中国科学院大学
University of Chinese Academy of Sciences

硕士学位论文

基于食材信息的食品图像识别方法研究

作者姓名：刘林虎

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

Ingredient Based Food 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

Liu Linhu

Supervisor: Professor Jiang Shuqiang

Associate Professor Min Weiqing

School of Artificial Intelligence

University of Chinese Academy of Sciences

June 2020

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

食物对人类生活至关重要，是人们生活的基础。随着社交网络、移动网络和物联网的快速发展，人们通常会上传、分享、记录食品图像、食谱、烹饪视频和饮食记录，这样可以轻松获得大规模的食品数据。研究人员可以利用这些食品数据在食品图像领域做各种研究，例如食品图像识别、食品检索等。食品图像识别是开展食品推荐、检索等领域的研究基础，并且有着广泛的实际应用，如多模态的食品记录和个性化的医疗健康等，因此食品图像识别越来越受到关注。

现有方法大多数都是使用深度卷积神经网络（Convolutional Neural Network, CNN）直接提取整张图像的视觉特征来进行食品图像识别，而没有考虑食品图像自身的特点。不同于一般的物体图像，食品图像通常不具有独特的空间布局，而且没有共同语义部分。因此，直接使用 CNN 很难捕获食品图像判别性信息。随着移动互联网的发展，用户不仅上传大量的食品照片，而且提供丰富的食材信息，就像物体对场景的重要性一样，食品图像中的食材对于食品识别同样非常重要。而且许多研究结果表明使用语义上有意义的食材可以作为食品图像识别的属性信息，它从不同的视角和粒度提供互补性信息来提高食品图像的识别性能。此外，尽管食品图像通常不具有明显的空间排列，但是可以使用不同尺度的食品图像块（Patch）融合成多尺度的特征表示。这样的表示可以将 Patch 特征从粗粒度尺度融合到细粒度尺度，因此它们的特征含有具有判别性图像区域的信息。而且这样多尺度融合可以对食品图像的几何变形变得更加鲁棒。因此，本文基于食品图像食材信息开展食品图像识别的研究，主要研究内容和贡献如下：

（1）本文提出了一种多尺度多视角特征融合（Multi-Scale Multi-View Feature Aggregation, MSMVFA）方法来进行食品图像识别。本文使用食材信息微调 CNN 来提取中层属性特征，从类别信息监督的 CNN 中提取高层语义特征和深层视觉特征。MSMVFA 可以对这三种类型的特征进行多尺度融合，并对具有不同粒度的各种类型特征进行多视角融合，以此产生更具区分性的细粒度特征表示。

（2）本文提出了一种食材指导的级联多注意力网络（Ingredient-Guided Cascaded Multi-Attention Network，IG-CMAN）来进行食品图像识别，IG-CMAN 能够以粗粒度到细粒度的多尺度方式，从类别信息和食材信息监督的子网络中顺

序定位到多个食品图像区域。这些在不同信息监督下生成的区域特征是非常互补的，融合这些区域特征可以形成更全面、更具区分性的特征表示。

（3）本文构建了一个与现有食品数据集非常互补并且含有食材的新食品图像数据集。该数据集包含 Wikipedia 列表中 200 种食品、大约 200,000 张食品图像和 319 种食材。它可以进一步推动食品图像识别领域的发展。

关键词：食品图像识别，卷积神经网络，食材信息，多尺度，多视角

Abstract

Food is very essential for human life and it is fundamental to the human experience. With the rapid development of social networks, mobile networks, and Internet of Things (IoT), people commonly upload, share, and record food images, recipes, cooking videos, and food diaries, leading to large-scale food data. Researchers can use these food data to do various research in the food filed such as food recognition, food retrieval, and so on. Food recognition is the basis of research in the food field and gained more attention in the many communities due to its various applications, e.g., multimodal foodlog and personalized healthcare.

Most of existing methods directly extract visual features of the whole image using popular deep networks for food recognition without considering its own characteristics. Compared with other types of object images, food images generally do not exhibit distinctive spatial arrangement and common semantic patterns, and thus are very hard to capture discriminative information. With the development of the mobile internet, users not only upload a large number of food photos but also provide the ingredient information. Just like the importance of objects to the scene, ingredients within food images are also very important for food recognition. Moreover, many research results proved that using semantically meaningful food ingredients can be used as attribute information for food image recognition. It provides complementary information from different perspectives and granularities to improve the recognition performance of food images. Furthermore, although food typically does not exhibit distinctive spatial arrangement, we can explore image patches from different scales and then fuse them into multi-scale representation. Such representation can fuse patch features from the coarse scale to the fine scale, and thus their features contain information from discriminative image regions. In addition, multi-scale fusion can be more robust to the geometrical deformation. Therefore, in this thesis, we make research on food recognition based on food image ingredient information. The main research contents and contributions are as follows:

(1) This thesis proposes a Multi-Scale Multi-View Feature Aggregation (MSMVFA) scheme for food recognition. We utilize additional ingredient information to fine-tune the deep network to extract mid-level attribute features. The high-level semantic features and deep visual features are extracted from class-supervised deep

neural network. MSMVFA can conduct two-level fusion, namely multi-scale fusion for each type of features and multi-view aggregation for various types of features with different granularity to produce more robust, discriminative and comprehensive fine-grained representation.

(2) This thesis proposes an Ingredient-Guided Cascaded Multi-Attention Network (IG-CMAN) for food recognition, which is capable of sequentially localizing multiple informative image regions with multi-scale from category-level to ingredient-level guidance in a coarse-to-fine manner. These regional features generated from the network under the supervision from different granularity are very complementary. Therefore, integrating diverse regional features can lead to more comprehensive and discriminative representation

(3) This thesis presents a new food dataset, which is very complementary to existing datasets for food recognition with ingredients. It contains 200 food categories from the list in the Wikipedia, about 200,000 food images and 319 ingredients. It will be made publicly available to further the development of scalable food recognition.

Key Words: Food recognition, Convolutional neural network, Ingredient information, Multi-scale, Multi-view

目 录

图目录

表目录

第 1 章　绪论

1.1　研究背景与意义

食物对人类生活、健康具有深远影响，但是现在越来越多的人却变得超重或肥胖。根据世界卫生组织的数据，年龄超过 18 岁的超重成年人超过 19 亿，其中有 6.5 亿以上的肥胖者。2016 年，全世界的肥胖率几乎是 1975 年的三倍。许多研究人员已经证实超重和肥胖是各种慢性疾病（如糖尿病和心血管疾病）的主要危险因素。据估计，2015 年全球有 4.15 亿人患有糖尿病。一个重要的原因是许多人普遍过着不健康的生活方式和不良的饮食习惯，比如高能量和高脂肪食物摄入量的增加[1]。食物不仅仅是生活的基础，它在定义我们的身份、社会地位、宗教意义和文化方面同样起着重要作用[2]，正如 Jean Anthelme Brillat-Savarin 所说：“告诉我吃什么，我会告诉你你是谁。”而且，我们如何烹饪和食用也是我们个人文化传承的深远影响因素。因此，与食物相关的研究[3,4]一直是一个研究热点，同时也受到各个研究领域的广泛关注。

早些年，研究人员从不同方面进行了与食物相关的研究，例如食物选择[5]、食物感知[6]、食品消费[7]、食品安全[8]和食品文化[9]。但是，这些研究基本上都是使用传统方法进行的。而且大多数方法都使用小规模的数据，比如问卷和食谱等。如今，社交网络、移动网络和物联网等各种网络的快速发展，用户可以通过这些网络共享食物图像、食谱、烹饪视频或记录食物的日记，从而形成大规模的食品数据集。这些食物数据意味着丰富的知识，它们可以为与食物相关的研究提供巨大的机会，例如发现食物感知[10]的原理、分析烹饪习惯[11]并控制饮食[4]。现将不同来源的食品数据分为以下几种类型：（1）食谱：食谱包含一组食物食材和烹饪操作方法。在较早的研究中，食谱是从烹饪书中收集，并手动输入到计算机中。现在，研究人员可以从许多烹饪网站上收集食谱，例如 epicurious.com 和 Allrecipes.com。我们可以直接将食谱嵌到网络的潜在空间进行食谱分析和其他任务。（2）食品图像：食品图像由于其视觉信息和语义内容成为最常见的多媒体数据。它们包含具有类别的食物图像，我们可以通过现有的深度学习方法来提取有意义的概念和信息来完成各种食品任务。大多数任务是对单一食物进行食品图像的视觉分析。还有一些是针对有多个食物的食品图像任务。（3）

烹饪视频：现在有很多的烹饪视频可以指导人们做饭。它们包含人们的烹饪操作和烹饪顺序信息。（4）食物属性：食物具有各种各样的属性，例如风味、味道、气味、食材、烹饪和切法等属性。我们可以使用丰富的食物属性来提高食品图像识别性能和其他任务。（5）食物日志：食物日志记录了食物图像、文本和卡路里等信息。随着移动技术和应用的快速发展，我们可以使用 FoodLog 这个 App 来保持健康饮食。（6）与餐厅有关的食物信息：现在越来越多的研究使用了餐厅有关的信息来识别食品图像。它们包括餐厅菜单和餐厅 GPS 信息。（7）健康：由于人们生活水平的提高，越来越多的人关注健康。健康包含丰富的信息，例如热量和营养。不健康生活方式和不良饮食习惯会引发超重、肥胖和其他疾病。研究人员可以利用食物的健康信息来研究饮食习惯。(8)其他食品数据:包括烹饪书、政府数据和调查表等。

