就无法很好地将已有的各种食品识别方法完美地应用于实际食品识别系统中去。总之，一个大规模食品图像的出现必将推动食品识别技术的发展和应用，也将为食品计算研究带来诸多益处。因此，食品数据集构建是一个具有重要基础性价值的研究工作。

（2）从方法层面说，食品图像在视觉上呈现出复杂性。成千上万的食品种类中，有些食品图像存在很明显视觉差别，比如面条和米饭。有些食品图像具有类间差异小、类内差异大和非刚性结构等属性，这些属性具有细粒度图像特点。详细地解释其视觉上的细粒度特性：首先，很多食品图像类内差异比较大，比如图 1.1 中的（a）组图片，这两张图片展示的都是炸鸡翅这道菜，但是它们形状上差别却很大，而不像一般实物图像数据集，同类图片在形状上呈现出高度一致性，

比如 ImageNet[9]的猫类和狗类等都各自呈现出相同的体型模样；其次，有些食品图像类内差异性较小，比如图 1.1 中（b）组图片，烤猪排和烤羊排看起来是一样的，连人眼都很难分辨。若用计算机从视觉角度去识别，也是会有一定难度；最后，很多食品图像是非刚性结构，即不存在固定的形状，典型例子如汤类，所呈现形状是根据容器形状而变化的，这种食品它本身没有形状。图 1.1.中（c）组酸辣汤图片，充分展示了这类食品的非刚性结构特点。总之，食品图像存在视觉上呈现出复杂性，如何结合食品图像特性充分利用食品图像的全局特征和局部特征得到判别性图像表示，这是一个有挑战性的研究工作。

(a) 炸鸡翅　(b) 上图：烤猪排；下图：烤羊排　(c) 酸辣汤

图 1.1　食品图像的视觉特性

Figure 1.1 Visual characteristics of food images

1.1.2 研究意义

食品为人们进行其他一切活动提供身体能量来源，食品识别技术在个人和社会生活中有着广泛的需求，比如食品生产、食品安全和个人饮食管理等。然而，现今的食品识别技术并没有大范围使用，其中一个原因是食品识别技术尚未完全成熟，比如识别准确性不好，识别种类较少等。虽然很多研究人员已经做了大量工作改善食品识别技术，但是研究成果与实际应用之间还有一定差距。因此，本文从食品数据集构建和识别方法研究两方面展开工作。首先，本文构建了一个大规模食品图像数据集 ISIA Food-500，包含更多食品种类和图片量，与现有食品数据集共同加强食品识别领域的数据基础，从而为食品识别技术发展提供数据支持。再者，本文提出了 Stacked Multi-Scale Multi-Attention Network 方案，以食品

图片为输入，联合学习全局特征和局部特征来进行有效地食品识别。本工作丰富了食品图像识别的方法，为实际应用提供多一种选择方案。

食品识别技术研究是一项既有实际需求价值也有重要科研价值的工作。本工作中食品图像数据集构建具有重要基础性价值，研究食品图像属性的识别方法创新具有一定挑战性。在后续部分将详细介绍本文提出的数据集和新方法。

1.2 国内外食品识别研究进展

1.2.1 现有食品图像数据集介绍

表达食品的方式有很多种，比如文字菜谱、视频记录和食品图片等，食品计算是利用各种模态的食品描述形式来进行食品内容分析和理解等研究。本工作重点探讨的食品图像数据是食品计算研究的重要资源之一，目前很多食品计算工作是在食品图像数据基础上进行实验建模：文献[10]结合食品图像和其他各种食品上下文领域知识来建立个性化食品计算模型研究食品推荐系统，可以根据食品特性推荐相应食品给特定人群；文献[11]结合食品图像和菜谱配方等信息构建了一个检索食品图片和分析菜谱模式的跨模态食谱分析模型；文献[12]利用食品图像、餐厅信息和位置信息构建了一个概率模型来进行食品识别。由此可见，食品图像为食品识别、食品推荐和食品检索等食品计算研究提供不可或缺的数据支持，其重要性好比是大厦之根基。

近来年，越来越多的食品图像数据集已经被发布[2][8][13][14][15][16]。食品图像数据集PFID数据集[17]是由Mei Chen等人在2009年提出，这是一个快餐类食品图像数据集，包含101类快餐食品的4,545张图片，食品数据量较少。之后，食品图像数据集规模逐步增大。在2014年时，Lukas Bossard 等人[2]公布了ETHZ Food-101食品图像数据集，这个数据集的食品种类主要来自西方，包括101类食品的101,000张图片，每类均衡地包含1,000张图像，这是一个典型用于食品识别的数据集，这个数据集是目前食品识别领域公认度比较高的一个基准食品图像数据集，但是它有只覆盖了西方食品、种类较少和数据量较少等问题，对未来食品识别研究起到的数据支持作用有限。在2016年时，Jingjing Chen等人[8]公布了Vireo Food-172食品图像数据集，这个数据集的食品主要来自中国，包括172类食品的110,241张食品图片。这个数据集每类数量是非均衡的，单类图片数量从191~1,061张不等，与ETHZ Food-101数据集一样，这也是一个用于食品识别的基

准数据集。以西方食品为主的ETHZ Food-101和以中国菜为主的Vireo Food-172是现在食品识别研究领域公认度比较高的两个基准食品数据集。Yoshiyuki Kawano等人[14]发布了一个以日本食品为主的数据集UEC Food256，这个数据集包含256类日本食品25,088图片。近些年，一直又有新的食品图像数据集不断发布。比如在2017年，Kaggle与Computer Vision and Pattern Recognition (CVPR)联合举办的Fine-Grained Visual Categorization (FGVC)挑战赛发布了FoodX-251[15]食品图像数据集，这个数据集包含251类食品158,846张图片。然而，这些数据集的种类覆盖范围和图片量都比较少，无法很好的满足未来食品识别需求。还有一些食品集包含多种模态食品信息（这里的多模态概念是指食品图像和食材文本形式等），比如Recipe1M食品数据集[18][19]和Instagram800K食品数据集[20]。其中比较出名的是Recipe1M食品数据集，它包含约100万个结构化烹饪食谱及其图片，用于在对齐多模态数据上训练大容量模型，从而进行跨模态检索[19][21]。相比之下，本文的ISIA Food-500数据集旨在研究视觉上的食品识别。ISIA Food-500食品图像数据集和Recipe1M等多模态食品数据集具有一定数据互补性，可以共同促进食品识别领域智能分析与理解的发展。此外，Saihui Hou等人[22]提出的Vegfru数据集是果蔬类食品数据集，它包含292类果蔬160,000张图片，这个数据集在食品种类上稍微不同于以菜品和快餐等熟食为主的食品数据集，扩充了食品数据集的类型。

然而，现有数据集的共有问题是：（1）无明确食品种类概念体系。这些数据集的构建过程里没有体现出完整食品种类概念体系，论文或者网站中没有提到对数据集构建时种类概念体系分析，只是单纯的把各种食品种类放在一起构建，看不出是按照何种概念体系来逐步构建的。而本工作则从构建食品种类概念体系、收集数据、清洗数据和检查数据来逐步构建新数据集。（2）数据集规模较小。现有食品数据集的食品种类和图片量都比较少，比如两个主流数据集（ETHZ Food-101和Vireo Food-172）的食品种类数不超过200类，图片量不超过12万。（3）食品种类地域覆盖性较低。很多现有食品数据集专注于某国食品，这种数据集能很好的适用于训练某国食品识别模型，但是未必能学到世界多样食品数据的模式，模型应用范围十分有限。

本文推出的大规模食品图像数据集ISIA Food-500包含500种食品的405,776张图像，食品种类来自于全球52个国家。ISIA Food-500数据集在图片量、种类数

和可拓展性等方面具有较大优势，与现有食品图像数据集的比较分析见第二章。

1.2.2 食品分类标准现状

维基百科网站包含丰富的食品资料，在维基百科食品列表[23]中有四种列表：基于食品原材料的“Lists of foods by ingredient”、基于食品所属国家的“Lists of foods by nationality”、基于食品类型的“Lists of foods by type”和基于食品品牌的“Lists of brand name foods”。这四种列表包含的食品种类数量不等，从多到少排序是：基于食品原材料的种类（包含 71 类食品）、基于食品类型的种类（包含 61 类食品）、基于食品所属国家的种类（55 类食品）、基于食品品牌的种类（26 类食品）。四种食品列表中的食品种类有重复但是彼此之间种类数不等，原因是有些食品无法分辨所属某国或某类型或某品牌。本工作计划构建一个大规模食品图像数据集，综合比较后选择基于食品原材料的列表。在基于食品原材料的食品列表中，食品名称按照英文首字母顺序排列，每类食品有丰富的文字和图片资料，主要包括食品外形、主要原材料、所属国家和食品图片等信息。维基百科的食品资料为本工作食品数据集构建提供了不可或缺的知识。

另外，作者对国内外食品分类标准进行了广泛调研。James Ireland 等人[24]针对欧洲食品提出从食品营养和能量角度构建食品分类系统。中国绿色食品发展中心的陈倩等人[25]在 2017 年的中国食物与营业期刊上发表了论文简要介绍了国内外食品分类标准，其中提到几个世界权威的分类标准的用途，比如国际食品法典委员会的《食品与饲料分类标准》（CAC）[26]主要作用是为农药残留限量标准制定参考标准。日本的“肯定列表制度”[27]也是为食品中农业化学品残留管理而制定的一项标准。在中国，两个主要的食品分类标准《食品安全国家标准　食品添加剂使用标准》（GB　2760-2014）[28]和《食品安全国家标准　食品中农药最大残留限量》（GB 2763-2014）[29]，也是关于添加剂和农药剂量的食品安全标准。

作者发现大多数的国际食品分类标准大多是为食品安全服务，依据食品添加剂残留量来制定分类标准。虽然国际食品及农业组织的《食品添加剂通用法典标准》（GENERAL STANDARD FOR FOOD ADDITIVES, GSFA）[30]也主要用于规定食品添加剂使用剂量和使用范围，但是它是根据食品大类来制定标准，其中

对各种食品分类进行了详尽描述，每个分类都有细致种类注解和主要食品的举例，这为本文构建食品种类概念提供了参考。

1.2.3 食品识别方法研究现状

食品识别研究已开展多年，涌现出越来越多的方法在一定程度上促进了食品识别技术发展。Weiqing Min 等人[1]发表了一篇食品计算综述，里面介绍了食品计算问题由来，其中提到食品识别是食品计算的研究方向之一。总体来说，智能化食品识别方法大致可以划分为：传统方法和深度学习方法。

（1）传统方法：

早些年，各种手工设计的特征被用来进行食品识别[2][31]。在 2010 年，Shulin Yang 等人[31]就提出了用统计方法来对食品图像的特征表示进行计算，从而实现食品识别。Lukas Bossard 等人[2]利用随机森林算法找到判别性图像局部区域，然后使用判别性特征表示来进行食品识别。Taichi Joutou 等人[32]利用颜色直方图方法来进行食品识别。

（2）深度学习方法：

（a）深度因其对高级特征强大的学习能力而备受关注，现在流行的食品识别方法多是基于深度学习框架[33][34][35]。在 2014 年，Hokuto Kagaya 等人[36]采用 AlexNet 网络提取食品图像的深层特征表示来进行食品识别。在 2016 年，Ashutosh Singla 等人[37]采用 GoogLeNet 网络对食品图像提取更深层次的表示，实现了食品和非食品图像的分类。这些方法是对图像全局视觉信息进行特征提取，没有充分考虑图像的非食品背景部分对分类性能的影响。为此，有人开始利用基于卷积神经网络（CNN）物体检测的方法提取图像的候选区域，然后对候选区域进行特征提取并经过多种特征融合生成图像的统一特征表示。例如在 2017 年时，梅舒欢等人[38]就是利用 Faster R-CNN 网络检测食品图像的候选食品区域，然后采用 CNN 的深度神经网络提取候选区域的视觉特征表示，最后利用这些特征表示进行食品图像分类。与找到食品图像候选区域的原理类似，在 2018 年，Niki Martinel 等人[34]对具有垂直结构特性的食品图像作切片来提取局部特征，然后融合这些局部特征得到全局特征表示来进行食品识别。

（b）另一类基于深度学习方法的食品图像识别是结合食品原材料的文本信息、食品图像的视觉信息及烹饪属性等构建一个多模态网络结构来进行食品识别。

它们利用附加的上下文信息来进行食品识别，比如文献[4][16]利用原材料信息和位置信息来提高食品识别准确率，Feng Zhou 等人[4]通过二分图探究原材料和餐馆之间丰富的信息来进行食品识别。最近，Weiqing Min 等人[16]利用丰富的原材料成分信息作为附加的监督信号，对多个不同尺度食品图像区域进行定位，并将这些区域特征融合到统一特征表示中来进行食品识别。后来，他们[5]用流行的深度学习模型融合高级语义特征、中级属性特征和抽象的深层视觉特征来进行食品识别，在两个公开的基准食品图像数据集 ETHZ Food-101 和 Vireo Food-172 上达到了目前最好识别性能。然而，当附加属性信息（原材料信息和地理信息等）不可用时，这些方法可能无法对食品识别起到很好作用。所以，本文提出一种以图像模态为输入的食品识别方法 SMSMANet 网络，本工作在深度学习框架下通过空间和通道注意力机制，结合多尺度策略，融合全局和局部特征来进行食品识别。后续章节会详细介绍本文的方法框架。

1.3 本文主要研究内容

本文主要研究食品识别领域相关问题，并希望这项工作能为食品计算发展做出一定的推进，从而更好地让智能化食品计算技术应用服务于社会。本文主要研究内容是围绕食品识别研究课题，构建了一个大规模食品图像数据集，并提出了一种食品识别方法：

（1）本工作构建了一个新的大规模食品图像数据集 ISIA Food-500 用于食品识别，它包含来自 500 类食品的大约 40 万张图像。这将是迄今最大可公开获取的食品图像数据集，能够为训练高容量食品识别模型提供帮助。

（2）本工作提出了一个堆叠多尺度多注意力网络（SMSMANet）共同学习全局特征和局部特征进行食品图像识别。此方法在基准数据集 ETHZ Food-101 和 Vireo Food-172 验证有效，同时在 ISIA Food-500 数据集上达到最好性能。

本文组织结构：

第 1 章 绪论。本章简要介绍本论文问题背景、当前相关领域研究现状和本文主要研究内容。

第 2 章 ISIA Food-500 食品图像数据集构建和分析。本章首先介绍 ISIA Food-500 数据集构建过程，从食品概念体系构建、数据获取，再到后来数据清洗及扩

充数据，详细介绍每一步数据集构建工作。接着对 ISIA Food-500 数据集进行多方面统计分析，展示了 ISIA Food-500 数据集的优势和特点。

第 3 章 基于 SMSMANet 网络的食品识别。本章主要介绍本文提出的新方法框架 SMSMANet 网络，包括全局特征学习子网络和局部特征学习子网络的详细介绍。