随着技术的发展，提出了各种新的数据分析方法，包括网络分析、计算机视觉、机器学习和数据挖掘。人工智能技术的最新突破，例如深度学习[12]，进一步激发了人们对大规模食品相关研究的兴趣。食品图像识别是开展食品推荐、检索等领域的研究基础，一旦我们识别到食物的类别或者食材，我们可以进一步进行各种健康相关的分析，例如卡路里摄入量的估计、营养分析和饮食习惯分析。而且，食品图像识别有着很广泛的实际应用，比如在自助餐厅，食品识别不仅可以监控食品消耗，而且可以自动地对客户用餐进行计费。人们也可以通过简单拍照来更好的了解他们不熟悉或以前从未见过的食物，并了解其细节，比如烹饪方法、食材、口味等。因此，食品图像识别研究有着很重要的实际意义。

1.2 技术挑战

现在大多数食品图像识别的方法都是仅仅使用 CNN 直接提取整张图像的视觉特征，并没有考虑食品图像自身的特点。食品图像识别和细粒度图像识别一样，最主要的关键点是提取最具有判别性特征。然而，食品图像却面临不同于细粒度图像的技术挑战。

（1）现实生活中的食品图像往往包含各种各样的复杂背景信息。如图 1.1 所示，图片中除了我们要识别的汉堡包外，还有饮料、薯条等其他食品背景信息，这些复杂背景信息影响了汉堡包的识别性能。

图 1.1 食品图像的样例展示

Figure 1.1 Some food samples

（2） 食品图像识别与一般的物体识别任务不同，许多类型的食品不具有独特的空间布局。如图 1.2 所示，食品图像通常是非刚性的，而且结构信息不容易被利用，可能会随着烹饪方法和切法而改变。比如青椒土豆，土豆切成块和切成丝，在视觉上就存在明显的差异性。因此，标准的物体识别方法在食品图像上识别性能不佳。现有方法仅限于具有某些视觉上独特空间布局的食物类型，例如垂直结构（比如汉堡包、蛋糕）。

图 1.2 食品图像非刚性样例

Figure 1.2 Non-rigid structures of food images

（3）如图 1.3 所示，同一类别中的食品图像可能具有明显的差异性，而不同类之间可能具有相似性。

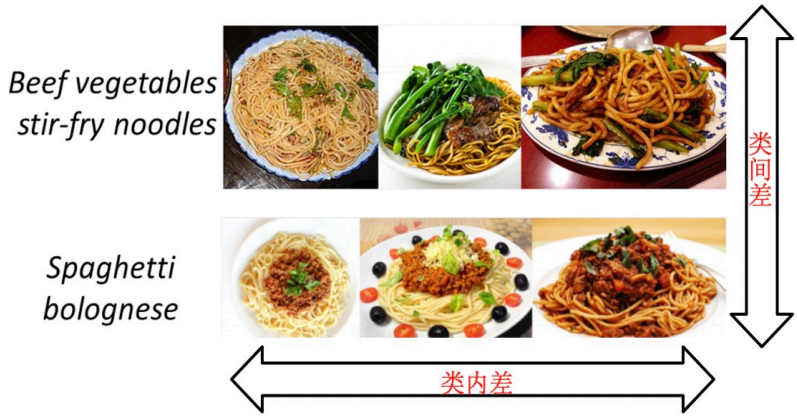


图 1.3 食品图像存在大的类间差和小的类内差

Figure 1.3 Food images with large inter-class variance and small intra-class variance

（4）不同于一般的物体识别，食品图像还具有各种几何变形，例如不同的视角，旋转和比例，如图 1.4 所示。它要求食品识别方法应具有几何不变性来识别食品图像。现有食品图像识别方法通常使用 CNN 直接从整张食品图像中提取视觉特征，并且没有考虑其几何不变性。这是因为 CNN 只能通过最大池化（Maxpooling）来处理具有小规模变形的图像。

图 1.4 一些具有不同几何变形的食品图像

Figure 1.4 Some food images with different geometrical deformations

（5）食品图像识别属于细粒度图像识别，细粒度图像识别[14][15]的第一步通常是发现某些主体的固定语义部分。但是，许多类型的食品图像中不存在共同的语义部分，因此，很难通过现有细粒度方法从食品图像中捕获语义信息。

（6）缺乏具有许多类别的大规模食品图像数据集。在计算机视觉中，大规模 ImageNet 数据集极大地促进了物体识别的发展。同样，发展食品图像识别，

也需要大规模的食品图像数据集。目前为止，确实存在一些标准食品数据集，例如 ETH Food101[13]和 VireoFood172[78]。但是，与 ImageNet 相比，这些数据集的类别和图像数量还不够。而且面向食品的数据集构建也面临着特殊的挑战。例如，由于区域差异，同一道菜可能有几种不同的名称。同样，有些食品被标记为相同的食品名称，但实际上属于具有不同食材的不同食品。这意味着很难像 ImageNet 这样根据食品名称来建立标准的数据集。

1.3　本文的研究内容与主要贡献

1.3.1　本文的研究内容

虽然食品图像识别存在一定的挑战，但是随着移动互联网的发展，用户不仅上传大量的食品照片，而且提供丰富的食材信息，如图 1.5 所示。就像物体对场景的重要性一样，本文作者认为使用语义上有意义的食材可以作为食品图像识别的属性信息，从而提高食品图像的识别性能。

因此，本文基于食品图像食材信息开展食品图像识别的研究。本文主要研究内容有以下 3 点：

（1）虽然食品通常不具有独特的空间排列，但可以尝试探索不同尺度的食品 Patch 图像，然后将它们融合成多尺度的特征表示。这种表示可以将特征从粗尺度融合到精细尺度，因此它们的特征包含来自判别性区域的信息。所以，多尺度融合可以对几何变形更加鲁棒。食品图像所特有的食材属性学习可以有助于食品图像识别，但是除了食材属性表示以外，CNN 的高层食品语义分布和深层视觉特征还可以从不同的视角和粒度提供互补信息。如果将这三种类型的特征融合在一起，可以最大可能地从食品图像中捕获语义信息，以应对食品图像的非刚性结构。所以，本文研究如何使用多尺度多视角来融合食品图像的特征，以提高食品图像识别性能。

图 1.5 一些包含食材信息的食品图像样例

Figure 1.5 Some food samples with rich ingredients

（2）食品图像识别属于细粒度图像识别，这些细粒度识别方法[14][15]大多数使用弱监督的方式在类别信息指导下，通过深层网络找到多个判别性区域，这些细粒度方法通常是发现某些主体的固定语义部分，例如鸟类和汽车。但是，许多类型的食品图像中不存在共同的语义部分，因此，可以使用食品图像特有的食材信息来明确地指导网络在不同粒度图像尺度上发现不同的语义区域。这些在不同信息监督下生成的区域特征是非常互补的，融合这些区域特征可以形成更全面、更具区分性的特征表示形式。因此，本文研究如何使用食品图像特有的食材信息来明确地指导网络在不同尺度图像上发现不同的语义区域。

（3）现在食品图像识别领域缺乏具有许多类别的大规模标准食品图像数据集，比如 ETH Food-101[13]和 VireoFood172[78]数据集在类别和图片数量上远远不够，而且它们也仅仅是专注于某个国家的菜品。构建数据集应该考虑到食品图像的地理分布，例如不同的美食，以覆盖整个世界。另外每个地区都有自己的特色美食和菜肴。现有研究工作已经证明融入食品图像上下文信息有助于识别任务。因此，本文进一步研究如何去构建一个包含食材信息的新食品图像数据集。

1.3.2　本文的主要贡献

本文主要贡献可归纳如下：

（1）本文提出了一种多尺度多视角特征融合（MSMVFA）方法来进行食品图像识别。本文将食品图像和食材信息相结合，从食材信息监督的食材网络提取中层属性特征，从类别信息监督的类别网络提取高层语义特征和深层视觉特征。MSMVFA 对每种类型的特征进行多尺度融合，并对具有不同粒度的各种类型特征进行多视角融合，以此产生更强大、更具区分性和更全面的细粒度特征表示。

（2）本文提出了一种食材指导的级联多注意力网络（IG-CMAN）来进行食品图像识别。该方法能够在食品图像特有的食材信息监督下，明确地指导网络在不同尺度图像上发现不同的语义区域。这些在不同信息监督下生成的区域特征是非常互补的，融合这些区域特征可以形成更全面，更具区分性的特征表示形式。

（3）本文作者在多个标准食品图像数据集上进行全面的实验评估，实验结果表明本文所提方法在当时达到了最好的识别性能。

（4）本文构建了一个新的食品数据集 ISIA Food-200[1]，该数据集包含 Wikipedia 列表中的 200 种食品、大约 200,000 张食品图像和 319 种食材。它与现有数据集非常互补，可以进一步推动食品图像识别领域的发展。

1.4　论文结构安排

论文具体研究工作在第 3 章和第 4 章中分别进行了阐述，其之间的关系如图 1.6 所示。论文各个章节内容安排如下：

第 2 章首先回顾了食品图像识别领域国内外研究现状。大部分研究表明 CNN 提取的深度特征比传统的手工特征（Hand-crafted Features）更具有判别性。因此，大多数食品图像识别的方法都是使用 CNN 直接提取整张图像的视觉特征进行食品图像的识别，但是这些研究工作并没有考虑到食品图像自身的特点，比如食品图像不具有独特的空间布局、食品图像没有共同语义部分等。

第 3 章提出了一种多尺度多视角特征融合方法。不同于一般物体图像，食品图像通常不具有独特的空间排列，而且还具有几何变形。基于以上两个特点，本文提出了一种多尺度多视角特征融合方法来进行食品图像识别。它能基于食材信

[1] https://github.com/minweiqing/Ingredient-Guided-Cascaded-Multi-Attention-Network-for-Food-Recognition

息，把三种不同类型的特征融合成统一的特征表示，而且实验结果验证了食材信息可以从另外一个视角来描述食品，并提供互补性特征来提高识别性能。

第 4 章提出了一种食材指导的级联多注意力网络。第三章中提出的方案预先使用了固定的 Patch，这 Patch 可能会包含一定的噪音，因此固定的 Patch 来表示判别性的区域不是最佳的选择。而且食品图像还具有不同的语义部分，因此基于第三章的实验结果，本文提出了一种食材指导的级联多注意力网络来进行食品图像识别。通过类别信息监督和食材信息监督的指导，它能够顺序定位到不同尺度图像上的不同注意力区域。该方法首先从类别信息监督的注意力子网中生成初始粗粒度的注意力区域，该区域能够去除原始图像一些复杂的背景信息。基于这个初始粗粒度的注意力区域，食材信息监督的注意力子网络能够找到食材对应的细粒度区域。最后，将这些区域的特征融合成最终的特征表示，用 Softmax 分类器进行食品图像识别。实验结果表明不同尺度的区域具有互补性，而且通过可视化分析，生成的食材细粒度区域还具有可解释性。

第 5 章对全文进行总结，并对未来的研究工作进行了展望。

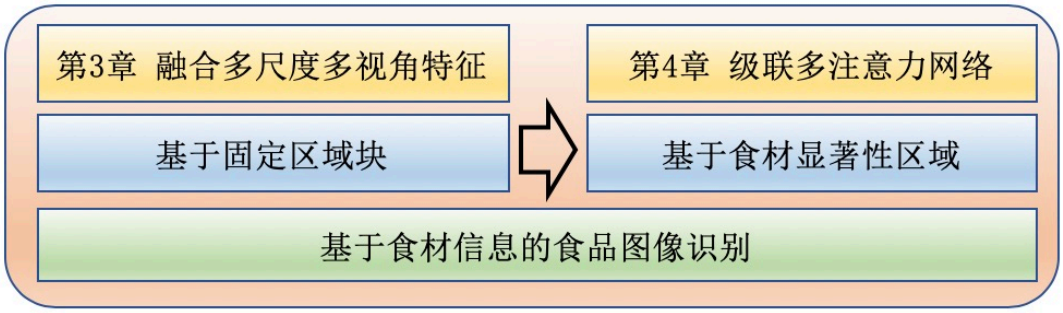


图 1. 6 两个研究工作之间的联系

Figure 1.6 The relation between two research works

第 2 章　研究现状概述

食品是人类生命生活的必需品，图像已成人类信息交流与服务的主要媒介，而食品图像的研究潜力巨大，所以这方面研究得到越来越多的关注。相关人员在国际上第一次系统提出了食品计算框架[90]，建立了食品计算分类体系，食品计算主要是从不同来源获取和分析异构多模态多属性食品数据，并用于食品的感知、识别、检索、推荐等任务中，可以被广泛用来解决医学，生物学，美食和农学中等多个领域的食品相关问题。此外，相关研究[88]基于食品多模态异构数据特点及食品推荐本身的独特性，第一次系统提出了食品推荐框架，该框架主要包括丰富上下文和知识的融合，用户模型构建及异构食品数据分析三个主要元素，且三者之间互为关联，构成一个有机整体。食品图像识别是开展这些食品图像领域的研究基础，一旦我们识别到食物的类别或者食材，可以进一步进行各种健康相关的分析或者其他任务，例如卡路里摄入量的估计、营养分析、饮食习惯分析和饮食推荐等。而且食品图像识别有着很广泛的实际应用，比如自助餐厅智能计费、炒菜机器人等。因此，食品图像识别是食品图像领域的研究热点。随着深度学习相关技术的发展，食品图像识别算法也迎来了爆炸式增长，本文现将其分为以下四种类型：（1）单标签食品图像识别，其目标是仅包含一种食物的食品图像；（2）针对包含多种食物的多标签食品图像识别；（3）面向移动端的食品图像识别；（4）融入上下文信息的食品图像识别。

2.1　单标签食品图像识别

关于食品图像识别的大多数研究工作都是假设一张食品图像中仅包含一种食物。其研究工作主要包括使用手工特征和深度特征来进行食品图像识别。

使用手工特征有两种方式，一种是单一特征，另一种是不同特征的组合。SIFT（Scale-Invariant Feature Transform）特征[16]被广泛用于食品图像识别的视觉特征[17,18]，文献[19]首先使用语义文本森林将所有图像像素分为几类，然后获得成对特征分布作为视觉特征。相反，大多数方法融合了不同类型的手工特征来提高食品图像识别性能。例如文献[20,21]通过多核学习融合各种类型的图像特征来进行食品图像识别，图像特征包括 SIFT，Gabor 和颜色直方图等特征。文献[22]

利用 LBP（Local Binary Pattern）特征和食品图像主体的全局结构信息。文献[23]则使用了 Garbor，LBP 和 GIST 等多种类型的特征，然后利用其子集来提高性能。

最近几年， CNN 已被广泛用于食品图像识别中的特征提取，并且和传统方法相比，其性能得到很大的提升[24]。现在大部分研究工作直接使用不同类型的 CNN 提取视觉特征进行食品图像识别。比如文献[25]使用 AlexNet 网络[24]提取整张食品图像视觉特征用来食品图像识别。文献[27]通过预训练和微调 AlexNet 网络[24]验证了 CNN 在食品图像识别任务中的有效性。文献[29]直接微调 Inception V3 网络来进行识别。文献[30]使用 ResNet-50 网络[31]直接提取视觉特征进行食品图像识别和营养分析。这些研究工作仅仅使用 CNN 提取视觉特征直接进行识别，并没有考虑食品图像其自身的特点，因此识别性能并不高。有一些研究工作则通过融合不同类型的特征来提高识别性能。比如文献[26]将 CNN 视觉特征与常规特征融合成最终的特征表示。文献[28]基于 GoogLeNet 网络的深度学习特征和语义标签推理的分层语义来进行食品图像识别。文献[32]融合了 AlexNet，GoogLeNet 和 ResNet 网络的视觉特征用于食品图像识别。文献[33]将 ResNet 网络与有监督的机器学习算法相结合，来进行食品图像的识别。文献[34]则融合 WRN（Wide Residual Networks）[35]的视觉特征和他们自己所提切片网络（Slice Network）的视觉特征来进行识别，整个网络由两个分支组成：WRN 网络分支和具有切片卷积层的切片网络分支。WRN 用于捕获一般的视觉特征，而切片网络用于捕获特定的食品垂直结构。通过融合这两个分支的视觉特征，达到了当时多个食品图像数据集上最好性能。文献[89]融合三元卷积神经网络（Triplet Convolutional Neural Network）与关系网络（Relation Network）进行小样本食品图像识别。另外，还有一些相关工作[36]融合了额外的属性特征，比如食品图像的食材和菜谱。这些研究结果表明，融合不同的特征类型有助于食品图像识别。

除了进行食品类别识别以外，还有一些研究工作则对食品图像的风味（Cuisine）或者食材进行识别。比如文献[37]将食材信息作为属性信息来进行食品图像的风味识别。文献[38]将食材视为特征，并构建不同的分类器来预测食谱的风味标签。文献[39]使用多模态深度玻尔兹曼机（Boltzmann Machine）来探索视觉和食材信息，进行多模态的食谱（Recipe）分类。文献[40]提出了一种深度迁移学习方案用于食品图像的食材识别。文献[41]使用了食谱中的类别、配料以及做法，来预测食谱属性（例如口味和风味）。

2.2　多标签食品图像识别

在现实生活中，食品图像中可能有多种食物。文献[42]是从一张食品图像中识别出多种食物的第一项工作，他们首先检测到候选区域，然后对其进行识别。文献[43]进一步使用了食物之间的共现关系信息来进行多种食物的识别。另外，食物检测和食物分割技术也被广泛用于具有多种食物的食品图像中来实现多标签食品识别。

食物检测在早期通常被定义为二分类问题，该算法仅用于区分给定一张图像是否为食品图像[25,44]。这些工作要么使用传统的手工特征表示，要么使用现在普遍流行的深度特征表示。相比于传统的手工特征，通过 CNN 提取的深度特征表现出更优的性能。许多研究人员都是基于 CNN 提取图像的视觉特征，或者直接使用 CNN 来进行端到端的识别。比如文献[45]直接使用了 GoogLeNet 网络来进行食品和非食品的分类实验。文献[46]则通过 GoogLeNet 网络提取视觉特征后，使用 PCA（Principal Components Analysis）降维，最后经过 SVM（Support Vector Machines）分类器来进行食品和非食品的分类。这些研究工作仅限于整张食品图像上的特征提取，然而整张食品图像可能包含一些复杂背景信息或者噪音，因此文献[47]首先通过输入图像形成激活图（Activation Map），然后生成边界框的建议（Proposals），接着使用 GoogLeNet 网络识别出每个边界框内存在的每种食物类型。文献[48]则微调物体检测算法 YOLOv2[49]来进行多种食物识别和检测。这些研究工作基于每个边界框进行特征提取来识别多种食物，由于这些边界框内的食物图像仅包含食物主体，没有复杂背景，因此表现出更好的识别性能。与食物检测相比，食物分割则是对食物图像的每个像素进行分类。文献[50]提出了一种新的方法用于自动分割和识别多种食物图像。它融合图像的颜色特征和纹理特征后，通过 SVM 进行分类。

作为代表性工作，文献[48]提出一种基于语义食物检测方法来实现托盘中多种食物识别。该方法由食物分割、食物检测和语义食物检测三个部分组成。食物分割使用完整的 CNN[51]生成二值图像，然后采用 Moore-Neighbor 跟踪算法进行边界提取，最后微调 YOLOv2[49]来实现食品检测，同时识别出托盘中多种食物。语义食物检测通过融合分割和检测的结果来消除主体检测中的一些错误，从而提高性能。