第 4 章 实验及分析。本章主要介绍实验设置以及在几个数据集上进行了 SMSMANet 网络和其他食品识别方法的实验评估。

第 5 章 结论与展望。本章主要介绍了本文研究结论及未来工作展望。

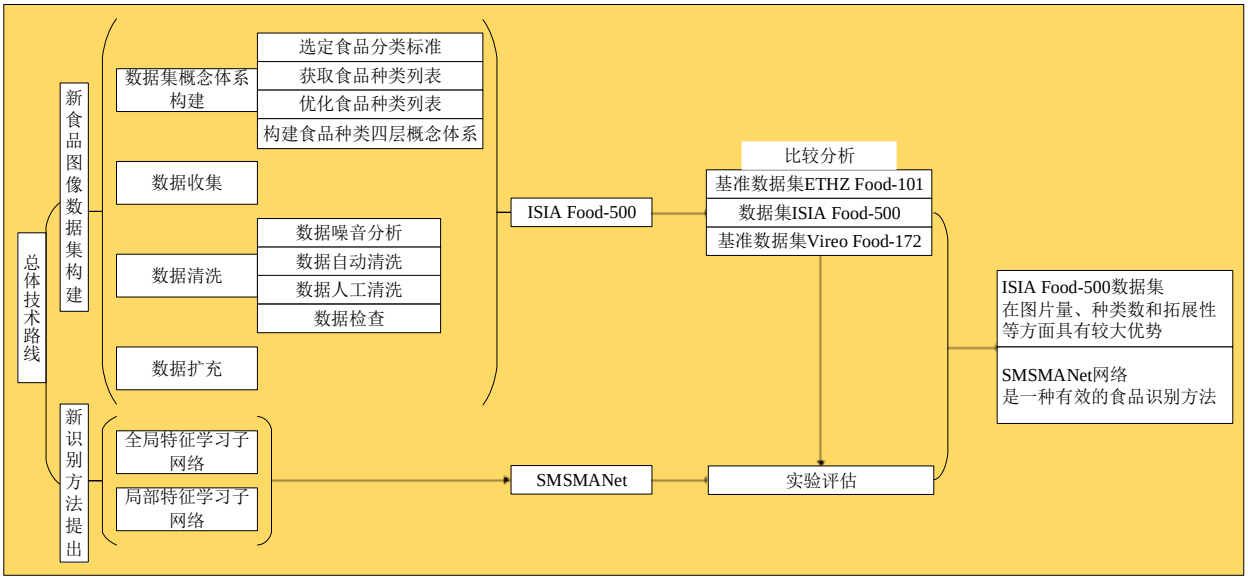


图 1. 2 本文总体技术线路图

Figure 1.2 Overall technology roadmap of this thesis

第 2 章 ISIA Food-500 食品图像数据集构建及分析

本工作重要贡献之一是构建了一个大规模食品图像数据集 ISIA Food-500，这个数据集具有数据覆盖性广、数据质量高、样本密度高和可拓展性良好等优点。覆盖性广是指数据集包含的种类覆盖广和数据来源广，即食品数据集包含世界各地菜品，而不仅包含是中方和西方菜品，还有多种肉类、蔬菜类和五谷类等种类；图片方面的广覆盖性是指图片来源多样，采集的图片来自多个网站、多种图片搜索引擎等。数据质量高是指食品图像数据集包含的食品种类是常见类，数据集中图片都是尺寸大于 100*100 像素、非动态图片和 RGB 三通道完整的图片等。样本密度高是指数据量大且每个单类包含数据量都在 500 张以上。此外，数据集可拓展性也很重要，在某个时间段构建的数据集可能只满足近几年业界需求，数据集是否可以指导其他研究人员持续扩充，是否在统一标准下构建并保持数据集结构一致性等，这都将影响数据集未来使用寿命。

大规模食品图像数据集 ISIA Food-500 构建主要有四个步骤：数据集食品种类概念体系构建、数据收集、数据清洗和数据扩充。最终，本工作构建的食品图像数据集 ISIA Food-500 包含世界各地 12 个大类 60 个食材小类的 500 种食品，图片总量为 405,776 张，每类图片量都在 500 张图片以上。

2.1 数据集概念体系构建

2.1.1 选定食品分类标准

良好数据集的一个重要特点是拥有高覆盖率数据种类。比如以对象为中心的数据集 ImageNet[9]和以场景为中心的数据集 Places[39]具有很高的类别覆盖率。ImageNet 数据集首次在 2009 年 CVPR 国际会议上发布时，包含了 12 个大类的 5,247 个子类约 320 万张图片。ImageNet 的类别概念是基于 WordNet[40]层次结构进行组织构建的。WordNet 是普林斯顿大学联合心理学、语言学和计算机学领域专家设计的一种认知语言学英文词典，它按照单词含义构建了一种单词网络。比如“wet”会与“moist”、“humid”和“damp”等近义词相联系。Jia Deng 等人[9]根据 WordNet 选定关键词从图片搜索引擎下载图片，通过亚马逊进行众包完成构建。Places[39]数据集发表在 2018 年 TPAMI 国际期刊上，其包含了 434 个场景语义

类的一千万张左右场景图片。与 ImageNet 相似，它也是基于 WordNet 来构建场景种类，然后通过亚马逊进行众包完成构建。这样构建数据集的好处是它们的种类概念来源于 WordNet 这个层级结构体系，这就保证了数据集种类的概念体系有理可依，使数据集的种类呈现出合理层次性和区别性。同时，因为数据集的种类名称来源于 WordNet 概念体系，也使得此数据集具有很好拓展性，后续的数据集构建者可以依据 WordNet 进行数据集种类的扩充，或者根据已有的数据类进行图片数量的扩充。

秉承这种依托某种体系概念来组织数据集种类概念体系的思想，作者决定选择维基百科基于食品原材料的“Lists of foods by ingredient”食品列表和国际食品及农业组织的《食品添加剂通用法典标准》（GSFA）分类标准作为构建图像数据集食品种类概念体系的分类依据，主要原因是：

（1）食品图像识别主要是直接根据图像中食品呈现的材料状态（材料烹饪加工后的样子）来进行视觉识别，无法从图中直接看出国家和品牌信息，因此放弃选择基于食品所属国家的“Lists of foods by nationality”和基于食品品牌的“Lists of brand name foods”。

（2）本文欲构建大规模食品图像数据库，“Lists of foods by ingredient”下的食品列表中食品最多，所以首选基于食品原材料的食品列表。

（3）在国际食品及农业组织的《食品添加剂通用法典标准》（GSFA）中，食品大类的分类结构有助于构建食品分类概念体系的食品大类层。

2.1.2 获取食品种类列表

针对维基百科基于食品原材料的“Lists of foods by ingredient”食品列表，作者使用深度优先算法程序来遍历网站链接，以便获得较完整的食品种类名称，得到了“列表-食材子列表-食品种类列表”这样的食品种类列表，形如“Lists of foods by ingredient(71 P)/List of cassava dishes/Cassava Pie”。这样，本工作获得 4,943 类原始的食品种类列表。

2.1.3 优化食品种类列表

随后，作者对获取的维基百科食品列表进行优化处理：

（1）优化重复食品类别：为保证每个类名在种类概念体系中唯一性和清晰性，作者删除了一些模糊性归类，比如 apple pie 同时在 apple dishes 和 fruit dishes 列表中，作者删除 fruit dishes 列表中的 apple pie 类名。

（2）优化类别粒度：本文对维基百科食品列表中的第二节点（食材层）和第三节点（食品种类层）做粒度优化。比如将 fruit dishes 改为 other fruit dishes，因为 fruit 包含 apple 类，苹果类及其他具体水果类（banana，pear 类等）已经存在于数据集种类概念体系中，为了使 List of other fruit dishes 列表中不重复包含已存在水果列表，作者用 other fruit 代指那些没有单独成食品列表的水果类。同样处理的还有 other vegetable dishes、other meat dishes 和 other fish dishes 等

（3）优化层级结构：从维基百科上下载来的食品列表大多都是三层节点，但有些是四层节点结构，作者统一将所有列表优化到三层结构。比如 Lists of foods by ingredient(71 P)/List of almond dishes/Marzipan/Königsberg marzipan 和 Lists of foods by ingredient(71 P)/List of almond dishes/Marzipan/Marzipan pig，作者删除这种条目的第三节点，这是为了保证整个概念体系的粗细粒度层次统一性，需要保留食材子类节点（第二节点）和食品单类名称（第四节点，叶节点）。假设如果保留第三节点 Marzipan 而删除第四叶节点，将 Marzipan 作为一个食品种类收集图片，则最后在 Marzipan 这个食品中将包含 Königsberg marzipan 和 Marzipan pig 两种食品图片，这类图片在视觉上将会呈现出多种模式，计算机在学习这类图像模式时，将很难确定究竟哪种模式是这种食品明确模式，导致模型性能效果不佳。

（4）优化规范食品类名：作者还对一些特殊符号（比如｛｝、*、|和!等）做了删除，但是保留的法语单词，意大利文单词等非英文单词的原样。因为这样在用食品类名做检索词时，才可以保证搜到准确的食品图片。

（5）优化食材子类层：维基百科食品类别的 List of rice varieties 和 List of edible flowers 中的类是食品原材料。rice varieties 中呈现的是各种品种的米，比如 Calrose rice，Ofada rice，Pecan rice 等；List of edible flowers 呈现的是各种食用花，比如 Althaea，Malus，Dianthus 等。作者删除这个两个食品食材子类之后，Lists of foods by ingredient 下的节点数由 71 个变成了 69 个。

2.1.4 构建食品种类四层概念体系

本文参照国际食品及农业组织的《食品添加剂通用法典标准》（GSFA）分类标准将维基百科的食品列表结构从三层扩到四层，比如 Lists of foods by ingredient(71 P)/List of chicken dishes/Ayam bakar 扩展为 Dish(GSFA)/ Meat and meat products/ Chicken dishes/Ayam bakar，在维基百科食品列表第一节点和第二节点之间增加了《食品添加剂通用法典标注》的分层，维基百科食品列表第一节点 Lists of foods by ingredient 是食品范畴名，维基百科食品列表第二节点 List of chicken dishes 是食材子类，这两层之间粒度跨越较大，如果加入一个中间层，则整个概念体系层次粒度逐渐细化，各层节点数均匀合理。

表2. 1　食品种类列表的层次统计表

Table 2.1 Hierarchical statistical table of food category list

第一层	第二层	第三层	第四层
Dish（GSFA）	Meat and meat products	Bacon dishes	Bacon_egg_and_cheese_sandwich
			Bacon_and_eggs
			…
			Bacon_sandwich
		…	…
		Beef dishes	Beef_bun
			Filet_mignon
			…
			Ginger_beef
	…	…	…
	Cereals and cereal products	Rice dishes	Claypot_chicken_rice
			…
			Congee
		…	…
		Noodle dishes	Fried_noodles
			…
			Wonton_noodles
1 个节点	11 个节点	69 个节点	3,309 个节点

结合国际食品及农业组织的《食品添加剂通用法典标准准》（GSFA），本工作将 69 个食材层列表分到了 11 个食品大类中。最后构建了一个四层的“食品范畴类-食品大类-食材子类-食品单类”的层次结构体系，第一层、第二层、第三层和第四层分别有 1、11、69 和 3,309 个节点。层次结构体系可视化的见表 2.1。

2.2 数据收集

本工作将数据集概念体系叶节点（即食品种类名）作为搜索图片的关键词，这样检索的食品种类就有 3,309 种。起初，本文直接用食品类名尝试性地在三个图片搜索引擎谷歌图片[41]、百度图片[42]和必应图片[43]上检索图片，发现：（1）有些食品名搜出很多非食品的图片，比如检索词“Utan”，既是一个食品名字，也是一个俄罗斯地名。于是，本文将食品类名关键词分别与“food”和“dish”组合来共同作为检索词，当用“Utan food”和“Utan dish”检索到的大多则是食品图像。（2）从图片搜索引擎集合了不同网站食品图片，它们可能来自美食网站，可能来自美食论坛，可能来自公开的美食分享等，这样直接从图片搜索引擎就可以找到不同来源的食品图像，保证了图片良好多样性和数据源更广泛覆盖率。相比于从单一某个美食网站检索图片，利用三大图片搜索引擎检索图片既保证了数据多样性又有较高效率。另外，有些美食网站（比如美团，大众点评等）具有很强的反爬虫设置，使用这种网站的食品图片，可能造成侵权等问题。虽然使用图片搜索引擎有些图片也无法爬虫下来，但凡是能爬取的图片均是公开发布并且可便利获取的图片，没有其他侵权等问题。从图片搜索引擎下载图片的可靠性早在 2009 年 Jia Deng 等人[9]构建 ImageNet 时就已验证，ImageNet 的工作也是从图片搜索引擎下载大量图片，然后进行清洗标注。同样地，Bolei Zhou 等人[39]在 2018 年构建 Places 千万级数据集时也是从图片搜索引擎上采集图片，他们明确提到是从谷歌图片、必应图片和 Flickr 上获取图片。但是因为 Flickr 上图片在中国大陆访问受限，所以图片资源无法顺利获得。这说明从综合性图片搜索引擎（尤其是谷歌图片和必应图片网站）中收集图片是目前高效且有效的主要方法之一。

确定图片检索词和图片数据源后，本工作利用 python 的 icrawler 库爬取图片，分别从三个不同的图片搜索引擎（谷歌图片、百度图片和必应图片）上爬取相关食品图片。icrawler 库中有分别对应于谷歌、百度和必应的程序包 GoogleImageCrawler、BaiduImageCrawler 和 BingImageCrawler 可以直接调用，根

据检索词爬取三大图片搜索引擎的相关图片并保存到本地。经过“食品类名+food”和“食品类名+dish”两轮爬虫，作者使用四台机器耗时一周，设置每个检索词在每个图片搜索引擎下爬取图片最大量为 3,000 张。最终，总共下载了 3,309 类食品 6,196,253 张候选图片。三个图片搜索引擎的两轮爬虫图片的基本统计情况见表 2.2，其中图片数量最多的类 Fillet of Beef Prince Albert 类中有 6,359 张图片，最少的类 Sha Phaley 食品类中有 13 张图片，平均每类爬取图片 1,872 张，单类图片数量大于 1,000 张以上的有 2,143 类。造成爬到的每类图片数量不均衡的可能原因是：（1）有些图片链接只供检索，却因下载受限导致无法下载。（2）有些食品是常见食品，在世界各地都是热门美食；有些食品本身不常见，可能是新创造出的菜品，也可能是来自小国地区食品不被世界其他地区的人熟知。每类食品可爬到的图片数量也在一定程度上反映了这款食品的热门程度。

值得注意的是，根据 ImageNet 构建者统计分析，爬虫的图片中大约只有 10%图片是符合数据集构建的图片。因此，这些爬取的图片中含有大量的噪音图片，需要进一步图片清洗才能作为构建数据集的正式图片。

表 2.2　两轮爬虫图片量统计

Table 2. 2 Statistics of two rounds of crawler images

单类中最大图片数	6,359 张
单类中最小图片数	13 张
平均每类图片数	1,872 张
前 1,000 类图片数范围	6,359~2,886 张
前 2,000 类图片数范围	6,359~1,059 张
大于 2,000 张的类数	1,268 类
大于 1,000 张的类数	2,143 类

2.3 数据清洗

2.3.1 数据噪音分析

对于下载的候选图片，作者在清洗之前进行了抽样观察，以便找出候选图片数据中噪音问题，有针对性采取清洗措施。候选图片中存在的问题如下：

（1）图片格式。作者观察到候选图片中有 png、jpg 和 gif 等多种图片格式，其中的 gif 动态图片格式不是常规图像识别数据集需要的格式。

（2）非 RGB 三通道图片。作者观察到有些图片是灰色而非彩色图片，于是用代码检查某类下载的所有图片，发现大部分是 RGB 三通道图片，有些是两通道或者四通道的图片。而抽样检测两个基准食品图像数据集 ETHZ Food-101 和 Vireo Food-172，发现都是 RGB 三通道图片。因此，要做一个食品图像数据集，应选用 RGB 三通道图片数据，不能多一个通道，也不能少一个通道，这样方便使用已有的食品识别方法代码来统一验证新数据的可信度，并与其他食品数据集公正比较。

（3）损坏图片。作者观察发现有些图片是下载过程中损坏，可能是网络不稳定导致没有下载完整或者存储写入时硬盘部分区域有问题等原因，这些不可避免。虽然这种问题图片的数量比较少，但这种问题图片是一定不可出现在食品图像数据集中。