除了食物类别识别以外，还有一些关于多标签食材识别的工作。比如文献[52,53]使用深度学习的方法来实现食品多标签食材识别。文献[54]进一步提出通过 Inception v3 和 Resnet-50 网络进行多标签食材预测的方法。文献[55]设计了一个多任务学习框架来进行多标签食材识别，该框架能够嵌入食材结构。文献[56]提出了一种多任务系统，该系统可以使用 CNN 从食物图像中识别菜肴类型，食物食材和食物烹饪方法。文献[57]提出了通过二分图（Bipartite-graph）标签体系来探索丰富的食材和标签之间的关系，然后将二分图标签和 CNN 进行融合，实现食品多标签食材识别和菜品（Dish）识别。

2.3 面向移动端的食品图像识别

随着智能便携式设备的迅速普及，将食品图像识别应用于移动环境，实现移动端食品图像识别越来越受到关注。内置传感器与食品图像识别相结合，不仅可以记录日常生活，而且可以为饮食评估和管理提供详细信息。文献[58]开发了一种使用移动设备获取单个食品图像识别的方法。他们首先自动确定图像中特定食物所在的区域，然后根据其特征（包括颜色和质地特征）进行食品图像识别。文献[59]提出了一种基于手机照相的应用程序 DietCam，该应用程序从多个角度考虑食物的外观来进行食物图像识别。文献[60]提出了一种半自动系统，该系统可以识别准备好的饭菜，该系统属于轻量型且可以轻松地嵌入到配备摄像头的移动设备中。文献[61]提出了一个能量消耗估算的实时食品识别平台。文献[62]提出了一种移动食品识别系统 FoodCam，能在智能手机上实现实时的食品图像识别。深度学习提供了一个强大的工具，可以自动生成复杂多媒体数据的高级特征表示。因此，许多深度学习网络，例如 DenseNet[63]，MobileNets[64]和 ShuffleNets[65]已适应移动设备。因此，基于深度学习的移动食品识别方法已经得到快速发展。例如，文献[66]通过引入用于视觉特征提取的深度学习网络，将 FoodCam 扩展到 DeepFoodCam。文献[67]提出了一种移动端食品识别系统，该系统可以识别餐桌上多种食物，例如餐桌上的牛排和土豆，然后进一步估计食品的卡路里和营养。

2.4　融入上下文信息的食品图像识别

除了手机移动端食品图像识别以外，最近的一些调查表明，大部分人选择在外面用餐而不是在家用餐。因此，越来越多的研究工作专注于特定的餐厅场景，实现特定餐厅的食品识别，有些研究工作[88]基于食品识别进一步可以实现餐厅的食品推荐或者个性化的食品推荐。在这种特定场景下，将融入额外的上下文信息，例如位置 GPS 信息和菜单信息。文献[68]提出了一种利用位置传感器信息和各种手工制作的视觉特征来自动识别餐厅食品的方法。文献[69]提出了在地理区域中实现特定餐厅的食品图像识别的框架，并介绍了地理区域模型的概念，其中利用了 DeCAF 的深度特征和餐厅位置信息。他们首先构建了一个餐厅的食品图像数据集，其中包括地理位置和菜单以及相应菜单的食物图像。然后，使用这些菜品图像对地理定位的模型进行训练，其中每个模型都与特定的地理位置相关。当测试时，查询的特定地理位置定义了一些候选餐厅的邻域。对于每个查询，选择相应的地理定位模型，并将其组合成适合该查询的新分类器。文献[70]提出了一个概率图模型来融合菜品、餐厅和位置信息来进行食品识别。文献[71]提出了一种多任务的 CNN，从食品图像中同时识别菜肴和餐厅。文献[72]将 CNN 与循环神经网络（Recurrent Neural Network, RNN）结合起来识别对应餐厅的菜品。此外，文献[73]提出了一个食品识别系统，该系统包含两个主要部分：用于图像特征学习的离线三元组网络和用于图像检索的最近邻搜索网络。当在线（Online）阶段时，使用检索算法，将查询的食品图像及其 GPS 与预定距离内的候选食品图像库进行匹配。这些研究结果表明，在特定场景下融入额外的上下文信息有助于食品图像识别。

2.5　小结

本章从四个不同的维度全面阐述了食品图像识别的国内外研究现状。随着深度学习的发展，许多实验表明 CNN 提取的深度特征比传统的手工特征更具有判别性。因此，现有大多数食品图像识别的方法都是使用 CNN 直接提取整张图像的视觉特征用来识别，但是这些研究工作并没有考虑到食品图像自身的特点，比如食品图像不具有独特的空间布局、食品图像没有共同语义部分等。此外，一些研究工作则引入食品图像特有的属性信息，比如食材、菜谱、餐厅的地理位置信

息等。研究结果表明，引入食品图像的属性信息能从不同视角和粒度来描述食品，并且它和食品图像视觉特征是互补的，因此能提高食品图像的识别性能。随着深度学习的发展，物体检测算法在各个领域得到应用，一些研究工作基于食物检测找到食物主体，然后通过 CNN 提取食物主体视觉特征进行食品图像识别。该方法有效的去除了食品图像中复杂背景信息，基于食品图像局部区域进行识别，从而提高识别性能。虽然基于食物检测能够提高识别性能，但是需要大量边界框的人工标注，而且检测算法需要更多的时间成本和资源。

第 3 章 融合多尺度多视角特征的食品图像识别

大部分研究工作直接使用 CNN 提取整张食品图像的视觉特征来进行识别，但是他们却没有考虑到食品图像自身特点，因此识别性能并不理想。此外，相关研究工作已经验证引入食品图像特有的食材信息可以提高其识别性能。因此在本章中，针对食品图像自身特点，同时融入食品图像的食材信息，提出了一种多尺度多视角特征融合方法来进行食品图像识别。

3.1　问题引出

和一般物体识别一样，食品图像识别的关键是提取具有判别性的视觉特征。但是，食品图像识别作为特殊的物体识别任务，因为以下几点原因，食品图像识别并没有完全被解决。（1）与一般物体识别不同，许多类型的食品没有表现出独特的空间布局和结构。它们通常是非刚性结构，并且不容易利用其结构信息。因此，标准物体识别的方法不适用于食品图像识别任务。现有的食品识别方法，例如文献[34]仅限于具有某些视觉上独特空间排列的食品图像，例如垂直结构（例如，汉堡包）。（2）食品图像识别也可以被看作是细粒度图像识别。细粒度图像识别通常是找到某些主体的固定语义部分，例如鸟类和汽车。但是，固定的语义部分在许多类型的食品图像中并不存在。因此，很难通过现有细粒度识别的方法从食品图像中捕获语义信息。（3）食品图像还具有各种几何变化，例如不同的视角，旋转和比例等。现有食品图像识别的方法一般都是使用 CNN 直接从整张食品图像中提取视觉特征，但是当几何变形较大时识别性能会比较差。因为 CNN 只能通过最大池化来处理具有小范围变形的图像。

食品图像有着其特有的食材信息，就像物体对于场景的重要性一样，食品图像中的食材对于食品识别也是非常重要。因此，中层食材属性学习可以有助于食品图像识别。除了中层食材表示之外，CNN 的高层食品语义分布和深层视觉特征还可以从不同的视角和粒度提供互补性信息。如果将这三种类型的特征融合在一起，则可以最大可能地从食物图像中捕获语义信息。此外，尽管食品图像通常不表现出明显的空间排列，但是我们可以探索不同尺度的 Patch，然后将它们融合成多尺度的特征表示形式。这样的表示可以将 Patch 特征从粗粒度尺度融合到

细粒度尺度，因此它们的特征含有具有判别性图像区域的信息。而且，多尺度融合可以对几何变形更加鲁棒。考虑到这些因素，本文提出了一种用于食品图像识别的多尺度多视角特征融合（MSMVFA）方法，其中多视角意味着不同类型的特征。考虑到食物通常不会表现出明显的空间排列方式，因此本文针对每种类型的特征采用了多尺度融合方法。最粗粒度的尺度是整张食品图像，因此保留了全局空间布局，而更细粒度的比例尺度可以捕获到食物图像的更多局部细粒度细节。因此，这样的融合特征对于食品图像几何变形更加鲁棒。基于每种特征类型的多尺度表示，MSMVFA 可以进一步将高层语义特征，中层属性特征和深层视觉特征融合成统一的特征表示。这三种类型的特征可以从不同的视角来表示食品图像。因此，融合的特征可以最大可能地捕获其语义信息。本文使用其食材信息微调 CNN 来提取中层属性特征，从类别信息监督的 CNN 中提取高层语义特征和深层视觉特征。

3.2 模型设计及实现

本文所提方法使用了两项关键技术。首先，本文作者将高层语义特征，中层属性特征和深层视觉特征融合成统一的特征表示。同类型的特征从不同的粒度描述食品图像。因此，融合的特征可以最大可能地捕获食品图像的语义信息。其次，与一般物体不同，食品图像通常不会表现出独特的空间布局。为了解决这个问题，本文对每种类型的特征经过多尺度融合来获得更鲁棒和有判别性的特征表示。这种多尺度特征表示不仅包含判别性图像区域的特征，而且对几何变形不敏感。

如图 3.1 所示，给定输入图像，MSMVFA 能够提取和融合具有各种比例尺度和不同粒度的三种类型特征。对于该方案，MSMVFA 中引入了两种类型的深度神经网络，即食材网络和类别网络。可以使用任何一种现有的主流 CNN 作为食材网络和类别网络的基础网络，例如 VGG，ResNet 和 DenseNet 网络。通过类别网络，可以提取类别的语义分布以及具有多尺度的更抽象的深层视觉特征。为了获得中层属性特征，本文使用额外的食材信息，还设计了食材网络以多尺度方式提取中层属性特征。与食品类别相比，每种食品类别的食材可以在局部层级上描述食品图像。因此，与全局语义分布相比，食材网络可以提取面向局部区域的中层属性特征。对于每种类型特征，然后通过多尺度融合的方式融合来自不同尺

度的特征。这三种不同类型的融合特征进一步归一化，并通过多视角特征融合方式融合成最终的特征表示。最后，这融合的特征经过 Softmax 分类器进行食品图像识别。在以下各节中，本文将详细介绍 MSMVFA 的主要组成。

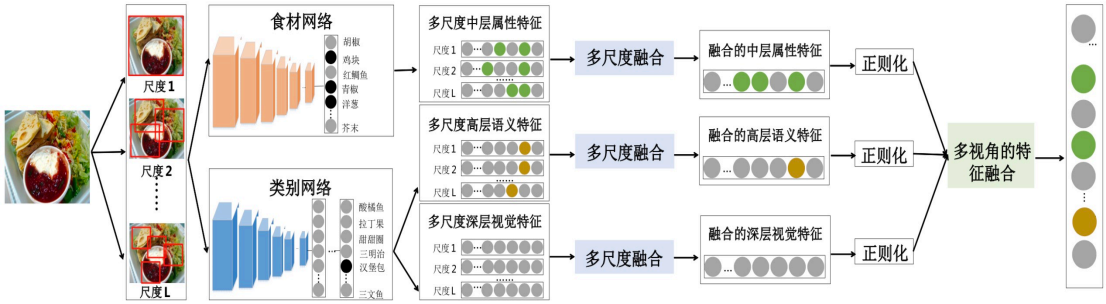


图 3. 1 MSMVFA 模型框架

Figure 3.1 Framework of Multi-Scale Multi-View Feature Aggregation

3. 2. 1 多视角特征融合

3. 2. 1. 1　中层属性特征

食品识别属于细粒度识别，并且由于其视觉复杂性而非常具有挑战性。仅仅探索食品类别信息可能不足以进行食品识别。食品图像有着其特有的食材信息，许多食材可以描述局部食品图像的视觉属性。因此，基于食材信息可以为食品图像提供更为细粒度的特征表示。

为了获得这样的中层属性表示，本文应该设计一个食材网络来提取食材的特征表示。可以采用不同类型的深度属性网络，例如 PANDA[82]和级联 CNN[83]。PANDA 结合了基于人体模型和级联 CNN 的属性预测，首先对面部区域进行定位，然后基于局部区域进行面部属性预测。实际上，在这些属性网络中，食材网络可以直接使用多标签食材信息作为监督来微调 CNN，然后融合所有的属性特征成为统一的特征表示。因此，本文的方法类似于 PANDA，但区别在于 PANDA 首先融合所有比例尺度的特征，然后将融合的特征输入到分类器来进行属性学习。

以第一个尺度为例，即整张图像，通过食材网络，可以得到每张图片x的属性分布：$\hat{a} = (\hat{a}^1, \ldots \hat{a}^m, \ldots \hat{a}^M)$。式中 M 代表的是整个数据集食材的个数。$\hat{a}^m \in [0,1]$代表的是第$m$个食材的预测概率，其概率可以通过 Sigmoid 函数计算得到：

$$\hat{a}^m = \frac{1}{1+e^{(-f(x))}} \tag{3.1}$$

其中，$f()$表示的是食材网络最终的特征。

食品图像食材属性学习其实质是多标签分类，故使用交叉熵损失函数来优化整个食材网络，使用随机梯度下降法来更新参数，直至使优化目标最小：

$$L_I = -\sum_{m=1}^{M}(a^m \log(\hat{a}^m) + (1 - a^m)\log(1 - \hat{a}^m)) \quad (3.2)$$

其中，a^m表示二值化的标签（0 表示当前图像或当前图像区域中不包含该食材，1 表示当前图像或当前图像区域中包含该食材）。通过食材网络，可以得到食材的预测概率$\hat{a}^m$。整个网络的优化目标是使得L_I最小，其作用是让食材预测概率$\hat{a}^m$约等于真实二值化标签a^m。通过食材网络，可以把预测的食材概率分布$\hat{a} = (\hat{a}^1, \dots, \hat{a}^m, \dots \hat{a}^M)$当作中层属性特征。

3.2.1.2 高层语义特征

以 CNN 为基础构建食品类别网络，它的基础网络可以为现有主流的 CNN 结构，比如 VGG，ResNet 和 DenseNet 网络，和食材网络不同的是类别网络使用类别标签去监督整个 CNN。类别网络最后一层输出的是语义概率分布（类别概率分布），它通常表示的是高层的语义信息。通常从头开始训练一个 CNN，需要大量的计算资源和时间开销，故而采用迁移学习的思想，使用在 ImageNet 上预训练的神经网络参数来初始化本文的类别网络。为了获得高层语义特征，微调以类别信息监督的 CNN。在类别网络中，CNN 通常使用 Softmax 函数输出最后一层的概率$\hat{y}^c$：

$$\hat{y}^c = \frac{e^{g(x_c)}}{\sum_{i=1}^{C} e^{g(x_i)}} \quad (3.3)$$

其中，C表示数据集类别的数量，c表示类别索引，$g()$表示的类别网络最终的特征。

最后利用食品类别的预测概率与真实类别标签的交叉熵来优化整个网络，使用随机梯度下降法来更新参数，直至使优化目标最小：

$$L^c = -\sum_x y\log(\hat{y}) \quad (3.4)$$

网络的优化目标是使得L^c最小，其作用是让预测的概率$\hat{y}$约等于真实的标签y。经过微调类别网络，直到网络优化目标L^c最小时，提取预测的类别概率分布$\hat{y} = (\hat{y}^1, \dots \hat{y}^c, \dots \hat{y}^C)$当作高层语义特征，$C$表示的是类别数。

3.2.1.3　深度视觉特征

靠近输出层的类别神经网络层也包含独立的类别相关信息。因此，除了高层语义特征，本文提取靠近输出层的特征当作深层视觉特征，比如 VGG-16 网络中 fc7 层的 4096 维特征。深层视觉特征可以表示为$\hat{h} = (\hat{h}^1, \dots, \hat{h}^d, \dots \hat{h}^D)$，其中$D$表示特征维度。

当获得所有类型的特征后，将其融合为统一的特征表示。考虑到不同特征之间的值会不一样，所以首先对这些特征进行归一化，然后进行特征融合：

$$\mathrm{F} = \mathrm{Agg}(\mathrm{Norm}(\hat{a}), \mathrm{Norm}(\hat{y}), \mathrm{Norm}(\hat{h})) \tag{3.5}$$

Norm()是归一化操作，比如 L2 归一化，Z-score 归一化。Agg()是特征融合方式，比如简单的串连操作或者前馈神经网络。

3.2.2　多尺度特征融合

不同类型的特征在不同比例尺度下效果更好。例如，本文作者可能会在较小的尺度下提取更具判别性的中层食材特征。另外，许多类型的食品图像没有明显的空间布局。每种类型的特征按各种比例尺度进行融合也是解决此问题的一种方法。而且，大部分研究工作提出的多尺度融合方法已被证明在场景识别，图像检索和图像恢复等任务中是用于特征表示的一种有效方法[85, 86, 87]。

对于每种类型的特征，可以采用多尺度 CNN 来提取每个尺度的特征，然后将不同尺度同一特征融合成统一的特征表示。以中层属性特征为例，L=1 代表整张食品图像，而 L=2 代表提取图像的 4 个 Patch，L=N 代表的是最细粒的尺度。对于每一个尺度 L，都训练一个食材网络去提取中层属性特征。例如对于 L=1 的尺度，将整张图像输入到食材网络；对于 L=2 的尺度，将一张图像的 4 个 Patch 输入到食材网络中，分别提取每个 Patch 的属性特征，然后使用最大池化的方式将 4 个 Patch 的特征表示成一张图像的特征表示。最终，可以得到不同尺度的属性特征$\{\hat{a}_L\}_{L=1}^{N}$。

同理，对于高层语义特征和深层视觉特征，依然使用多个尺度：L=1 代表整张食品图像，而 L=2 代表提取图像的 4 个 Patch，L=N 代表的是最细粒的尺度。对于每一个尺度 L，都训练一个类别神经网络去提取高层语义特征和深层视觉特征。对于 L=1 的尺度，本文将整张图像输入到类别网络。对于 L=2 的尺度，将一张图像的 4 个 Patch 输入到类别网络中，分别提取每个 Patch 的属

性特征，然后使用最大池化的方式将 4 个 Patch 的特征表示成一张图像的特征表示。最终，可以得到不同尺度的高层语义特征$\{\hat{y}_L\}_{L=1}^{N}$和深层视觉特征$\{\hat{h}_L\}_{L=1}^{N}$。

3.2.3 多尺度多视角特征融合

当所有特征提取完，就可以分别对上述三种类型特征的各个尺度特征进行多尺度融合，然后得到三种多尺度融合后的特征。三种类型特征的多尺度融合可以表示成$\mathrm{Fus}(\{\hat{a}_L\}_{L=1}^{N})$、$\mathrm{Fus}(\{\hat{y}_L\}_{L=1}^{N})$、$\mathrm{Fus}(\{\hat{h}_L\}_{L=1}^{N})$。融合操作 Fus()可以是简单的串连或者前馈神经网络操作，在实际操作中可以采用简单的串连方式进行不同尺度间的融合。经过多尺度融合后，可以获得多尺度融合后的特征。在本文的方法中，有三种不同的特征，所以采用多视角融合方式，就可以得到一张图像的特征表示。多视角融合可以表示成：

$$\mathrm{F}=\mathrm{Agg}(\mathrm{Norm}(\mathrm{Fus}(\widehat{\{\hat{a}_L\}_{L=1}^{N}})),\mathrm{Norm}(\mathrm{Fus}(\{\hat{y}_L\}_{L=1}^{N})),\mathrm{Norm}(\mathrm{Fus}(\{\hat{h}_L\}_{L=1}^{N})) \quad (3.6)$$