（4）图片尺寸不一致。作者观察到候选图片间尺寸不一致，有些图片尺寸长或宽只有几十像素，有些长或宽可达一千多像素。这些图片来源于网络各个网站和论坛，拍摄设备可能包含手机、相机和摄影机等，所以出现各种尺寸照片是可以理解的。但是，像 ImageNet 数据集和 Places 数据集都对图片尺寸做了适当清洗，原因是图片在输入模型时，都会把原始图片缩放到同一尺度，比如缩放到 512*512 像素大小或者 256*256 像素大小等，这样可以减小图像尺寸变化对模型性能的影响。如果图片尺寸太小，在模型输入放大图片时，图片显示的主体内容会被严重变形模糊化，影响模型对图片数据内容学习。所以，一个好的数据集中图片应该有适当图片尺寸，尤其不能有像素尺寸较小的图片。这种道理就仿佛给你一张小图片，你越放大图像内容越模糊，最后可能看不清图像内容；但是给你一个大尺寸图片，缩小时则不会使图像内容模糊。

（5）非食品图片。作者观察到候选图片中存在很多非食品图片，比如有涂鸦画、文字菜谱和各种广告等。一个好的食品图像数据集追求的是来自真实世界真实摆拍的食品图片，这种非食品图片是需要删除的噪音图片。

（6）非本类食品图片。作者发现虽然有些图片是食品图片，但并不是这个类的图片。原因可能是上传者在上传此图时给定标签是错误的。由于候选图片是代码程序根据关键字（即原始图片的标签）来进行下载的，所以如果图片标签是错误的，收集的图片就是与真实参照图片不符。

（7）图片内容被遮掩。作者发现有些候选图片虽然是这个种类的图片，但

是图片上真正展示食品的区域却有限，图片上有文字、明显 logo 水印和部分广告等占据了相当大图片版面，这种给人一种不“干净”的感觉，不是那种“纯净”食品图片。所以，这种有较多无关内容占用太多图片版面的图片也不宜存在于一个高质量数据集中，除非研究的是那种部分遮掩状态的食品图像识别课题，但至少这种图片目前不适合食品识别研究。

（8）重复图片。候选图片中出现重复图片的问题来源于两方面：第一，收集的图片来源于多个图片搜索引擎。第二，同一图片搜索引擎下也可能出现重复图片。这主要是因为上传者会把自己的图片分享到多个网站，其他网友也可能把图片转载到不同的网站。而爬虫程序只是对检索到的图片链接进行下载，并不会考虑图片内容是重复，这就造成了爬虫下载的图片中存有大量重复图片。

（9）其他问题，比如从多个图片搜索引擎下载的图片命名有重复，则在合并时需要对图片重名等。

2.3.2 数据自动清洗

面对几十万数量候选图片，作者希望高效地完成噪音图片的清洗工作，所以在细致观察候选图片数据后，优先考虑使用代码程序自动清洗，其次再考虑人工标注清洗。自动清洗噪音图片的程序有三个：

（1）清洗程序一：删除 gif 格式图片；删除非 RGB 三通道图片；删除长或宽尺寸小于 100 像素的图片；删除程序不能读取的损坏图片；把多个图片搜索引擎获取的同类图片移动到同一个文件夹并重命名图片。

（2）清洗程序二：本工作借助 VGG-16 模型[44]训练了一个食品/非食品二分类器，用来对非食品图像进行清洗。具体方法是，作者结合两个基准食品图像数据集 ETHZ Food-101（西方菜）和 Vireo Food-172（中国菜）训练集的图片作为训练集正样本。然后，从 ImageNet 和 Places 数据集中随机选择了约 40 万张非食物图像作为训练集负样本。这些正样本和负样本共同构成了训练集。同理对于测试集，一部分来自 ETHZ Food-101 和 VireoFood-172 的所有测试样本作为正测试集，而另一部分则来自 ImageNet 和 Places 数据集的大约 100,000 个随机选择的非食品图像作为负测试集。最终，训练 VGG-16 模型二分类器的数据集包括约 530,000 张训练图像和 168,000 张测试图像。利用这个训练集对分类器进行了训练，在这个构建的测试集上进行了测试，获得了分类精度为 99.48％的二分类器。

接着，使用训练好的二分类器清洗候选图片中大部分非食品图片。

（3）清洗程序三：这一步主要是清洗重复图片。本工作用一种哈希算法来清洗候选图片中的重复图片。具体地解释，首先将图片调整到长*宽尺寸为 32*32 像素的大小。为了便于计算处理，把图片数据转化为灰度图像，并得到图片像素 0~225 的灰度值。然后通过比较像素之间的像素值，得到差异值序列，并计算转化为图片的哈希码。最后，采用暴力计算两两图片之间的汉明距离来判断两个图片是否为同一图片。如果两张图片之间的汉明距离低于阈值，将删除其中一张。这样，在每个食品类别中就不会包含相同图像。在本工作中，作者将阈值设置为 150，这是经过多次尝试得到的经验值。作者发现，如果阈值太严（阈值大），则会过度删除一些图片，比如同一摆拍的食品，只是角度稍微旋转了，也会被删除，但其实这两张图片是不同的，不算重复图像。反之，如果阈值太松（阈值小），则会保留一些重复的图片。作者取一个适中经验值，是为了减轻后续人工清洗工作量，同时保证合适图片不被代码程序过度删除。

经过几个代码程序自动清洗后，作者抽取了部分数据进行了统计分析，如表 2.3 所示，总共抽样 43 万多张图片进行程序处理，清洗后剩余 27 万多张，剩余了大概原数据量的 64%，具体数据可见表 2.3 所示。

表 2. 3　程序自动清洗噪音图片数据统计

Table 2.3 Statistics of noisy images automatically cleaned by the program

图片来源	抽样初始图片量	程序一清洗后图片量	程序二清洗后图片量	程序三清洗后图片量	最终剩余图片量	剩余比例
谷歌图片	214,598	6,996	63,572	1,143	142,887	67%
必应图片	82,608	1,780	25,117	1,168	54,543	66%
百度图片	135,676	3,479	51,629	1,713	78,855	58%
总计	**432,882**	**12,255**	**140,318**	**4,024**	**276,285**	**64%**

2.3.3 数据人工清洗

使用代码程序清洗噪音图片固然高效省事，但是机器毕竟是机器，有些噪音图片计算机无法清洗彻底，这时候就需要人工标注清洗，人工标注清洗虽然慢，但是清洗比较准确，知道什么样图片是应该保留，什么样图片必须删除。对于那些只能定性处理而不能定量处理的图片，人的主观能动性发挥了机器无可代替的作用。如果说程序清洗是粗糙的和定量的操作，那么人工标注清洗图片就相当于比较精细的和定性的清洗操作。

为了统一规范人工标注清洗，本工作提前制定了适合本工作数据集构建的清洗标准，简要步骤如下：

（1）每个标注人员从维基百科中搜索该类食品真实样例图片，细读维基百科对这类食品的讲解和注释，确保清楚这类食品是什么，长什么样，有什么区别性的特征等。

（2）使用自主开发的 Java 标注工具软件进行人工标注清洗。参照真实样例图片，对标注工具读取的当前候选图片进行判断：如果是本类食品图片，则选“下一张”；如果不是本类食品图片，则选“删除”，标注工具自动读取下一张。

（3）明确强调需要删除和暂且保留的图片类型，比如涂鸦画、广告等。

■需要删除的类型有：

a.“多合一”图片类型需要删除。所谓“多合一”就是一个图像版面上同时合成了多个小照片。

b.非食品图片类型需要删除。虽然，训练的二分类器清除了大部分的非食品图像，但是候选图片中还是存在一些非食品图像，这时就需要人工使用标注工具去清除。

c.涂鸦图片类型需要删除。不管涂鸦内容是食品内容还是非食品内容，这种图片都需要删除。

d.包装画类型的图片需要删除。本文构建的数据集是真实摆拍的照片，不能含有包装画类型的图片。

e.内容上有较多文字的图片类型需要删除。有些食品图片展示的主体内容上有较多打印文字、logo 和非食品内容在上面，或者图片上既有食品部分也有非食品部分，非食品部分占了大部分图片版面。本工作规定这种非食品内容不能超过图片版面的四分之一，且不遮掩食品主体内容，否则应给予删除。

f.食品名是一种原材料的图片类型需要删除。有些类名是一种原材料，比如 sardine，维基百科查询是沙丁鱼，本工作构建的是可直接食用的或简单加工的餐桌或快餐食品，而不是需要复杂加工的食品材料。

g.不是本类食品的图片需要删除。

■暂且保留的图片类型：

需要保留的图片基本就是参照真实样例图片的内容，呈现的是同种食品图片。这里强调的保留是指那种图片四周有小 logo 或者文字的图片，且这些非食品内

容占图片版面不到四分之一。

2.3.4 数据检查

本工作组织多名本实验室同学来进行人工标注，这些组员对食品计算研究有一定了解，对食品识别的知识掌握更加到位，所以对食品图像标注结果相对比较可靠。但是，人们在工作难免会有一些小纰漏。为了尽可能构建优质食品图像数据集，本工作还是增加了最后的检查工作。作者对人工标注的每类图片进行一一检查，并选择合适类来构建食品图像数据集。

所谓检查，其实也是作者使用标注工具对组员已标注的每类数据做再次标注。因为已经人工标注过一次，当再次标注时，基本都是符合该类的图片，所以检查标注速度会比较快。对于检查过程中出现标注疏漏太多的类，会分派给组员再次标注；而对于类中少数残留的不合规图片，检查标注时就顺便清除。

在检查数据的过程中，作者发现存在一些食品名称不同但视觉上重复的类别，这主要是维基百科食品列表中没有甄别不同食品名称可能是同一种食品，比如 Tonkotsu ramen 和 Hakata ramen 虽然是两个不同食品名字，但都是豚骨拉面，属于同一种食品，于是合并了这两类。如图 2.5 所示，这两个名称不同的食品也呈现出相同视觉形态。这种重复类在起初构建食品种类概念体系时，并不能发现，因为是不同名字，对于不熟悉这两道食品的人来说，无法从食品名称上判断这两类食品是否重复。

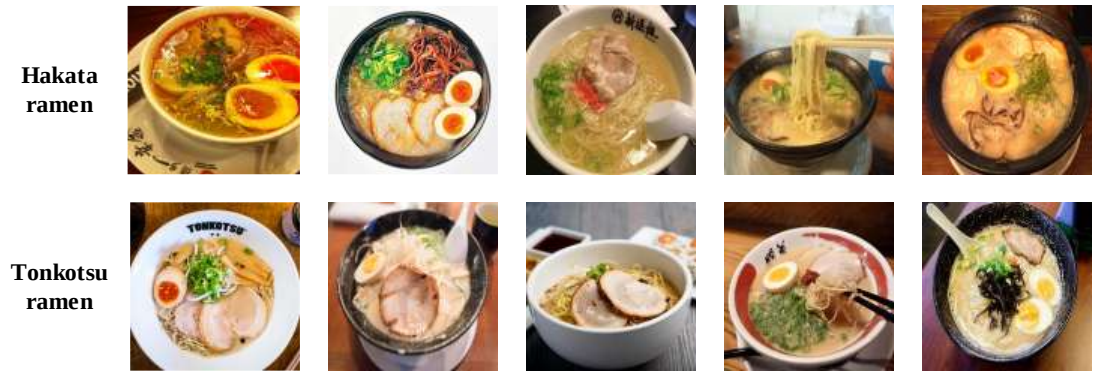


图 2. 1　不同食品名称的同类食品

Figure 2.1 Similar food with different food names

2.4 数据扩充

经过上一轮的数据收集和清洗之后，作者发现许多食品种类的图片还是比较

少。于是对这些食品种类图片量进行了扩充。

首先考虑第一轮标注后，对单类图片数量接近 500 张的食品种类进行扩充，即对标注后剩余数量在 300~500 张的类进行扩充。扩充数据的操作主要有两个方面：

（1）检索词扩充。这里的检索词扩充是指将食品类名翻译成中文、法文、土耳其文和西班牙文，然后在图片搜索引擎上进行爬虫。之所以翻译成这几国语言，是因为作者在多处网站调研世界美食较多的国家，虽然没有一个权威机构的美食排名，但是从网友的帖子或者其他分享中，作者综合考虑选定这几国语言，后续如有需要，可以再多翻译成其他语言。在翻译时发现并不是所有食品名称都可翻译成这四国语言，其中会有些无法翻译的名称。但是，这并不影响大多数类的数据量扩充，此环节的目的是尽可能多扩充数据。

（2）数据源扩充。这里的数据源是指可以爬虫下载食品图片的网站。文献[1]是一篇比较全面的食品计算综述性文献，其中提到了很多美食网站。作者对其中的美食网站进行了调研，文献[1]中的网站大部分主要是已知食品数据集网址链接。其余网站有些是菜谱形式网站（比如 www.yummly.com），这种网站有详细菜谱，但每类食品仅有几张图片，且食品种类较少；有些是提供餐馆位置的网址（比如 foodspotting.com）等。另一方面，作者调研了其他的美食网站并从中下载图片进行数据集扩充，比如网站 www.meishij.net 和 foodgawker.com 等。

扩充收集的数据，经过图片清洗后，再与上一轮清洗后的图片合并，并再次进行一遍重复图片清洗。因为两轮清洗是分开进行的，在两轮检查完后合格的图片中可能有些是彼此重复。最后，再对数据集中的图片进行可读性检查并按顺序重新命名。

目前，本文得到了一个 500 类的食品图像数据集，其中每类图片数量均大于 500 张，数据集图片总数 405,776 张，本文命名它为 ISIA Food-500。

2.5 数据集统计与分析

在大规模食品图像数据集 ISIA Food-500 构建完成后，需要对数据集特点进行梳理。作者希望通过与现有的两个食品图像基准数据集（ETHZ Food-101 和 Vireo Food-172）各方面数据分析比较，体现出本工作数据集 ISIA Food-500 的优势和特点。

2.5.1 数据量分析

在众多的食品图像数据集中，ETHZ Food-101 和 Vireo Food-172 是两个基准食品图像数据集，主要食品图像数据集统计情况见表 2.4。ISIA Food-500 和两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）的数据量比较见图 2.2。

表 2. 4　公开的食品图像数据集

Table 2.4 Public food image datasets

数据集	发布时间	种类数	图片数	覆盖范围
PFID[18]	2009	101	4,545	日本
Food50[32]	2010	50	5,000	综合类型
Food85[45]	2010	85	8,500	综合类型
UEC Food100[13]	2012	100	14,361	日本
UEC Food256[14]	2014	256	25,088	日本
ETHZ Food-101[2]	2014	101	101,000	西方
Diabetes[46]	2014	11	4,868	综合类型
UPMC Food-101[47]	2015	101	90,840	西方
UNIMIB2015[48]	2015	15	2,000	综合类型
UNIMIB2016[49]	2016	73	1,027	综合类型
Food11[37]	2016	11	16,643	综合类型
ChineseFoodNet[50]	2017	208	192,000	中国
Vegfru[22]	2017	292	160,000	综合类型
FoodX-251[15]	2017	251	158,846	综合类型
Vireo Food-172[8]	2017	172	110,241	中国
ISIA Food-200[16]	2019	200	197,323	综合类型
ISIA Food-500(本文工作)	**2020**	**500**	**405,776**	**综合类型**

相比之下，ISIA Food-500 数据集拥有比两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）更多的图片总量，大概是它们的 3~4 倍；ISIA Food-500 数据集更多食品种类数，大概是基准食品图像数据集的 3~5 倍。ETHZ Food-101 有 101 个 1,000 张以上的类，Vireo Food-172 只有 6 个 1000 张以上的类，而 ISIA Food-500 数据集有 125 个 1000 张以上的类。