Norm()是归一化操作，比如 L2 归一化，Z-score 归一化。Agg()是特征融合操作，比如简单的串连或者前馈神经网络。

对于第一级融合，本文进行了多尺度特征融合，融合的特征包含来自食品图像判别性区域的特征，并且可以适应几何变形。对于第二级融合，本文融合三种不同类型的特征，以最大可能地捕获食品图像的语义特征。因此，本文提出的两级融合即 MSMVFA 适用于食品图像。

在测试阶段，首先基于训练数据集中每个图像的融合特征使用 Softmax 分类器进行训练。给定一张测试图像，首先获得预测的中层属性特征，预测的高层语义特征和深层的视觉特征，然后通过 MSMVFA 将这些不同类型的特征融合为最终的特征表示。最后，将融合的特征输入到分类器中获得预测结果。为了在后续实验中进行公平比较，因此本文采用 Softmax 分类器对最终的融合特征表示进行分类。

3.2.4 分析

MSMVFA 的优势可以从两个方面获得。首先，MSMVFA 可以在不同的监督信号下获得不同类型的特征。类别监督的 CNN 可以提供高层语义特征和深层视觉特征，而食材监督的 CNN 可以提供细粒度的中层属性特征。从不同的角度和粒度来看，它们是互补的。其次，MSMVFA 可以探索具有不同比例尺度的判别

性图像区域。将这些区域特征（从粗尺度到细尺度）融合在一起，就可以最大可能地包含判别性信息。另外，这样的融合特征也可以对几何变形更加鲁棒。因此，MSMVFA 的最终融合特征是具有互补性和判别性的。

3.3　实验验证与分析

3.3.1　实验数据

ETH Food-101[13]是一个包含101种食品和101,000张食品图像的数据集。每个类别有1,000张图像，其中包括750张训练图像和250张测试图像。相关研究[54]进一步提供了相应的227种食材。

VireoFood-172[78]包含172种食品、110,241张食品图像和353种食材。为了实验的公平性，其数据集的划分和文献[78]一样，在每种食品类别中，分别随机选择60%、10%、30%的图像进行训练、验证和测试。

ChineseFoodNet[84]包含 185,628 张食品图像和 208 种中国食品。 整个数据集划分为 145,065、20,253 和 20,310 张图像，分别用于训练、验证和测试。 但是，该数据集并未提供测试集的标签信息。因此，本文将验证集分为两部分：约 20%（4,050）用作验证集，其余 80%（16,503）用作测试集。该数据集不提供相关的食材信息。考虑到 ChineseFoodNet 和 VireoFood-172 都属于中国菜，本文使用来自 VireoFood-172 的食材列表作为该数据集的食材。

图 3.2 展示了这三个数据集一些食品图像样例，我们可以看到，VireoFood-172 和 ChineseFoodNet 的食品类别有重叠，比如 Shredded cabbage。这是因为这两个数据集都属于中国菜。

图 3.2　三个数据集食品图像样例

Figure 3.2 Some food examples from three datasets

3.3.2 评测指标

本文将使用Top-1和Top-5分类准确率作为评价指标。Top-1分类准确率表示测试图像中最有把握的预测标签正确的百分比。Top-5分类准确率表示真实标签在排名前5位的预测标签中的测试图像所占的百分比。因为多项研究工作都使用Top-1和Top-5分类准确率，为了公平对比，本文也将Top-1和Top-5分类准确率作为评价指标。

3.3.3 实现细节

VGG、ResNet 和 DenseNet 网络是当前三种比较主流的 CNN 结构。为了验证 MSMVFA 的有效性和鲁棒性，本文基于这三种基础 CNN 进行相关实验。在实验中，本文选择 VGG-16、ResNet-152 和 DenseNet-161 网络当做 MSMVFA 基础网络。VGG-16 网络的初始学习率设置为 0.0001，而 ResNet-152 网络和 DenseNet-161 网络的初始学习率设置为 0.001。在类别网络和食材网络上，学习率都在 10 个周期后除以 10。VGG-16、ResNet-152 和 DenseNet-161 网络的批量（Batch Size）分别设置为 48、8 和 8。每个网络都训练 30 个周期。对于 VireoFood-172 和 ChineseFoodNet 数据集，本文选择验证集上 Top-1 分类准确率最高的模型用来测试。而 ETH Food-101 数据集没有提供验证集，本文选择训练损失函数不再变化的模型用来测试。所有 CNN 都使用 0.9 的动量（Momentum）和 0.0001 的权重衰减，并且使用随机梯度下降法进行整个网络优化。本文在 Nvidia GPU Titan X 上使用 Caffe 训练所有 CNN。每种模型都在 ImageNet 上进行了预训练。

在本文实验中，采用三种不同尺度 L1、L2、L3，L1 对应于全局 256×256 图像，L2 对应于 128×128 的 Patch，L3 对应于 64×64 的 Patch。对于 L1 尺度，本文直接使用整张图像对模型进行微调。对于 L2 尺度，一张图像分为 4 个 Patch，这些 Patch 共享相同的食品类别和食材标签。本文将所有 Patch 二线性插值到 256×256 大小来微调模型。L3 尺度也采用相同的策略。对于多尺度融合 Fus()和多视角融合 Agg()，本文均采用简单的串连操作。而且采用最大池化进行不同 Patch 的特征融合。尽管可以选择其他尺度类型和特征融合方法，但在本项研究工作中只强调所提方法的有效性。

对于深层视觉特征，本文从 VGG-16 网络的 FC7 层中提取 4096 维特征特征，从 ResNet-152 网络中提取 2048 维特征，从 DenseNet-161 网络中提取 2208 维特

征。本文从类别网络中提取类别概率预测层的高层语义特征。同样，本文从食材网络中提取食材预测层的中层属性特征。

3.3.4　性能分析

3.3.4.1　ETH Food-101 的性能分析

在本小节中，本文首先对多尺度特征融合进行性能比较，然后对多视角特征融合进行性能比较。最后，本文将与最新技术进行性能比较。

多尺度特征融合的性能比较。表 3.1 至表 3.9 分别展示 VGG-16、ResNet-152、DenseNet-161 网络结构在 ETH Food-101 数据集上对深层视觉特征、中层属性特征和高层语义特征不同尺度融合的结果。我们可以看到（1）对于单一尺度，在所有三种类型的网络中，L2 尺度 128×128 的 Patch 在三种类型特征上的性能都优于 L1 和 L3 尺度。原因是不同类型的特征在不同比例尺度下效果更好。当采用深层视觉特征时，与 L1 和 L3 相比，L2 可能包含更多的判别性信息。大尺度可能包含更多背景，而 L3 尺度则包含不完整的外观和深层视觉特征。（2）在表 3.1 到表 3.3 中，对于所有这三种类型的特征，大多数情况下两种尺度之间的融合性能要高于单个尺度。在三种类型的特征中，所有三个尺度融合一起在 Top-1 分类准确率上达到最佳性能。当表 3.4 到表 3.9 中采用不同的 CNN 时，可以看到类似的趋势。性能之所以提高得益于三种不同尺度之间的互补性。因此，我们可以得到结论，多尺度融合可以提高识别性能。（3）在这三种类型的 CNN 中，DenseNet 上的多尺度融合性能最高，是因为 DenseNet 提出了一个更激进的密集连接机制：即互相连接所有的层，具体来说就是每个层都会接受其前面所有层作为其额外的输入。

表 3. 1　VGG-16 深层视觉特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.1 The performance comparison of different combinations of scales for deep visual features on the ETH Food-101 using VGG-16 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	4,096	78.76	94.19
L2	4,096	84.73	96.47
L3	4,096	83.03	96.03
L1+L2	8,192	83.26	96.02
L1+L3	8,192	81.34	95.31
L2+L3	8,192	83.95	96.27
L1+L2+L3	12,288	**85.89**	**96.98**

表 3. 2　VGG-16 中层属性特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.2 The performance comparison of different combinations of scales for mid-level attribute features on the ETH Food-101 using VGG-16 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	227	77.67	88.65
L2	227	76.32	91.32
L3	227	74.97	**92.27**
L1+L2	454	78.50	90.06
L1+L3	454	76.61	88.42
L2+L3	454	75.80	90.09
L1+L2+L3	681	**78.60**	90.36

表 3. 3　VGG-16 高层语义特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.3 The performance comparison of different combinations of scales for high-level semantic features on the ETH Food-101 using VGG-16 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	101	78.38	94.07
L2	101	81.08	95.38
L3	101	76.66	94.18
L1+L2	202	84.37	96.38
L1+L3	202	82.54	95.75
L2+L3	202	81.83	95.56
L1+L2+L3	303	**84.94**	**96.68**

表 3. 4　ResNet-152 深层视觉特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.4 The performance comparison of different combinations of scales for deep visual features on the ETH Food-101 using ResNet-152 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	2,048	83.61	95.72
L2	2,048	87.02	97.13
L3	2,048	82.71	95.32
L1+L2	4,096	88.39	97.70
L1+L3	4,096	86.07	96.85
L2+L3	4,096	88.15	97.39
L1+L2+L3	6,144	**89.00**	**97.86**