总之，在图片据量和种类数方面，ISIA Food-500 数据集比两个基准数据集有更大优势，这样在需要大量数据来构建复杂模型时，ISIA Food-500 数据集将更好地满足这个需求。

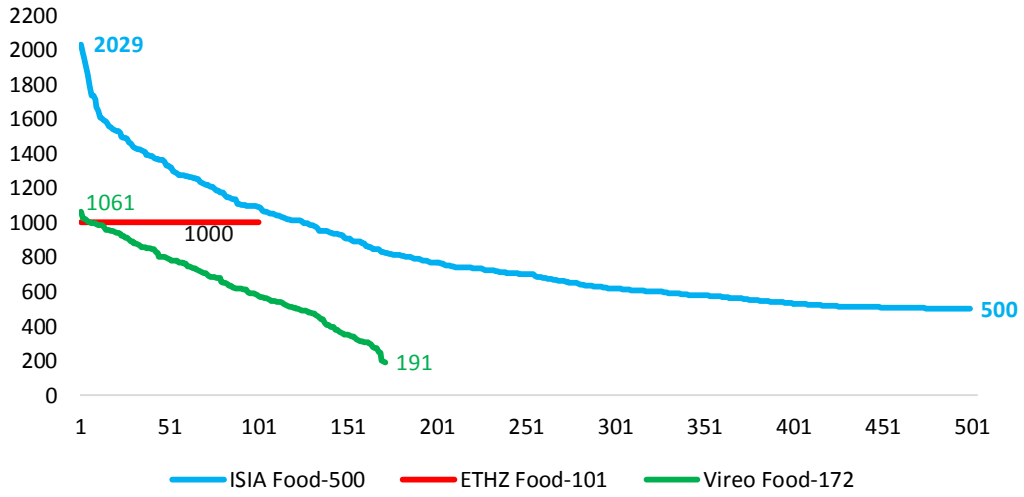


图 2. 2　ISIA Food-500 与基准食品数据集的数据量统计图

Figure 2.2 Statistics of ISIA Food-500 and benchmark food datasets

2.5.2 覆盖性分析

广泛的覆盖性是指数据集种类在世界各个地区的分布情况。在这里着重分析比较 ISIA Food-500 和两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）。对食品所属地区判断是依据维基百科中该食品的描述，维基百科中写明了食品所属国家或地区。

从洲际情况来看，ISIA Food-500 数据集中最多的是亚洲（Asia）食品 200 种，最少的是大洋洲（Oceania）食品 2 种，其中有 58 种是维基百科中没有明确所属地区的食品。ETHZ Food-101 数据集中最多的是欧洲（Europe）食品 44 类，最少的是非洲（Africa）食品 0 类，其中维基百科中没有明确所属地区的有 10 类；Vireo Food-172 数据集的食品种类全部是亚洲中国食品。总体来看，ISIA Food-500 数据集在任何大洲的食品种类覆盖数目都比两个基准食品图像数据集多。三个数据集洲际情况的比较见图 2.3，图中的 uncertain 表示维基百科中没有明确给出食品所属洲际范围。

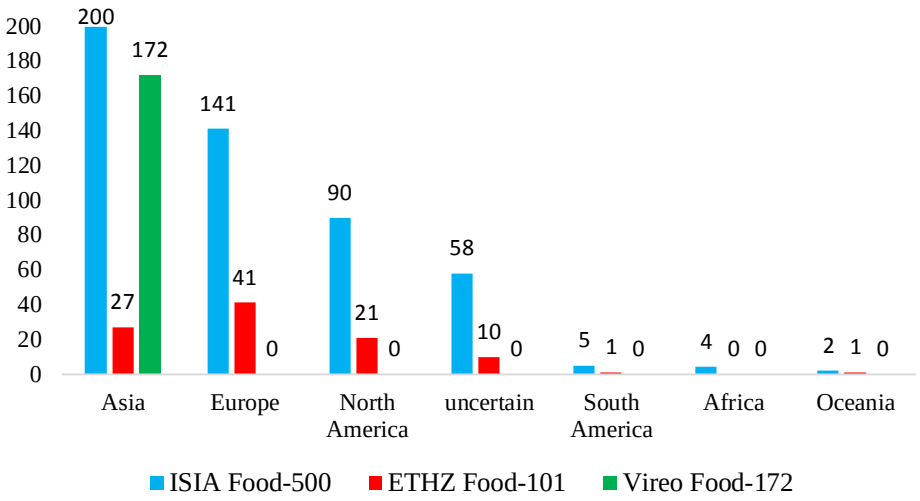


图 2.3 ISIA Food-500 与基准食品数据集食品种类的洲际覆盖情况比较

Figure 2.3 Comparison of intercontinental coverage of food categories between ISIA Food-500 and benchmark food datasets

从国家地区来看，ISIA Food-500 数据集中食品种类覆盖了 52 个国家或地区，而 ETHZ Food-101 数据集覆盖了 22 个国家或地区，Vireo Food-172 数据集覆盖 1 个国家（中国）。三个数据集各国家的比较见图 2.4，其中 unknown 表示维基百科中没有明确给出食品所属国家。

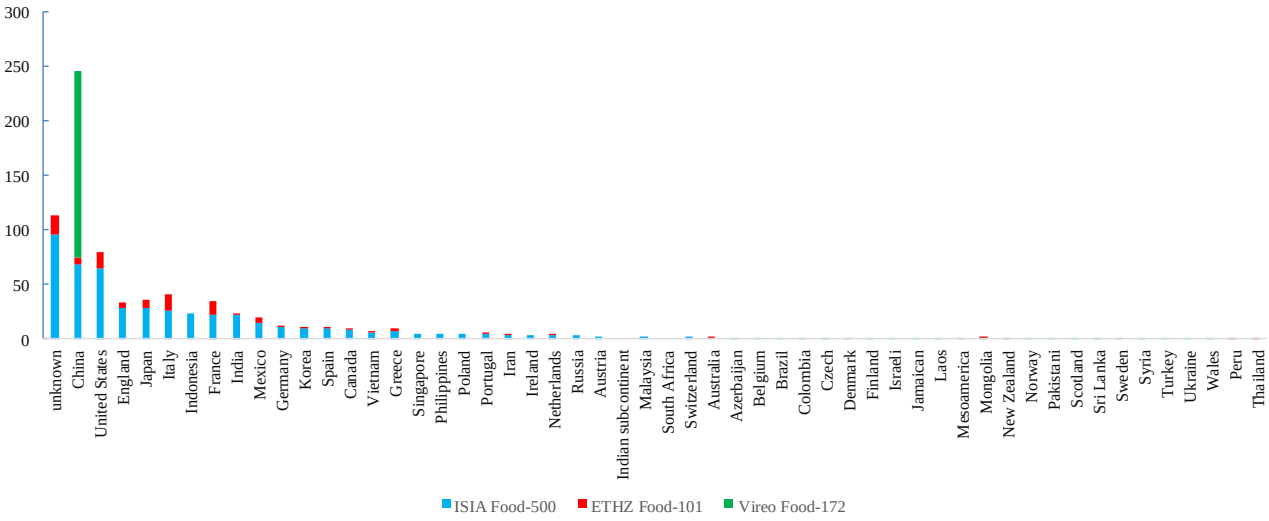


图 2.4 ISIA Food-500 与基准食品数据集食品种类的国家覆盖情况比较

Figure 2.4 Comparison of national coverage of food categories between ISIA Food-500 and benchmark food datasets

由此可见，ISIA Food-500 数据集比 ETHZ Food-101 数据集和 Vireo Food-172 数据集的食品种类覆盖性更广泛。由于，各个大洲和不同国家食品各具特色，覆盖的地域越多，在构建高级食品分类模型时学到的食品模式知识越多，这样得到

的食品识别模型应用地域也会更广泛。ISIA Food-500 数据集更加广泛的覆盖性将有助于构建的食品识别模型更加符合各个地域需求。

2.5.3 多样性分析

多样性主要指食品种类在本工作构建“食品范畴类-食品大类-食材子类-食品单类”的层次结构概念体系中第二层和第三层的分布情况。因为第一层是食品范畴类，ISIA Food-500、 ETHZ Food-101 和 Vireo Food-172 数据集都是食品范畴的数据集，第四层是具体食品种类，而这两层进行比较意义不大，看不出主要区别，所以本文将重点比较第二层和第三层分布情况。

第二层是在国际食品及农业组织的《食品添加剂通用法典标准》（GSFA）标准下划分的食品大类层。三个数据集食品种类在该层的分布情况见图 2.5，从图中可以清晰看到 ISIA Food-500 和 ETHZ Food-101 数据集在 11 个大类中都有分布，而 Vireo Food-172 数据集只在其中 6 个大类中有分布。同时，还可以看出 ISIA Food-500 在绝大部分大类中分布的数量都比其他两个数据集多，因此如果要更好地学到各个大类的食品图像模式，则推荐首选 ISIA Food-500 数据集来进行训练。

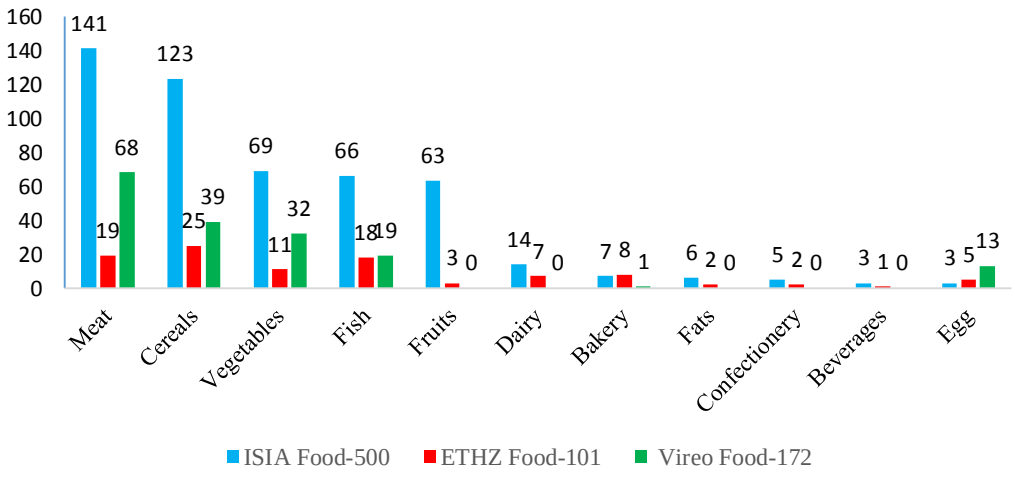


图 2. 5　ISIA Food-500 与基准食品数据集食品种类在第二层分布情况比较

Figure 2.5 Comparison of food categories distribution of ISIA Food-500 and benchmark food datasets in the second layer

从第二层分布情况来看（图 2.5 所示），ISIA Food-500 比 Vireo Food-172 数据集有明显的食品大类多样性，但是 ISIA Food-500 和 ETHZ Food-101 数据集在多样性的区别不能明显体系，因为两个数据集都在 11 个大类上有分布。于是作者进一步探究在第三层（食材子类层）的分布情况。从图 2.6 容易看出 ISIA Food-

500 和 ETHZ Food-101 数据集间的类别多样性区别：在第三层，ISIA Food-500 数据集食品种类分布在 60 个原材料类中，而 ETHZ Food-101 数据集食品种类分布在 41 个原材料类中，Vireo Food-172 数据集食品种类分布在 26 个原材料类中。这充分体现了 ISIA Food-500 数据集含有更多食材种类，可以为食品识别模型训练提供其他两个数据集没有的食品类别， ISIA Food-500 数据集食品种类的多样性也更接近现实世界食品多样性复杂的真实情况。

综上，ISIA Food-500 数据集在第二层国际 GSFA 分类标准下和在第三层食材子类层上都有更好的多样性，可以为复杂的食品识别模型提供更贴合实际应用的高性能数据。

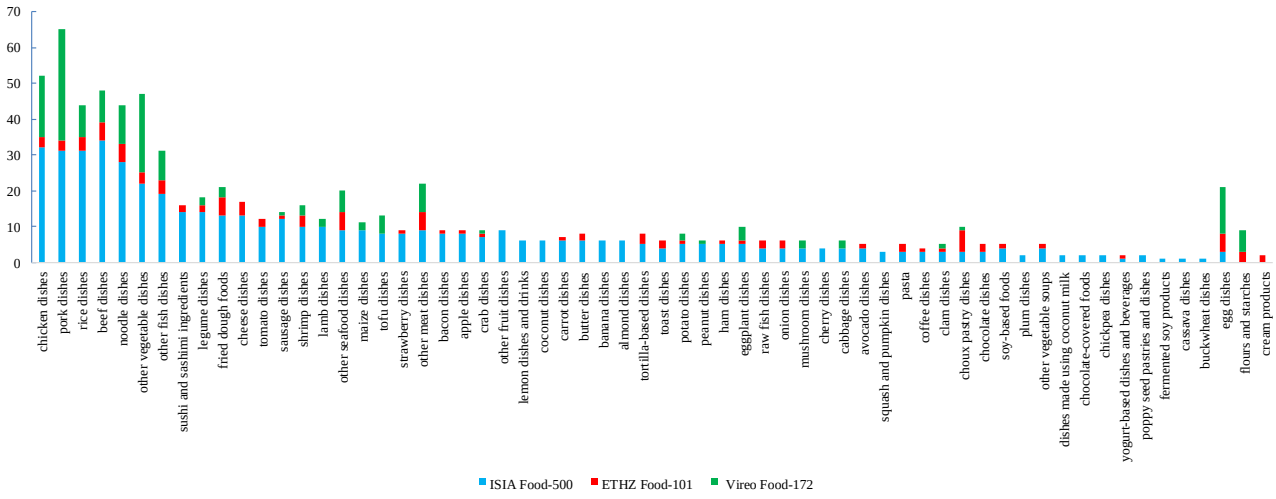


图 2. 6　ISIA Food-500 与基准食品数据集食品种类在第三层分布情况比较

Figure 2.6 Comparison of food categories distribution of ISIA Food-500 and benchmark food datasets in the third layer

2.5.4 可拓展性分析

评价数据集另一个重要指标是数据集的可拓展性。一个数据集不但要满足现在当下研究需要，还应具备后续研究人员拓展扩大数据集的可行性。ETHZ Food-101 和 Vireo Food-172 这两个数据集在文献中没有提到数据集食品种类概念体系，没有能够清晰地为后续他人扩充数据集种类提供一个概念指导。而本工作的 ISIA Food-500 数据集是在国际食品及农业组织的《食品添加剂通用法典标准》（GSFA）标准和维基百科食品列表结合基础上，构建了一个四层的食品种类概念体系，该体系由粗到细层次鲜明地逐步划分出 3,309 个食品种类，这个庞大概念体系即有国际标准的详细解释，又有维基百科真实样例示例。丰富的食品种类可以让后续研究者在该概念体系指引下，对食品图像数据集食品种类进行增加，

对种类中图片进行扩充。从这个角度来讲，ISIA Food-500 数据集比两个基准食品图像数据集 ETHZ Food-101 和 Vireo Food-172 更具有明确的可拓展性，便于广大研究人员在本工作基础上扩建食品图像数据集。

2.6 小结

本章主要讲的是大规模食品图像数据集的构建及分析，本文从数据集概念体系构建、数据收集、数据清洗、数据扩充，详细介绍每一步操作，并解释其中一些必要原因。最终，本工作构建了一个 500 种食品 405,776 张图像的食品图像数据集 ISIA Food-500。