表 3. 5　ResNet-152 中层属性特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.5 The performance comparison of different combinations of scales for mid-level attribute features on the ETH Food-101 using ResNet-152 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	227	80.42	90.23
L2	227	83.37	94.76
L3	227	76.88	92.99
L1+L2	454	82.71	92.40
L1+L3	454	82.00	92.05
L2+L3	454	82.82	94.36
L1+L2+L3	681	**83.81**	**95.70**

表 3. 6　ResNet-152 高层语义特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.6 The performance comparison of different combinations of scales for high-level semantic features on the ETH Food-101 using ResNet-152 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	101	83.41	95.67
L2	101	82.86	95.74
L3	101	76.68	93.17
L1+L2	202	88.08	97.53
L1+L3	202	87.36	97.22
L2+L3	202	85.80	96.56
L1+L2+L3	303	**89.05**	**97.79**

表 3. 7　DenseNet-161 深层视觉特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.7 The performance comparison of different combinations of scales for deep visual featureson on the ETH Food-101 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	2,208	86.94	97.03
L2	2,208	89.08	97.91
L3	2,208	85.93	96.95
L1+L2	4,416	89.57	97.94
L1+L3	4,416	88.64	97.69
L2+L3	4,416	90.04	98.11
L1+L2+L3	6,624	**90.14**	**98.11**

表 3. 8　DenseNet-161 中层属性特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.8 The performance comparison of different combinations of scales for mid-level attribute features on the ETH Food-101 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	227	82.84	93.30
L2	227	84.50	**95.30**
L3	227	78.10	93.98
L1+L2	454	84.88	94.82
L1+L3	454	83.57	94.06
L2+L3	454	83.74	95.04
L1+L2+L3	681	**84.89**	94.83

表 3. 9　DenseNet-161 高层语义特征多尺度融合在 ETH Food-101 数据集的性能（%）

Table 3.9 The performance comparison of different combinations of scales for high-level semantic features on the ETH Food-101 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	101	86.66	96.89
L2	101	86.28	97.36
L3	101	82.16	96.15
L1+L2	202	88.76	97.79
L1+L3	202	88.28	97.50
L2+L3	202	86.10	97.31
L1+L2+L3	303	**89.32**	**97.94**

多视角特征融合的性能比较。在本文实验中，使用了三种类型的特征，即深层视觉特征，中层属性特征和高层语义特征。表 3.10 至表 3.12 展示了在不同网络结构下不同类型特征的多视角融合结果，其中 F1 表示具有多尺度融合的深层视觉特征。F2 和 F3 分别表示具有多尺度融合的中层属性特征和高层语义特征。考虑到不同类型的特征在不同取值范围内，本文首先对每种类型特征进行归一化，然后将它们串连起来。在本文实验中，将每种类型的特征都归一化到[0，1]区间，然后使用 z-score 方法进行归一化。从表 3.10 至表 3.12 中，可以看出（1）对于这三个网络，两种或三种类型特征之间融合的性能通常都比单一类型特征的性能要高。所有三种类型特征串联一起在 Top-1 和 Top-5 分类准确率上达到最佳性能。因此，可以得出结论，不同类型的特征可以从不同视角来描述食品图像，因此这些类型的特征是非常互补的。（2）对于所有三种类型的 CNN，通过两级融合获得的特征性能均优于通过多尺度融合获得的特征性能。这证明本文所提的 MSMVFA 通过两级融合可以获得更好性能。(3)对于三种类型的 CNN，DenseNet-161 网络上的多尺度多视角特征融合性能最好。因此，在其他数据集的实验中，本文将 DenseNet-161 网络用作另外两个数据集 VireoFood-172 和 ChineseFoodNet 的基础网络。

表 3. 10　VGG-16 多视角特征融合在 ETH Food-101 数据集的性能（%）

Table 3.10 Comparison of Top-1 and Top-5 accuracy from multi-view feature aggregation under VGG-16 network architecture on ETH Food-101 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
F1	12,288	85.89	96.97
F2	681	78.60	90.36
F3	303	84.94	96.68
F1+F2	12,969	87.66	97.43
F1+F3	12,591	87.41	97.33
F2+F3	984	84.30	95.88
F1+F2+F3	13,272	**87.68**	**97.45**

表 3. 11　ResNet-152 多视角特征融合在 ETH Food-101 数据集的性能（%）

Table 3.11 Comparison of Top-1 and Top-5 accuracy from multi-view feature aggregation under ResNet-152 network architecture on ETH Food-101 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
F1	6,144	89.00	97.85
F2	681	83.81	95.70
F3	303	89.05	97.79
F1+F2	6,825	90.36	98.12
F1+F3	6,447	89.80	98.00
F2+F3	984	89.15	97.99
F1+F2+F3	7,128	**90.37**	**98.15**

表 3. 12　DenseNet-161 多视角特征融合在 ETH Food-101 数据集的性能（%）

Table 3.12 Comparison of Top-1 and Top-5 accuracy from multi-view feature aggregation under DenseNet-161 network architecture on ETH Food-101 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
F1	6,624	90.14	98.11
F2	681	84.89	94.83
F3	303	89.32	97.94
F1+F2	7,305	90.51	98.17
F1+F3	6,927	90.57	98.19
F2+F3	984	89.54	98.02
F1+F2+F3	7,608	**90.59**	**98.25**

与最新技术性能比较。本文与现有最新技术进行性能比较。表 3.13 展示了现有方法和 MSMVFA 在 ETH Food-101 数据集上性能比较。表 3.13 展示了 AlexNet，Inception V3，ResNet-200 和 WRN 等不同的 CNN 性能。从表 3.13 中，我们可以看到（1）WRN 的性能优于其他单个 CNN。（2）WISeR [34]通过增加带有切片卷积层的切片网络分支来改善 WRN 性能，该网络可以捕获食品图像特定的垂直结构。通过融合两个分支的输出进行食品识别。（3）使用 ResNet 网络时，MSMVFA（ResNet）在 Top-1 分类准确率上优于 WISeR 的性能。当采用 DenseNet 网络时，MSMVFA 在 Top-1 分类准确率上达到最好性能，并且比专门为食品识别设计的 WISeR 网络在 Top-1 分类准确率上性能提高 0.3%。尽管性能提升较

弱，但 MSMVFA 仍达到 Top-1 分类准确率的最好食品识别性能。实验结果验证了本文所提 MSMVFA 方法的有效性。

表 3. 13　MSMVFA 在 ETH Food-101 数据集的性能（%）

Table 3.13 Comparison of our model and state-of-the-art methods on ETH Food-101 (%)

方法	Top-1 分类准确率	Top-5 分类准确率
AlexNet-CNN[13]	56.40	-
DCNN-FOOD[27]	70.41	-
DeepFood[79]	77.40	93.70
FCAN[80]	86.50	-
CurriculumNet[81]	87.30	-
Inception V3[29]	88.28	96.88
ResNet-200[31]	88.38	97.85
DenseNet-161[63]	86.94	97.03
WRN[34]	88.72	97.92
WISeR[34]	90.27	**98.71**
MSMVFA(ResNet-152)	**90.37**	98.15
MSMVFA(DenseNet-161)	**90.59**	98.25

3.3.4.2　VireoFood-172 和 ChineseFoodNet 的性能分析

表 3.14 至表 3.16 展示了 DenseNet-161 网络在 VireoFood-172 数据集上多尺度特征融合的识别准确率。和 ETH Food-101 不同，VireoFood-172 数据集属于中国菜。如表 3.14 至表 3.16 所示，对于每种类型的特征，与单尺度或两个尺度融合相比，三种不同尺度的融合特征在 Top-1 和 Top-5 分类准确率上均达到了最好性能。表 3.19 至表 3.21 展示了 ChineseFoodNet 的多尺度特征融合的识别准确率。ChineseFoodNet 的类别和样本数量多于 VireoFood-172 数据集。由于 ChineseFoodNet 不提供食材信息，因此本文直接从 VireoFood-172 数据集的食材网络中提取食材特征。我们可以再次看到，将所有三个尺度的特征串联一起在 Top-1 和 Top-5 分类准确率上均达到最佳性能。

表 3.17 和表 3.22 分别展示了在 VireoFood-172 和 ChineseFoodNet 两个数据集上多视角融合的实验结果。在 VireoFood-172 数据集上，融合三种类型特征实现了 Top-1 分类准确率的最佳性能。

表 3. 14　DenseNet-161 深层视觉特征多尺度融合在 VireoFood-172 数据集的性能（%）

Table 3.14 The performance comparison of different combinations of scales for deep visual featureson on the VireoFood-172 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	2,208	87.40	97.25
L2	2,208	89.70	98.02
L3	2,208	83.86	96.02
L1+L2	4,416	89.96	98.10
L1+L3	4,416	88.63	97.67
L2+L3	4,416	90.23	98.14
L1+L2+L3	6,624	**90.28**	**98.20**

表 3. 15　DenseNet-161 中层属性特征多尺度融合在 VireoFood-172 数据集的性能（%）

Table 3.15 The performance comparison of different combinations of scales for mid-level attribute features on the VireoFood-172 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	353	82.84	95.55
L2	353	83.15	96.35
L3	353	77.86	94.44
L1+L2	706	84.96	96.75
L1+L3	706	84.51	96.46
L2+L3	706	83.85	96.62
L1+L2+L3	1,059	**85.87**	**97.13**

表 3. 16　DenseNet-161 高层语义特征多尺度融合在 VireoFood-172 数据集的性能（%）

Table 3.16 The performance comparison of different combinations of scales for high-level semantic features on the VireoFood-172 using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	172	86.93	97.17
L2	172	87.53	97.53
L3	172	77.86	94.28
L1+L2	344	89.34	97.98
L1+L3	344	88.29	97.54
L2+L3	344	88.29	97.73
L1+L2+L3	516	**89.75**	**98.08**