本章通过将 ISIA Food-500 数据集和两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）进行分析比较，体现了 ISIA Food-500 数据集特色和优势。本章从数据量、覆盖率、多样性和拓展性四个方面逐一对三个数据集的情况进行了分析，展示了 ISIA Food-500 数据集图片数量最多、覆盖率最广、种类最多样和良好拓展性的优点，这个数据集适合世界多国范围应用，能识别更多食品种类，也能加接近现实世界食品分类情况；同时，也更能给食品科研带来挑战。实际生活中没有广泛应用食品识别技术，其中一个主要原因就是现有食品图像数据集规模小、种类少和多样性有限导致食品识别模型不能满足实际需求。后续章节会介绍一个基于图像的食品识别方法，并从实验方面说明 ISIA Food-500 数据集具有一定的挑战性，能够刺激食品识别技术持续接力发展，促进食品识别技术更快地落地及广泛地实际应用。

第 3 章 基于堆叠多尺度多注意力网络（SMSMANet）的食品识别

随着手机和数码相机等各种具有拍照功能的便携设备普遍使用，生活中出现了大量的美食图片，通过这些食品图像进行食品识别研究是一项重要的基础性工作，可以为多模态食品识别和食品检索等任务提供帮助。在本章中介绍的 SMSMANet 网络是通过利用空间和通道注意力机制，结合多尺度策略，共同学习食品图像的全局特征和局部特征来进行食品识别。

3.1 问题描述

首先，有些食品外形在全局视觉形式上存在很大类内差异，如图 3.1 中 Butter pecan 这种食品类内形状差异大。虽然针对食品识别问题已经发展出很多方法，但是这些方法大多集中于提取特定类型或某些类型的特征，而忽略了其他方面的特征，例如 Vinay Bettadapura 等人[51]通过提取颜色特征来进行食品识别，Niki Martine 等人[34]设计了一种网络可以捕获特定垂直结构来进行食品识别；其次，有些食品图像的判别性细节很难捕捉，如图 3.1 所示，全局特征不足以区分 Corn stew 和 Leek soup 这两种食品，需要借助局部特征来辨别。为了提高识别性能，一些工作[4][51]利用辅助属性信息（比如位置信息和原材料信息）进行食品识别。但是，当原材料信息和地理信息等辅助信息不可用时，这类方法可能不起作用。

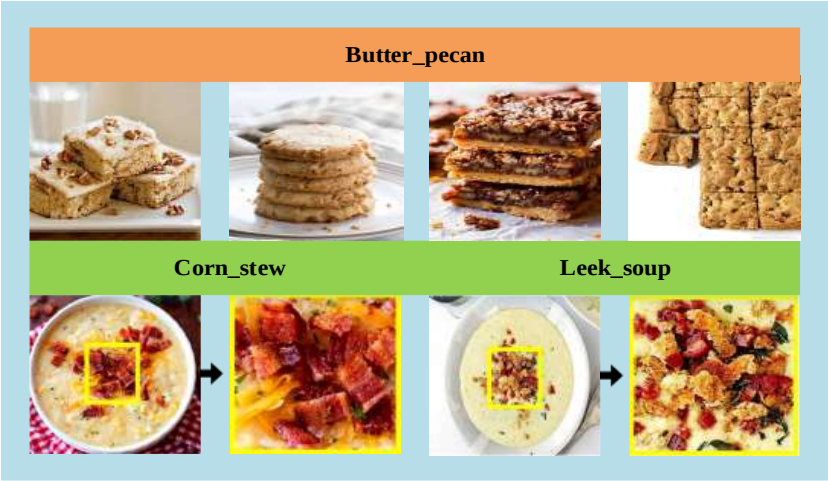


图 3. 1　食品图像类内差异大和类间差异小的示意图

Figure 3.1 Food images with large intra-class difference and small inter-class difference

为了解决以上两个问题，本文提出了一种新的食品识别方法：堆叠多尺度多注意力网络（Stacked Multi-Scale Multi-Attention Network, SMSMANet）。该方法以图片为输入，通过空间和通道注意力机制，结合多尺度策略，融合全局特征和局部特征来进行食品识别。接下来本章各小节将详细讲述 SMSMANet 网络结构。

3.2 SMSMANet 网络框架

本文提出了一种堆叠多尺度多注意力网络（Stacked Multi-Scale Multi-attention Network, SMSMANet）来共同学习互补性全局特征和局部特征进行食品识别。该方案结构图如图 3.2 所示，SMSMANet 主要有两部分构成：全局特征学习子网络（Global Feature Learning Sub-network, GloFLS）和局部特征学习子网络（Local-Feature Learning Sub-network, LocFLS）。

食品图像从 GloFLS 子网络端输入，经过卷积神经网络操作后，对某些卷积层（图中示意的是第二卷积层，第三卷积层和第四卷积层）的特征地图使用空间和通道注意力模块（Spatial-Channel Attention, SCA）[52]学习，SCA 模块可以从这些特征地图中获取更多的全局区分性特征。融合这些来自不同层的区分性特征就可以捕获不同类型全局性特征，例如这些全局特征可能是有关食品形状、食品颜色和食品纹理等。图中全局平均池（Global Average Pooling，GAP）作用与全连接层（Fully connected layer, FC）相似，它们都是把二维的特征地图转化成一维的特征向量，而 GAP 操作使用较少参数量，可以有效缓解过拟合发生，实验效果并不比 FC 差。本工作使用的融合方法是特征串接，即一个特征的尾部和另一个特征的首部串联起来。

LocFLS 子网络主要采用级联的空间变换模块（Spatial Transformer, ST）[53]来定位特征地图的不同局部区域（例如与原材料相关的区域）。LocFLS 子网络根据两层特征地图，来实现局部区域的定位。后层的特征地图通过 ST 模块找到局部区域，然后和前层特征地图计算定位局部区域在全局特征地图中的位置，这样就可以找到整个特征地图中多个判别性区域。接着再将每个区域通过 Inception 模块进行特征学习。同时，该方法是从多组特征地图中来学习不同层的判别性局部区域（本工作中是第一层与第二层、第二层与第三层、第三层与第四层这三组特征地图），最终多组特征地图和多个区域就构成了级联式网络。每组特征地图学到的多个局部特征通过 Maxpooling 操作，将这些不同区域的特征融合为统一

特征表示，于是得到了该组特征地图对应的局部特征表示。最后，将多组局部特征通过特征串接融合表示为该图像的局部特征。这里 GAP 操作和 GloFLS 子网络中一样，都是降低参数量和缓解过拟合。LocFLS 子网络中 Maxpooling 操作可以使来自不同位置的融合特征更具区分性。

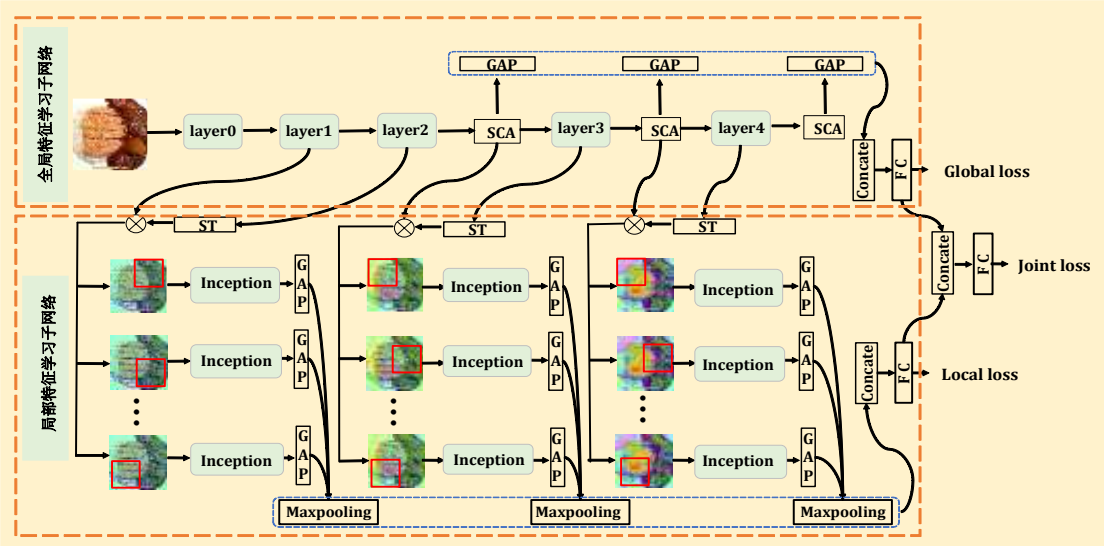


图 3. 2　本文 SMSMANet 网络框架图

图注：GAP 表示全局平均池化层，SCA 表示空间和通道注意力模块，ST 表示空间变换模块，FC 表示全连接层。

Figure 3.2 The proposed SMSMANet framework.

Note: GAP denotes Global Average Pooling layer, SCA denotes Spatial-Channel Attention, ST denotes Spatial Transformer, FC denotes Full-Connected layer.

最后，SMSMANet 网络结合了全局特征和局部特征来识别食品图像。此外，SMSMANet 网络还采用了一种多损失的学习方法，具有不同类型的损失，包括全局损失，局部损失和联合损失，SMSMANet 网络以端到端的方式进行培训，最大程度地发挥其判别性优势。下面将结合形式化公式详细讲述该框架。

3.3 全局特征学习子网络（GloFLS）

对于整张输入图像，GloFLS 子网络首先通过对多层特征地图进行 SCA 操作，来学习更多全局判别性特征，然后通过多层特征融合将不同层的判别性特征融合到全局级别表示中。由于从不同层次提取低层、中层和高层特征，因此 GloFLS 子网络可以捕捉到各种类型的全局特征，如食品图像形状、纹理和边缘线索等。

SCA 操作是将空间和通道上的注意力权重结合起来，可以从不同维度全面地捕捉判别性特征，它已经成功地应用于许多计算机视觉任务中，如图像描述[52]和行人再识别[54]。文献[52]认为卷积神经网络所提取的特征地图具有空间、通道和多层结构等属性，现有的一些深度学习方法主要考虑特征地图空间上的属性，其中主要运用注意力机制在空间上的注意力权重。因此，文献[52]充分利用卷积操作空间、通道和多层结构等属性，结合注意力机制的应用提出了 SCA 模块。在本文的工作中是为了来捕获食品图像中的全局判别性特征，将 SCA 模块应用到食物识别任务中。

SCA模块的输入是一个3-D张量$X^l \in R^{h\times w\times c}$，其中w表示宽度，h表示高度，c表示通道数，$l$表示 GloFLS 子网络的层。SCA 模块的学习目标是生成一个与 X 大小相同的显著性权值映射$A^l \in R^{h\times w\times c}$。与文献[54]一样，本工作采用分解方式进行 SCA 模块学习，如公式 3.1 所示：：

$$A^l = S^l \times C^l \quad \ldots(3.1)$$

其中，$S^l \in R^{h\times w\times 1}$和$C^l \in R^{1\times 1\times c}$分别表示空间和通道注意力映射。

通过 GAP 操作计算空间注意力，如公式 3.2 所示：

$$S^l = \frac{1}{c}\sum_{i=1}^{c} X^l_{1:h,1:w:i} \quad \ldots(3.2)$$

压缩和激励模块[55]的通道注意力计算，如公式 3.3 所示：

$$C^l = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w} X^l_{i,j,1:c}$$
$$C^l_{ex} = ReLU(W^{ca}_2 \times ReLU(W^{ca}_1 C^l)) \quad \ldots(3.3)$$

其中，$W^{ca}_1 \in R^{\frac{c}{r}\times c}$和$W^{ca}_2 \in R^{\frac{c}{r}\times c}$分别表示两个全连接层（FC）的参数矩阵，$r$ 表示瓶颈降低率。

多层特征融合是通过从多个层次提取显著性特征，得到低层、中层和高层的特征，包括各种类型全局特征，如纹理、形状和边缘信息等[56]，这些全局特征是食物识别的重要线索。因此，本工作将不同层次的判别特征融合到全局特征表示中用于食品识别。

3.4 局部特征学习子网络（**LocFLS**）

LocFLS 子网络可以定位具有不同位置和尺度的判别性区域，以获取判别性局部特征。对于每一层， LocFLS 子网络利用堆叠的空间变换模块（ST）来定位 T 个局部图像区域，并通过 Maxpooling 操作得到融合后的局部表示，本工作中 T 取值为4。在进行图像识别时，经典图像处理方法会考虑输入图像的平移不变性、缩放不变性、旋转不变性和聚焦局部性等来提高识别准确度，其实就是对图像裁剪、平移、缩放和旋转等操作。这些不变性的本质是对图像进行空间坐标变换，众所周知的一种空间变换就是仿射变换。文献[53]提出了一种空间变换网络模型，该网络中的 ST 模块不需要关键点标定，就可以根据任务需求自动学习参数来进行空间变换。这种模块对于特征地图空间类内差异较大的情况很适用。考虑到局部图像区域包含来自空间和通道注意力的特征地图，最大池化操作可以将这些特征融合为更具判别性的特征表示。LocFLS 子网络最终将不同层次融合的局部特征聚合到最终的局部特征统一表示中。

对于每一层，本文采用 ST 模块来定位潜在的 T 个区域，并通过变换矩阵将该区域进行注意力机制建模，如公式 3.4 所示：

$$A^l = \begin{bmatrix} s_h & 0 & t_x \\ 0 & s_w & t_y \end{bmatrix} \quad \ldots(3.4)$$

这样通过改变两个比例因子（s_h，s_w）和二维空间位置（t_x，t_y）进行图像裁剪、平移和各向同性缩放操作。

3.5 多损失学习

SMSMANet 网络是多损失训练的学习形式，它由三种类型的损失共同优化，即全局损失L_{Glo}，局部损失L_{Loc}和联合损失L_{Joi}，得到的最终损失函数L，如公式 3.5 所示：

$$\mathrm{L} = L_{Joi} + \gamma_1 L_{Glo} + \gamma_2 L_{Loc} \quad \ldots(3.5)$$

其中，γ_1和γ_2是平衡参数，三种损失均采用交叉熵分类损失函数。

这种多损失学习可以最大限度地提高全局特征和局部特征的互补性优势，以便增加特征融合后的判别能力。

3.6 小结

本文工作不仅完成了一个大规模食品图像数据集构建，而且还提出一种新的食品识别方法。本章讲述了本工作提出的食品识别方法：一种堆叠多尺度多注意力网络（SMSMANet）。该方法以图像为输入来进行食品识别，没有用到食品其他辅助信息（比如原材料信息、位置信息和 bounding box 信息等）。SMSMANet 网络是通过将全局特征学习子网络（GloFLS）取得的判别性全局特征和局部特征学习子网络（LocFLS）取得的判别性局部特征融合，从而进行食品图像识别。在 GloFLS 子网络中，本文运用 SCA 模块对多层特征地图在通道和特征地图空间上赋予不同的显著性权重，加强了某些对食品图像识别有利的空间和通道，从而得到全局的判别性特征；在 LocFLS 子网络中，本文使用 ST 模块从多组特征地图中找到多个判别性局部区域，最后对多组特征地图和多个局部区域的特征融合得到局部的判别性特征。最终，结合全局特征和局部特征来进行食品识别。在模型训练过程中，本文使用三种损失函数来控制调整参数，这三种损失是：全局损失、局部损失和联合损失。当这三种损失的组合值达到最优时，才反映出 SMSMANet 网络模型的训练效果达到了最好。本文 SMSMANet 网络的实验效果将会在下一章进行讲述，该方法将在两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）及本文大规模食品图像数据集 ISIA Food-500 上进行验证，以此证明本文所提识别方法的有效性。

第 4 章 实验与分析

在完成 ISIA Food-500 数据集构建和提出了一种堆叠多尺度多注意力网络（SMSMANet）后，作者将在本章进行实验验证。一方面，对多种食品识别方法及 SMSMANet 网络在 ISIA Food-500 食品图像数据集上进行实验，来验证 ISIA Food-500 食品图像数据集的效果。另一方面，将在 ISIA Food-500、ETHZ Food-101 和 VireoFood-172 数据集上进行广泛实验，证明 SMSMANet 网络对食品图像识别的有效性。

4.1 实验设置

本文 SMSMANet 网络是在 Pytorch 平台上实现。输入图像大小调整为 224*224 像素。模型训练采用随机梯度下降法进行优化，批量大小为 80，动量为 0.9。学习率最初设置为10^{-2}，以后每 30 个周期阶段除以 10。对于全局特征学习子网络（GloFLS），本工作选择 SENet-154 网络模型[55]作为骨干网络，其中压缩和挤压模块中瓶颈减少率 r=16；对于局部特征学习子网络（LocFLS），本工作选择简单的 Inception-B 单元作为基本构建模块。在 LocFLS 子网络每一层中，选取的判别性区域为 4 个，即 T=4。空间变换模块（ST）中比例固定为s_h=s_w=0.5；在多损失学习中，平衡参数γ_1=γ_2=0.5。

评价指标采用图像识别通用的准确率标准，即 Top-1 accuracy（Top-1 acc.）和 Top-5 accuracy（Top-5 acc.）进行评价，前 Top-n accuracy 是指排名前 n 的类别中正确判断结果的比率。

4.2 实验结果和分析

（1）对 ISIA Food-500 数据集实验评估

ISIA Food-500 数据集中的图片分别按照 60%，10%，30%比例随机分成训练集、验证集和测试集三部分，这三部分图片没有互相重复。进一步地，将测试集平分成公共测试集和私有测试集。公共测试集会和训练集、验证集共同发布公开，私有测试集计划用于未来大规模食品图像识别挑战赛提交结果评估。

1）在 ISIA Food-500 数据集上，本工作将 SMSMANet 网络与不同的食品识

别方法进行了实验比较，结果如表 4.1 所示。这些方法不仅包括各种典型的深度学习网络框架，如 VGG-16 网络[44]和 SENet-154 网络[55]，还包括一些最近被提出的细粒度图像识别方法，比如 NTS-NET 网络[57]和 WS-DAN 网络[58]。从表中实验结果可看出，本工作提出的 SMSMANet 网络性能无论是 Top-1 准确率还是 Top-5 准确率都比其他方法高。从关键指标 Top-1 准确率来看，相比于最好的基准方法 SENet-154[55]，SMSMANet 网络性能在公共测试集上高出 1.82%和在私有测试集上高出 2.65%，这表明本文 SMSMANet 网络联合全局特征和局部特征策略的优越性。此外，从表 4.1 实验结果可以看出，SENet-154 模型比其他网络模型有更高的食品图像识别准确率，因此本文选择 SENet-154 模型作为骨干网络来达到尽可能好的识别性能。根据混淆矩阵，本文在图 4.1 中样例展示了 SMSMANet 网络在 ISIA Food-500 食品图像数据集上预测的视觉上相似的 Top-5 类别和准确率。这些样例是作者从第 1 类开始按照等距 50 类选择的类别，其中 Aburaage 是 ISIA Food-500 数据集里的第 1 类，Zurcher_Geschnetzeltes 是 ISIA Food-500 数据集里的第 500 类。Æbleflæsk 是随机选取一类来保证作图补齐。图 4.1 可以帮助大家理解视觉属性相似的类别范畴，为后续再构建优质食品图像数据集和探讨食品识别特别方法提供参考。

表 4. 1　在 ISIA Food-500 数据集上的性能比较（%）

Table 4.1 Performance comparison on ISIA Food-500 dataset (%)

方法	公共测试集性能		私有测试集性能	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
VGG-16[44]	51.97	81.26	52.16	81.03
GoogLeNet[3]	53.24	81.59	53.09	81.32
ResNet-152[59]	52.94	81.59	52.94	81.41
WRN-50[60]	54.42	82.74	54.61	82.86
DenseNet-161[61]	55.21	83.19	55.05	82.83
NAS-NET[62]	57.25	84.64	57.42	84.53
SE-ResNeXt101 32x4d[55]	58.18	84.69	58.48	84.60
SENet-154[55]	59.94	87.02	60.05	86.82
NTS-NET[57]	54.31	81.93	54.48	82.21
WS-DAN[58]	57.40	84.84	58.12	85.16
SMSMANet(本文工作)	**61.76**	**87.58**	**62.70**	**87.79**

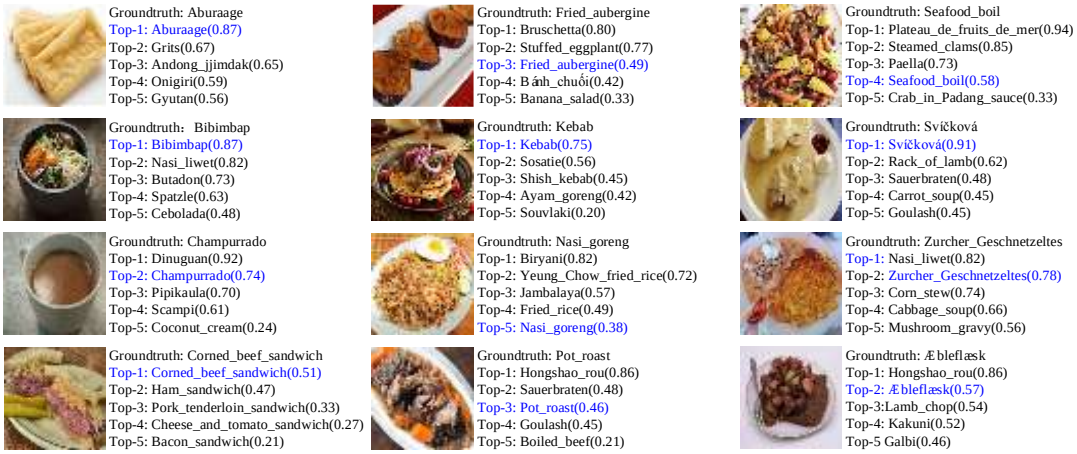


图 4. 1　SMSMANet 预测结果 Top-5 类别示例图

Figure 4.1 Top-5 category examples of SMSMANet prediction results

2）本工作通过消融实验评估融合全局特征和局部特征策略的效果。如表 4.2 所示：全局特征学习子网络（GloFLS）要比局部特征学习子网络（LocFLS）的性能高，这是因为全局特征包含了更多更全面的图像信息使模型学到更多判别性知识，而局部特征学习子网络只是选择性学习其中的部分局部特征，而不是所有的判别性区域，模型学到的关于食品的判别性知识相对较少，若仅仅依靠局部特征来进行食品识别，可能会陷入局部聚焦的盲区，忽略了其他部分重要的判别性信息，因此 LocFLS 的识别性能是这三者最低的。SMSMANet 网络融合了具有全面信息的全局特征和具有细节信息的局部特征，综合全局特征和局部特征的优势，故而在识别性能上有所增益。SMSMANet 网络在关键性能指标 Top-1 acc.上均高于 GloFLS 和 LocFLS；私有测试集上 GloFLS 的 Top-5 acc.性能比 SMSMANet 高，这是因为 LocFLS 性能较低从而拉低了 SMSMANet 性能。总体来说，SMSMANet 网络融合了图像的判别性全局信息和判别性局部信息，对食品图像识别性能提升有所帮助，也验证了联合全局特征与局部特征的互补性效果。

表 4. 2　在 ISIA Food-500 数据集上的消融实验结果（%）

Table 4.2 Ablation experiments on ISIA Food-500 dataset (%)

方法	公共测试集性能		私有测试集性能	
	Top-1 acc.	Top-5 acc.	Top-1 acc.	Top-5 acc.
GloFLS	61.73	87.42	62.45	**97.71**
LocFLS	59.92	86.43	60.66	86.65
SMSMANet	**61.76**	**87.58**	**62.70**	87.79

3）本工作在 SMSMANet 网络的三个不同特征层上对来自 GloFLS 子网络的 SCA 模块和来自 LocFLS 子网络的 ST 模块进行了可视化分析。如图 4.2 所示，

在 GloFLS 子网络中， SCA 模块在不同特征地图层捕捉不同的全局性特征，比如 Boiled beef 的食品形状信息和 Pumpkin bread 的食品纹理信息。同时，随着 SMSMANet 网络加深，SCA 模块捕获了更加集中和更加有判别性的全局信息；在 LocFLS 子网络中，ST 模块可以从不同特征地图层捕获背景较少的不同局部区域。总之，SMSMANet 网络联合 GloFLS 子网络和 LocFLS 子网络进行学习，既学到了判别性的全局信息又学到了判别性局部信息，图 4.2 从可视化图上验证了联合全局特征和局部特征学习的互补性效应。

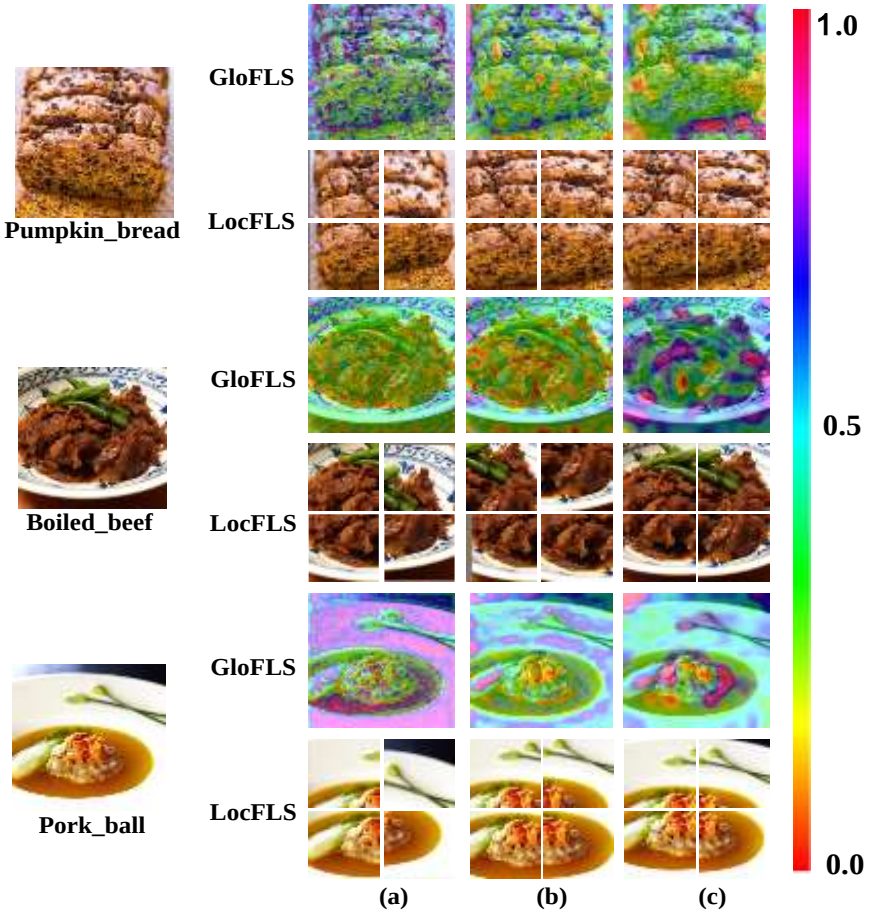


图 4. 2　GloFLS 中 SCA 模块和 LocFLS 中 ST 模块的可视化图

Figure 4. 2 Visualization of SCA in GloFLS and STs in LocFLS

4）作者从 ISIA Food-500 数据集中选择了 20 类来测试进行定性分析。本文列出在 Top-1 准确率上 10 个表现最好的类（如图 4.3 所示）和 10 个表现最差的类（如图 4.4 所示）。

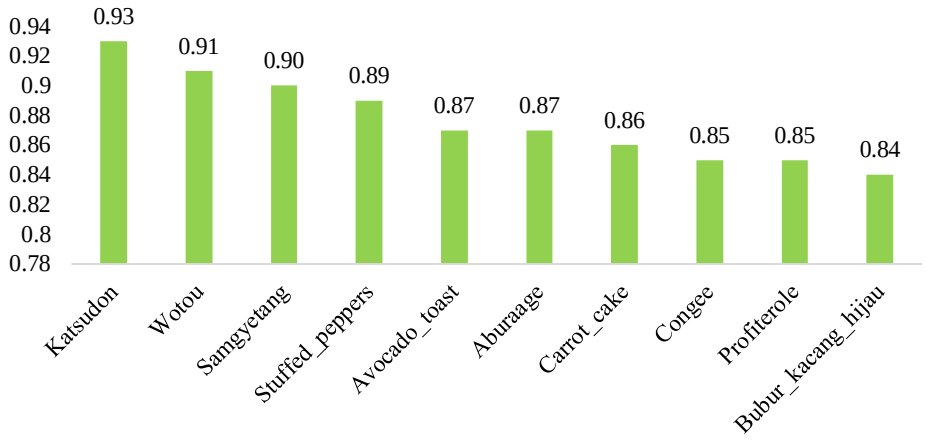


图 4. 3　选定分析类中 Top-1 准确率最前的 10 类结果图

Figure 4.3 The 10 best Top-1 accuracy categories of selected analysis categories

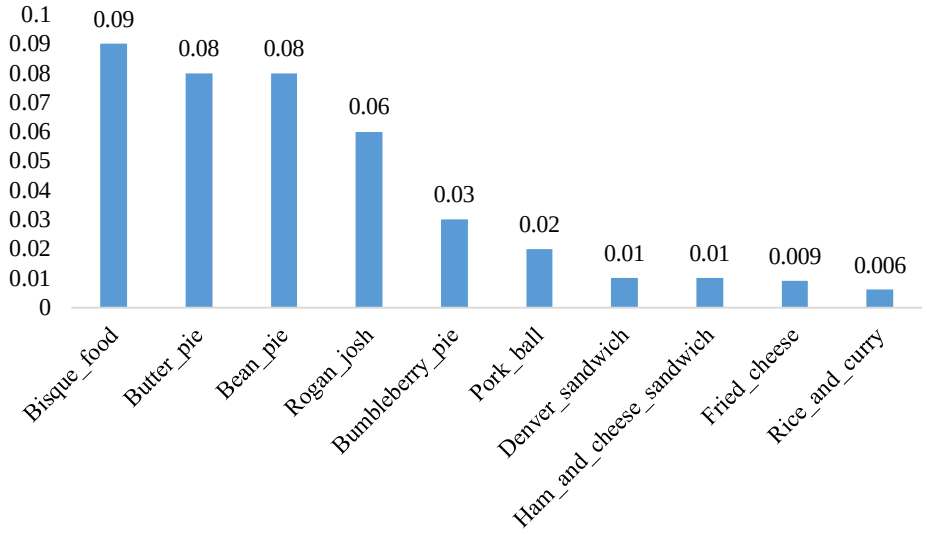


图 4. 4　选定分析类中 Top-1 准确率最后的 10 类结果图

Figure 4.4 The 10 worst Top-1 accuracy categories of selected analysis categories

从图4.3和图4.4中可以看到，有些类很容易识别，比如Katsudon和Wotou，它们的Top-1准确率都在90%以上；然而，有些食品类之间不容易识别，比如Bisque food 和 Butter pie，它们的Top-1准确率均在1%之下。为此，本文进一步分析不易识别的食品种类，作者挑选了3个Top-1准确率最差的类，并从其Top-5准确率的类中找到3个与之相似的类进行研究。从图4.5可以看出，类间差异太小是导致性能下降的主要原因。

这些实验评估和分析证明：（1）食品图像数据集比一般实物数据集（比如ImageNet和Places数据集）更具挑战性，ISIA Food-500是一个有挑战性的食品数据集；（2）现有的大多数基准识别方法在大规模食品图像数据集上进行识别的

效果是并不理想，识别准确率只有60%左右，而本文的SMSMANet网络性能高于这些基准方法，证明SMSMANet网络对食品识别的有效性。同时，大规模数据集的食品识别是一个需要持续研究和探讨的工作。

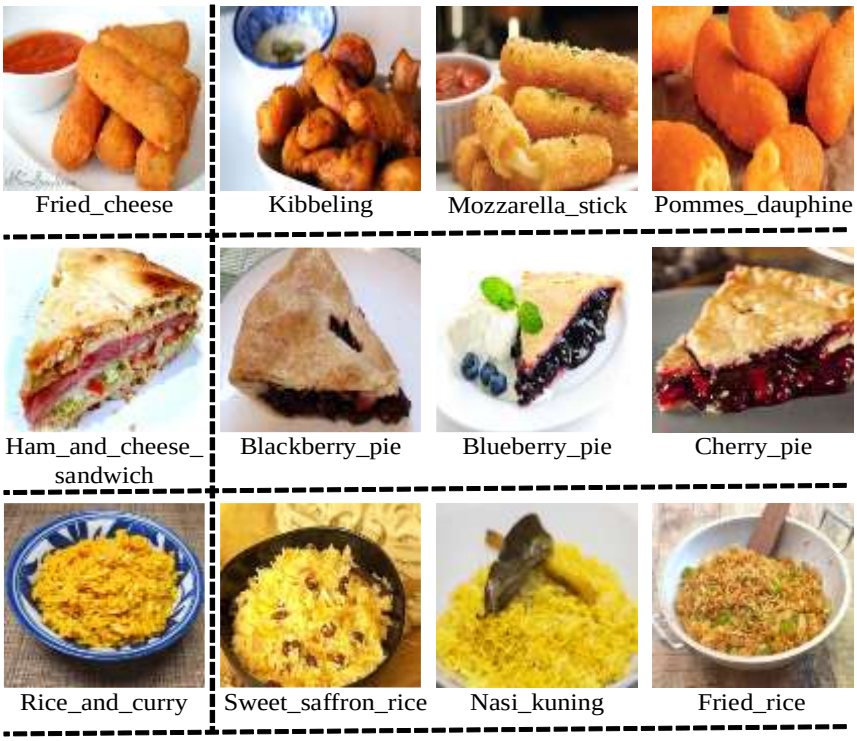


图 4.5　易混淆类展示

图注：其中第一列表示 3 个最差性能的类，然后每个类后是 3 个易混淆的类。

Figure 4.5 Confusing class display

Note: The first column denotes the 3 worst performing categories, followed by the three most confusing categories for each row.

（2）对 ETHZ Food-101 数据集实验评估

ETHZ Food-101数据集包含来自101类食品101,000张图像，每个类都是1000张图片。每个类的1000张图片都被分成750张训练图片和250张测试图片。本工作在ETHZ Food-101上实验了SMSMANet网络和现有的20个图像识别方法。表4.3所示：除了MSMVFA方法[5]、IG-CMAN方法[16]、WISeR方法[34]和Inception-Resnet-v2 SE方法[63]，本文提出的SMSMANet网络超出大部分基准方法。MSMVFA方法[5]和IG-CMAN方法[16]需要多阶段操作来提取特征，并且引入额外原材料信息作为监督信息，而本文提出的SMSMANet网络只使用图片，没有使用其他辅助信息。WISeR方法[34]采用了光度畸变和AlexNet型色彩增强。Inception-Resnet-v2 SE[63]方法使用了附加数据转移学习方法。而SMSMANet网

络没有使用这些数据增强或者附加数据的训练技巧。

同样，本工作对SMSMANet网络在ETHZ Food-101数据集上进行了消融实验，从表4.4可以看出，联合全局特征和局部特征的学习，对性能提升有一定的帮助。

表 4. 3　在 ETHZ Food-101 数据集上的性能比较（%）

Table 4.3 Performance comparison on ETHZ Food-101 dataset (%)

方法	Top-1 acc.	Top-5 acc.
AlexNet-CNN[2]	56.40	-
SELC[64]	55.89	-
ResNet-152+SVM-RBF[65]	64.98	-
DCNN-FOOD[66]	70.41	-
LMBM[67]	72.11	-
Ensemble Net[68]	72.12	91.61
GoogLeNet[3]	78.11	-
DeepFOOD[69]	77.40	93.70
ILSVRC[70]	79.20	94.11
WARN[71]	85.50	-
CNNs Fusion(I2)[72]	86.71	-
Inception V3[73]	88.28	96.88
SENet-154[55]	88.62	97.57
WRN[34]	88.72	97.92
SOTA[74]	90.00	-
DLA[75]	90.00	-
WISeR[34]	90.27	**98.71**
IG-CMAN[16]	90.37	98.42
Inception-Resnet-v2 SE[63]	90.40	-
MSMVFA[5]	**90.59**	98.25
SMSMANet(本文工作)	89.83	97.91

表 4. 4　在 ETHZ Food-101 数据集上的消融实验结果（%）

Table 4.4 Ablation experiments on ETHZ Food-101 dataset (%)

方法	公共测试集	
	Top-1 acc.	Top-5 acc.
GloFLS	89.61	**97.99**
LocFLS	88.89	97.78
SMSMANet	**89.83**	97.91

（3）对 Vireo Food-172 数据集实验评估

Vireo Food-172 数据集包含 172 类食品 110,241 张图片，每类图片数量不均衡，单类图片量在 191~1,061 张之间。每类数据按照 60%，10%，30%比例随机分成训练集、验证集和测试集。本工作选取了一些基准方法和 SMSMANet 网络在 Vireo Food-172 数据集上进行了实验，实验结果见表 4.5。从表 4.5 中可以看出，本文 SMSMANet 网络性能仅次于 IG-CMAN 方法[16]性能，原因是 IG-CMAN

方法[16]是多阶段学习而非端到端学习，并且引入了额外原材料辅助信息。本文的 SMSMANet 网络是端到端学习，且只有图片输入这一种模态。表 4.5 再次验证了 SMSMANet 网络的有效性。

作者对 SMSMANet 网络在 Vireo Food-172 数据集上也进行了消融实验，从表 4.6 可以得出一致性结论：融合全局特征和局部特征对 Top-1 性能精度始终有所改进，再次验证了联合 GloFLS 子网络和 LocFLS 子网络的互补性。

表 4.5　在 Vireo Food-172 数据集上的性能比较（%）

Table 4.5 Performance comparison on Vireo Food-172 dataset (%)

方法	Top-1 acc.	Top-5 acc.
AlexNet-CNN[2]	64.91	85.32
VGG-16[44]	80.41	94.59
DenseNet-161[61]	86.93	97.17
MTDCNN(VGG-16)[8]	82.06	95.88
MTDCNN(DenseNet-161)[8]	87.21	97.29
SENet-154[55]	88.71	97.74
IG-CMAN[16]	**90.63**	**98.40**
MSMVFA[5]	90.61	98.31
SMSMANet(本文工作)	89.77	97.97

表 4.6　在 Vireo Food-172 数据集上的消融实验结果（%）

Table 4.6 Ablation experiments on Vireo Food-172 dataset (%)

方法	公共测试集	
	Top-1 acc.	Top-5 acc.
GloFLS	89.62	**98.03**
LocFLS	89.14	97.89
SMSMANet	**89.77**	97.97

4.3 小结

本章主要是对本文提出的 ISIA Food-500 数据集和 SMSMANet 网络进行了实验验证。通过与多种基准方法比较，证明了 SMSMANet 网络将全局特征与局部特征联合学习的有效性。SMSMANet 网络在 ISIA Food-500 数据集上达到了识别性能最好。虽然，SMSMANet 网络在 ETHZ Food-101 数据集和 Vireo Food-172 数据集上的性能没有达到最好，但是也很接近目前最好性能。目前最好性能的 IG-CMAN 等方法是利用了额外原材料辅助信息和数据增加方式，然而本文 SMSMANet 网络是仅以图片为输入来进行模型训练学习，这证明了本文 SMSMANet 网络的有效性。同时，本工作还进行了消融实验，证明融合全局特征

和局部特征的策略要比单独使用全局特征和单独使用局部特征性能要好，从 Top-1 准确率数据得出，SMSMANet 网络中联合 GloFLS 子网络和 LocFLS 子网络的策略对食品识别性能提升是有帮助的。从 SMSMANet 网络的可视化分析看出，随着 SMSMANet 模型的卷积层不断加深， GloFLS 子网络捕获的全局信息更加集中和更加有判别性； LocFLS 子网络捕获的局部信息含有较少背景干扰信息。此外，本工作还对 ISIA Food-500 数据集做了定性分析，说明了 ISIA Food-500 数据集具有挑战性和大规模数据集的食品识别需要持续研究和探讨。

第 5 章 结论与展望

饮食健康开始受到越来越多人的关注，现实生活中对智能化食品识别的需求推动食品识别成为一个热门的研究方向。食品识别属于细粒度图像识别范畴，它与其他的一般实物识别有一定区别，不能生搬硬套其他识别技术。这主要是因为食品种类成千上万，很多食品图像数据类间差异小和类内差异大等特点与一般实物图像数据集（比如 ImageNet 和 Places 数据集）有很大不同。基于食品数据独特的属性，一般实物图像识别方法也不完全适用，而需要探索食品本身图像数据特性，找到适合食品自身的方法模型来进行食品识别。为此，本文主要从食品图像数据集和食品识别方法两个方面展开研究。

5.1 结论

（1）现有食品图像数据集中食品种类较少、图片量较少和覆盖范围有限，这无法满足大规模复杂食品识别统计模型构建的需求。现有的两个基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）只有一百多类和十几万张图片，这远远不能满意现实世界应用的需求。于是，本文工作欲构建一个大规模食品图像数据集，以便更好满足食品识别的需要。本文构建大规模食品图像数据集步骤主要包括：（a）数据集概念体系构建。本文构建了一个“食品范畴类-食品大类-食材子类-食品单类”的四层食品种类概念结构体系，用于指导 ISIA Food-500 数据集的食品种类选择。（b）数据收集。根据数据集概念体系中食品种类，使用爬虫程序从图片搜索引擎下载食品图片。（c）数据清洗。针对候选图片，本工作通过程序自动清洗和人工标注清洗噪音图片，最后进行数据检查。（d）数据扩充。作者通过专门的美食网站和把食品名称翻译成多国语言文字的方式再次爬虫来扩充图片。最终，本文构建了一个大规模食品图像数据集 ISIA Food-500，这个数据集包含 500 类食品 405,776 张图片，每类图片数据均在 500 张以上。ISIA Food-500 数据集具有数据种类多、数据量大、覆盖范围广和良好的可拓展性等优势和特点。

（2）本文提出了一种食品识别方法：堆叠尺度多注意力网络（SMSMANet），可以共同学习图像的全局特征和局部特征来进行食品识别。这个方法框架主要分

为两个部分：全局特征学习子网络（GloFLS）和局部特征学习子网络（LocFLS）。全局特征学习子网络（GloFLS）主要运用空间和通道注意力模块力（SCA）给特征地图空间和通道上分配显著性权重，让模型学到判别性全局特征表示。局部特征学习子网络（LocFLS）主要运用空间变换模块（ST）从特征地图中找到多个判别性局部区域来提取局部特征。最后，融合两个子网络提取的全局特征和局部特征来进行食品识别。

（3） 本文在 ISIA Food-500 数据集和两个已有的基准食品图像数据集（ETHZ Food-101 和 Vireo Food-172）上进行了充分实验，将提出的 SMSMANet 网络和多个食品识别方法进行了比较，最后表明本文 SMSMANet 网络的有效性和 ISIA Food-500 是一个大规模的、具有挑战性的食品图像数据集。

5.2 展望

虽然，本文工作在食品图像数据集和食品识别方法上取得了一定成果，但仍然有很多食品识别相关问题值得更多地研究，包括食品图像数据集和食品识别方法的思考。展望未来食品识别工作，作者认为可以从以下几个方面继续探究：

（1）食品数据集扩建：ISIA Food-500 数据集继续扩充，目标是 2,000 个食品种类、100 万张以上的图片量，届时将发布一个百万级的超大规模食品图像数据集，并尽力推动这个大规模食品图像数据集能成为一个新的基准食品数据集，成为食品数据集道路上的一个里程碑；对 ISIA Food-500 数据集不仅可以图片量和食品种类数扩充，还可以增加其他食品信息，比如食品原材料、烹饪方式和口味等，这样可以为更丰富、更多种模态食品识别方法的出现提供数据支持。

（2）融合食品多种属性的食品计算研究：在现实世界中，人们对美食会从色、香和味多个方面进行评价。因此，作者思考，“色”意为食品的样子，是指人眼看到的部分，其实从某种角度上可以把计算机视觉等同于人眼。同理，对于“香”，是指人鼻子闻到的气味，让计算机来运用气味知识（肉香味、臭味和酸气味等）辅助食品识别。气味信息可以用气味传感器获得或者人亲历闻然后标注。“味”是指食品的酸、甜、苦、辣和咸的口感，可以用味觉传感器获得数据，也可以根据原材料的味道，加上佐料的味道等进行推断合成。

（3）结合心理学的食品计算研究：在注重身体健康的同时，心理健康问题也颇受关注。大家都知道心情不好或者很好的时候会选择吃某类食品，比如有人

生气时吃辣，悲伤时吃糖，开心时吃烧烤等。这反映出食品可以调节心情的作用，食品的五颜六色、五味俱全和心情的喜怒哀乐之间存在微妙的关系。作者认为，或许将来会在食品或者心理学等领域会出现食品心情相关的研究，根据顾客心情，食品计算系统可以智能地推荐相关食品类型调节心情。

（4）结合多元文化的食品计算研究：饮食文化由来已久，饮食与文化交织在一起共同发展。在研究食品识别、食品推荐等食品计算问题时，借助民族文化、地域文化和气候条件等多种知识来丰富模型可以学到更多相关知识，从而提高模型计算性能，也是一个值得探索的方向。

（5）小样本食品计算研究：当下是大数据时代，此次人工智能兴起也得益于大数据基础，但是数据量指数式暴增似乎永远没有尽头。另一方面，并不是每种食品都有足够多图片量，比如有些菜只有某个地方人吃，并不大众流行，所以分享的图像数据资源可能比较少。面对这种情况，在大数据基础上训练出的食品识别模型可能就不受用。于是，近几年开始一些研究人员开始思考小样本食品图像识别技术，比如文献[76]提出在相对少量食品图像基础上，融合三元神经网络和关系网络来进行食品计算，也是一个有趣的研究方向。

总之，食品识别技术发展离不开食品数据集的构建，也离不开适合食品识别方法创新。通过不断提升食品识别模型的能力，可以有效解决食品识别在实际场景中广泛利用和效果提升问题。由此可见，针对食品识别数据集和识别方法的研究，无论是从科学研究方面，还是从实际应用方面，都具有极大探讨和研究价值。

参考文献

[1] Weiqing Min, Shuqiang Jiang, Linhu Liu, et al. A survey on food computing [J]. In ACM Computing Surveys, 2019, 52: 1–36.

[2] Lukas Bossard, Matthieu Guillaumin, Luc Van Gool. Food-101–mining discriminative components with random forests [C]. In European Conference on Computer Vision, 2014: 446–461.

[3] Austin Meyers, Nick Johnston, Vivek Rathod, et al. Im2calories: towards an automated mobile vision food diary [C]. In IEEE International Conference on Computer Vision, 2015: 1233–1241.

[4] Feng Zhou, Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1124–1133.

[5] Shuqiang Jiang, Weiqing Min, Linhu Liu, et al. Multi-scale multi-view deep feature aggregation for food recognition [J]. IEEE Transactions on Image Processing, 2020, 29(1): 265–276.

[6] Nitish Nag, Vaibhav Pandey, Ramesh Jain. Health multimedia: Lifestyle recommendations based on diverse observations [C]. In ACM International Conference on Multimedia Retrieval, 2017: 99–106.

[7] Eduardo Aguilar, Beatriz Remeseiro, Marc Bolanos, et al. Grab, pay and eat: Semantic food detection for smart restaurants [C]. IEEE Transactions on Multimedia, 2018, 20: 3266–3275.

[8] Jingjing Chen, Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval [C]. In ACM International Conference on Multimedia, 2016: 32–41.

[9] Jia Deng, Wei Dong, Richard Socher, et al. ImageNet: a Large-Scale Hierarchical Image Database[C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2009: 20–25.

[10] Weiqing Min, Shuqiang Jiang, Ramesh Jain. Food Recommendation: Framework, Existing Solutions and Challenges [J]. IEEE Transactions on Multimedia, 2019: 10.1109/TMM.2019.2958761.

[11] Weiqing Min, Bingkun Bao, Shuhuan Mei, et al. You Are What You Eat: Exploring Rich Recipe Information for Cross-Region Food Analysis [J]. IEEE Transactions on Multimedia, 2018, 20(4): 950–964.

[12] Luis Herranz, Shuqiang Jiang, Ruihan Xu. Modeling Restaurant Context for Food Recognition [J]. IEEE Transactions on Multimedia, 2017, 19(2): 430–440.

[13] Yuji Matsuda, Keiji Yanai. Multiple-food recognition considering cooccurrence employing manifold ranking [C]. In International Conference on Pattern Recognition, 2012: 2017–2020.

[14] Yoshiyuki Kawano, Keiji Yanai. Automatic expansion of a food image dataset leveraging

existing categories with domain adaptation [C]. In European Conference on Computer Vision, 2014: 3–17.

[15] Parneet Kaur, Karan Sikka, WeijunWang, et al. Foodx-251: A dataset for fine-grained food classification [J]. arXiv, 2019:1907.06167.

[16] Weiqing Min, Linhu Liu, Zhengdong Luo, et al. Ingredient-guided cascaded multi-attention network for food recognition [C]. In ACM International Conference on Multimedia, 2019: 1331–1339.

[17] Mei Chen, Kapil Dhingra, Wen Wu, et al. PFID: Pittsburgh fast-food image dataset [C]. In IEEE International Conference on Image Processing, 2009: 289–292.

[18] Javier Marin, Aritro Biswas, Ferda Ofli, 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, 2019: 10.1109/TPAMI.2019.2927476.

[19] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, et al. Learning cross-modal embeddings for cooking recipes and food images [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3020–3028.

[20] Jaclyn. Rich, Hamed. Haddadi, Timothy M. Hospedales. Towards bottom-up analysis of social food [C]. In International Conference on Digital Health Conference, 2016: 111–120.

[21] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, et al. R2GAN: Cross-modal recipe retrieval with generative adversarial network [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11477–11486.

[22] Saihui Hou, Yushan Feng, Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization [C]. In IEEE International Conference on Computer Vision, 2017: 541–549.

[23] Wikipedia [EB/OL]. [2018-11-09].https://en.wikipedia.org/wiki/Category:Lists of foods.

[24] James Ireland, Van Erp-Baart, Charrondiere, et al. Selection of a food classification system and a food composition database for future food consumption surveys [J]. European Journal of Clinical Nutrition, 2002, 56(s2):S33–S45.

[25] 陈倩，张志华，滕锦程，唐伟.国内外食品分类系统对绿色食品产品分类体系构建的借鉴 [J]. 中国食物与营养, 2017, 23(10): 11–14.

[26] CAC/MIST 4—1003, Classification of Foods amd Animal Feeds.

[27] 刘璇, 姚晗瑁, 章强华等. 日本“肯定列表制度”中食品分类体系研究 [J].农业质量标准, 2009, 5: 53–56.

[28] GB 2760--2014. 食品安全国家标准食品添加剂使用标准[S].

[29] GB 2763--2014. 食品安全国家标准食品中农药最大残留限量[S].

[30] CODEX STAN 192-1995, General Standard for Food Additives [S]. Food and Agriculture Organization of the United Nations, 2018.

[31] Shulin Yang, Mei Chen, Dean Pomerleau, et al. Food recognition using statistics of pairwise local features [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2010:

2249–2256.

[32] Taichi Joutou, Keiji Yanai. A food image recognition system with multiple kernel learning [C]. In IEEE International Conference on Image Processing, 2010: 285–288.

[33] Hokuto Kagaya, Kiyoharu Aizawa, Makoto Ogawa. Food detection and recognition using convolutional neural network [C]. In ACM International Conference on Multimedia, 2014: 1085–1088.

[34] Niki Martinel, Gian Luca Foresti, Christian Micheloni. Wide-slice residual networks for food recognition [C]. In IEEE Winter Conference on Applications of Computer Vision, 2018: 567–576.

[35] Shota Horiguchi, Sosuke Amano, Makoto Ogawa, et al. Personalized classifier for food image recognition [J]. IEEE Transactions on Multimedia, 2018, 20(10): 2836–2848.

[36] Hokuto Kagaya, Kiyoharu Aizawa, Makoto Ogawa. Food detection and recognition using convolutional neural network [C].In ACM International Conference on Multimedia, 2014: 1085–1088.

[37] Ashutosh Singla, Lin Yuan, Touradj Ebrahimi. Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model [C]. In ACM International Workshop on Multimedia Assisted Dietary Management, 2016: 3–11.

[38] 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, 2017, 9(6): 635–641.

[39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, et al. Places: A 10 million image database for scene recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452–1464.

[40] Christiane Fellbaum, George Miller. WordNet: An Electronic Lexical Database (Language, Speech, and Communication) [M]. A Bradford Book, 1998.

[41] Google Image [EB/OL]. [2018-11-13]. https://www.google.com.hk/imghp?hl=zh-CN&tab=wi&ogbl.

[42] Baidu Image [EB/OL]. [2018-11-13]. https://image.baidu.com/.

[43] Bing Image [EB/OL]. [2018-11-13]. https://m2.cn.bing.com/images/trending?form=Z9LH.

[44] Christian Szegedy, Wei Liu, Yangqing Jia, et al. Going deeper with convolutions [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1–9.

[45] Hajime Hoashi, Taichi Joutou, Keiji Yanai. Image recognition of 85 food categories by feature fusion [C]. In IEEE International Symposium on Multimedia, 2010: 296–301.

[46] Marios M Anthimopoulos, Lauro Gianola, Luca Scarnato, et al. A food recognition system for diabetic patients based on an optimized bag-of-features model [J]. IEEE Journal of Biomedical and Health Informatics, 2014, 18(4): 1261–1271.

[47] XinWang, Devinder Kumar, Nicolas Thome, et al. Recipe recognition with large multimodal food dataset [C]. In IEEE International Conference on Multimedia and Expo Workshops, 2015: 1–6.

[48] Gianluigi Ciocca, Paolo Napoletano, Raimondo Schettini. Food recognition and leftover estimation for daily diet monitoring [C]. In International Conference on Image Analysis and Processing, 2015: 334–341.

[49] Gianluigi Ciocca, Paolo Napoletano, Raimondo Schettini. Food recognition: a new dataset, experiments, and results [J]. IEEE Journal of Biomedical and Health Informatics, 2016, 21: 588–598.

[50] Xin Chen, Hua Zhou, Liang Diao. ChineseFoodNet: A large-scale image dataset for Chinese food recognition [J]. arXiv, 2017: 1705.02743.

[51] Vinay Bettadapura, Edison Thomaz, Aman Parnami, et al. Leveraging context to support automated food recognition in restaurants [C]. In IEEE Winter Conference on Applications of Computer Vision, 2015: 580–587.

[52] Long Chen, Hanwang Zhang, Jun Xiao, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6298–6306.

[53] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks [C]. In Advances in Neural Information Processing Systems 28, 2015:2017–2025.

[54] Wei Li, Xiatian Zhu, Shaogang Gong. Harmonious attention network for person re-identification [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2285–2294.

[55] Jie Hu, Li Shen, Gang Sun. Squeeze-and-excitation networks [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132–7141.

[56] Songfan Yang, Deva Ramanan. Multi-scale recognition with DAG-CNNs [C]. In IEEE International Conference on Computer Vision, 2015: 1215–1223.

[57] Ze Yang, Tiange Luo, Dong Wang, et al. Learning to Navigate for Fine-Grained Classification [C]. In European Conference on Computer Vision, 2018: 438–454.

[58] Tao Hu, Honggang Qi. See better before looking closer: Weakly Supervised Data Augmentation Network for fine-grained visual classification [J]. arXiv, 2019: 1901.09891.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep residual learning for image recognition [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.

[60] Zagoruyko Sergey, Komodakis Nikos. Wide residual networks [C]. In British Machine Vision Conference, 2016: 87.1–87.12.

[61] Gao Huang, Zhuang Liu, Laurens van der Maaten, et al. Densely connected convolutional networks [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261–2269.

[62] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, et al. Learning transferable architectures for scalable image recognition [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8697–8710.

[63] Cui Yin, Song Yang, Sun Chen, et al. Large scale fine-grained categorization and domain-specific transfer learning [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4109–4118.

[64] Niki Martinel, Claudio Piciarelli, Christian Micheloni. A supervised extreme learning committee for food recognition [J]. Computer Vision and Image Understanding, 2016, 148: 67–86.

[65] Patrick McAllister, Huiru Zheng, Raymond Bond, et al. Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets [J]. Computers in Biology and Medicine, 2018, 95:217–233.

[66] Keiji Yanai, Yoshiyuki Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning [C]. In IEEE International Conference on Multimedia and Expo Workshops, 2015: 1–6.

[67] Hui Wu, Michele Merler, Rosario Uceda-Sosa, et al. Learning to make better mistakes: Semantics-aware visual food recognition [C]. In ACM Multimedia Conference, 2016: 172–176.

[68] Paritosh Pandey, Akella Deepthi, Bappaditya Mandal, et al. FoodNet: Recognizing foods using ensemble of deep networks [J]. IEEE Signal Processing Letters, 2017, 24: 1758–1762.

[69] Chang Liu, Yu Cao, Yan Luo, et al. Deepfood: Deep learning based food image recognition for computer-aided dietary assessment [C]. In International Conference on Smart Homes and Health Telematics, 2016: 37–48.

[70] Marc Bolanos, Petia Radeva. Simultaneous food localization and recognition [C]. In International Conference on Pattern Recognition, 2017: 3140–3145.

[71] Pau Rodr´ıguez L´opez, Diego Velazquez Dorta, Guillem Cucurull Preixens, et al. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition [J]. IEEE Transactions on Multimedia, 2020, 22: 502-514.

[72] Eduardo Aguilar, Marc Bolanos, Petia Radeva. Food recognition using fusion of classifiers based on CNNs [C]. In International Conference on Image Analysis and Processing, 2017: 213–224.

[73] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, et al. Food image recognition using very deep convolutional networks [C]. In ACM International Workshop on Multimedia Assisted Dietary Management, 2016: 41–49.

[74] Simon Kornblith, Jonathon Shlens, Quoc Le. Do betterImageNet models transfer better? [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2019: 2661–2671.

[75] Fisher Yu, Dequan Wang, Evan Shelhamer, et al. Deep layer aggregation [C]. In IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2403–2412.

[76] 吕永强, 闵巍庆, 段华等. 融合三元卷积神经网络与关系网络的小样本食品图像识别[J]. 计算机科学, 2020, 47(1): 136–143.

致 谢

岁月匆匆，春夏秋冬。蓦然回首间硕士生活即将结束，三年日月星辰的研究生道路上，自己已从一个机械相关的学士转变成一个计算机专业的硕士，无论是科研上的进步，还是生活上的成长，这一步步的蜕变都离不开老师的指导、同学的帮助、朋友的鼓励和家人的支持。在这论文完成之际，谨借此处向大家致以最诚挚的感谢。

首先，由衷地感谢我的导师蒋树强研究员和闵巍庆副研究员。“师者，所以传道受业解惑也”，两位导师耐心细致的带我进入计算机视觉领域，指导我从零起步开始食品计算的相关学习研究，培养我认真踏实的科研习惯，帮助我学习和生活的方方面面。两位导师精深的专业知识、勤恳的工作方式和严谨的科研态度是我终身学习的榜样。

感谢人工智能学院的刘芳老师、李芳老师和付静老师等在学业上给予的无微不至地辅导与关怀。

感谢曾在 ISIA 组一起学习的师兄弟们：宋新航、朱耀辉、朱永清、陈恭巍、郭晓倩、张天予、杨嘉豪、李润泽、刘林虎、王致岭、王静、张延超、李辉、李卓、王博瀚、刘晨龙、曾海涛、张思贤、朱瑛、梅舒欢、孙健、刘培、梁思斯、王轶和吕永强等，感谢大家对我学习上的帮助，共同相处使我感受到科研生活的快乐。

感谢同班同学：董俊川、徐梦洋、李丹峰、张超、王晨浩、李捷、李佳妮、张莲莲、陈军航、荆秀秀、潘唯一、郑群、李凌云和潘浩楠等，大家不仅在学习上相互帮助，还使我在学习之余体验到轻松愉快的生活。

感谢我的弟弟罗正斐和我的父母对于我这么大年纪还读研的无条件支持，是家人给我提供了稳固的大后方，让我在一次次坎坷面前勇往直前。

感谢饭友王晨浩同学和黎向阳师兄，是他们无论怎样的严寒酷暑都带我去食堂吃饭，感谢朱耀辉、朱永清和陈恭巍等师兄经常带我吃宵夜……这些饭友成功地让我在三年的硕士生活中拥有了壮硕的身体。

再次感谢食品计算小组的同学：刘林虎、梅舒欢、吕永强、李卓、李辉、杨嘉豪、王致岭、刘春霖、张天予、张延超、朱瑛、王静、刘宇昕、程宝艳和侯强

等在本文数据集构建方面提供的帮助。

种一棵树最好的时间是十年前，其次是现在，三年的研究生时光美好而又短暂，祝福这三年里在学习、生活和校招中认识的所有人健康快乐！也祝愿我们组的食品计算研究再接再厉，不断进步！

罗正东

2020 年 3 月于北京

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历：

2009 年 9 月——2013 年 6 月，在合肥工业大学大学机械与汽车工程学院获得工业工程学士学位。

2017 年 9 月——2020 年 6 月，在中国科学院大学人工智能学院攻读计算机技术硕士学位。

获奖情况：

中国科学院大学人工智能学院 2017-2018 学年“三好学生”。

工作经历：

2013 年 7 月——2016 年 4 月，在合肥京东方显示光源有限公司工艺技术部从事工业工程师。

2016 年 5 月——2017 年 7 月，在北京延锋海纳川汽车饰件有限公司总部运作管理部从事持续改进工程师。

已发表（或正式接受）的学术论文：

[1] Weiqing Min, Linhu Liu, **Zhengdong Luo**, and Shuqiang Jiang. Ingredient-guided cascaded multi-attention network for food recognition [C]. In ACM International Conference on Multimedia, 2019: 1331–1339.

[2] Shuqiang Jiang, Weiqing Min, Linhu Liu, and **Zhengdong Luo**. Multi-scale multi-view deep feature aggregation for food recognition [J]. IEEE Transactions on Image Processing, 2020, 29(1): 265–276.

申请或已获得的专利：

[1] 蒋树强, 罗正东, 闵巍庆. 一种多尺度融合的食品图像分类模型训练及图像分类方法. 中国, 申请号: 2019113737606(申请受理中).