表 3. 17　DenseNet-161 多视角特征融合在 VireoFood-172 数据集的性能（%）

Table 3.17 Comparison of Top-1 and Top-5 accuracy from multi-view feature aggregation under DenseNet-161 network architecture on VireoFood-172 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
F1	6,624	90.28	98.20
F2	1,059	85.87	97.13
F3	516	89.75	98.08
F1+F2	7,683	90.56	98.22
F1+F3	7,140	90.55	98.22
F2+F3	1,575	90.06	**98.40**
F1+F2+F3	8,199	**90.61**	98.31

表 3. 18　MSMVFA 在 VireoFood-172 数据集的性能（%）

Table 3.18 Comparison of our model and state-of-the-art methods on VireoFood-172 (%)

方法	Top-1 分类准确率	Top-5 分类准确率
AlexNet	64.91	85.32
VGG-16	80.41	94.59
DenseNet-161	86.93	97.17
MultiTaskDCNN(VGG-16)[78]	82.06	95.88
MultiTaskDCNN(DenseNet-161)[78]	87.21	97.29
MSMVFA(DenseNet-161)	**90.61**	**98.31**

在 ChineseFoodNet 中，融合三种类型特征的性能与当前最好性能相差无几。因为本文只采用了 VireoFood-172 食材网络中的特征，而这样的食材特征对于 ChineseFoodNet 并不是最佳的。最后，本文在 VireoFood-172 和 ChineseFoodNet 上将 MSMVFA 与其他方法进行比较。表 3.18 和表 3.23 分别展示了在 VireoFood-172 和 ChineseFoodNet 上的识别结果。MultiTaskDCNN [78]是一种多任务学习方法，具有两种类型的输出层，一种是类别预测层，另一种是食材预测层。为了公平比较，本文同样实现了基于 DenseNet-161 网络的 MultiTaskDCNN 版本。如表 3.18 所示，尽管 VireoFood-172 与 ETH Food-101 数据集有很大不同，但观察到相似的结果，验证了本文方法的有效性，同样 MSMVFA 也达到了最佳性能。在表 3.23 中，ChineseFoodNet 可以观察到类似的实验结果。

表 3. 19　DenseNet-161 深层视觉特征多尺度融合在 ChineseFoodNet 数据集的性能（%）

Table 3.19 The performance comparison of different combinations of scales for deep visual features on the ChineseFoodNet using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	2,208	75.49	94.33
L2	2,208	81.11	96.60
L3	2,208	79.02	95.72
L1+L2	4,416	81.04	96.56
L1+L3	4,416	79.35	96.04
L2+L3	4,416	81.94	96.82
L1+L2+L3	6,624	**81.96**	**96.92**

表 3. 20　DenseNet-161 中层属性特征多尺度融合在 ChineseFoodNet 数据集的性能（%）

Table 3.20 The performance comparison of different combinations of scales for mid-level attribute features on the ChineseFoodNet using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	353	63.56	88.44
L2	353	63.06	88.41
L3	353	59.00	85.72
L1+L2	706	66.03	**90.36**
L1+L3	706	66.01	90.23
L2+L3	706	64.09	89.10
L1+L2+L3	1,059	**66.41**	90.32

表 3. 21　DenseNet-161 高层语义特征多尺度融合在 ChineseFoodNet 数据集的性能（%）

Table 3.21 The performance comparison of different combinations of scales for high-level semantic features on the ChineseFoodNet using DenseNet-161 (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
L1	172	75.22	93.97
L2	172	78.67	95.75
L3	172	75.17	96.07
L1+L2	344	79.07	95.87
L1+L3	344	77.26	95.31
L2+L3	344	78.28	95.78
L1+L2+L3	516	**79.47**	**96.25**

表 3. 22　DenseNet-161 多视角特征融合在 ChineseFoodNet 数据集的性能（%）

Table 3.22 Comparison of Top-1 and Top-5 accuracy from multi-view feature aggregation under DenseNet-161 network architecture on ChineseFoodNet (%)

方法	维度	Top-1 分类准确率	Top-5 分类准确率
F1	6,624	81.96	96.92
F2	1,059	66.41	90.32
F3	624	79.47	96.26
F1+F2	7,683	81.91	96.82
F1+F3	7,248	**81.99**	96.89
F2+F3	1,683	81.17	96.86
F1+F2+F3	8,307	81.94	**96.94**

表 3. 23　MSMVFA 在 ChineseFoodNet 数据集的性能（%）

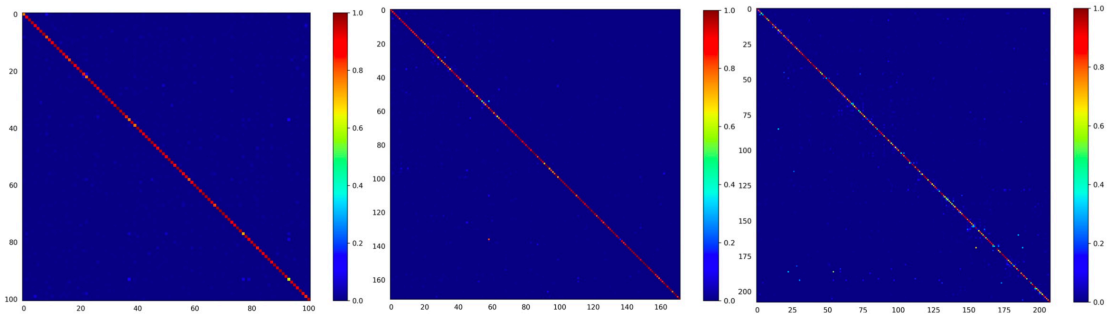
Table 3.23 Comparison of our model and state-of-the-art methods on ChineseFoodNet (%)

方法	Top-1 分类准确率	Top-5 分类准确率
DenseNet-121	78.07	95.42
DenseNet-161	78.87	95.80
DenseNet-201	79.05	95.79
DenseNet Fusion	80.47	96.26
MSMVFA(DenseNet-161)	**81.94**	**96.94**

3.3.5　讨论

本章提出的方法在三个主流大规模食品数据集上均实现了食品识别的最好性能。但是，仍然有一些食品图像难以被识别出来。本小节列出了从方法中得到的结果，尝试以此来找到难以识别的原因。

图 3.3 展示了 MSMVFA 在三个数据集上各个类别的混淆矩阵。我们可以看到，对于某些食品类别，MSMVFA 仍无法完全识别。基于图 3.3 观察到一些混淆食品类别，图 3.4 进一步展示了三个数据集中一些混淆的食品类别。我们可以看到这些食品类别在外观上非常相似，甚至人类也不容易在这些食品类别之间进行区分。我们可能需要设计更为细粒度的视觉特征来进行食品图像的识别。



(a) ETH Food-101　(b) VireoFood-172　(c) ChineseFoodNet

图 3. 3 MSMVFA 在三个数据集上的混淆矩阵

Figure 3.3 The detailed comparison over each individual food category for MSMVFA via the confusion matrix

(a) ETH Food-101　(b) VireoFood-172　(c) ChineseFoodNet

图 3. 4 三个数据集上混淆的图像样例

Figure 3.4 Some confused food categories from three datasets

3. 4　小结

本章提出了一种 MSMVFA 方法用来食品图像识别。MSMVFA 将食品图像和食材上下文信息相结合在一起，将高层语义特征，中层属性特征和深层视觉特征融合成统一的特征表示，从而最大可能地捕获食品图像的语义信息。此外，通过对每种类型的特征进行多尺度融合，MSMVFA 能够学习到更加强大和判别性的特征来处理食品图像的几何变形。通过两级融合，即针对每种类型的特征进行多尺度特征融合，并在三种类型的特征之间进行多视角融合，MSMVFA 可以生成更强大、更具区分性和综合性的特征表示来应对具有独特复杂视觉的食品图像。大量的实验结果表明，MSMVFA 在 Top-1 分类准确率上优于三个主流食品数据集的所有基线（Baseline）模型。

第 4 章　基于级联多注意力网络的食品图像识别

在第 3 章中，提出了一种 MSMVFA 方法来识别食品图像，但是在这个方案中预先使用了固定的 Patch，这个 Patch 可能会包含一定的噪音。因此，固定的 Patch 来表示判别性的区域不是最佳的选择。在本章节中，本文将使用弱监督的方式来定位出可判别性的区域而不是使用固定 Patch 来提高识别性能。

4.1　问题引出

食品图像识别属于细粒度识别，它是区分下属类别（例如鸟类和汽车）的任务。解决此问题的关键是提取具有判别性区域的特征。现有细粒度方法大多数是通过类别信息监督下，以弱监督的方式定位到多个语义区域。但是，图像的类别标签仅能提供弱监督的信息。因此，使用类别标签训练的 CNN 可能定位不到具有互补性的细粒度食物区域，而且可能不是监督多个区域定位的最佳方法。此外，现有的细粒度图像基本上都具有固定的语义部分，并且语义部分与整体之间存在明显的关系。在这种假设下，这些方法主要定位于固定的语义区域，同时利用这种关系约束来去除不合理的区域。但是，与这些细粒度图像相比，许多类型的食品都是非刚性的，并且没有表现出独特的空间结构和固定的语义模式。因此，很难通过现有的细粒度方法从食物图像中捕获判别性语义信息。

食品图像所特有的食材信息提供了一个新的思路，通过食材信息来明确地指导网络在食品图像不同尺度上发现不同的语义区域，这些网络在不同粒度的监督下生成的区域特征是非常互补的。因此，融合这些区域的特征可以形成更全面，更具区分性的特征表示。所以，本文提出了一种食材指导的级联多注意力网络（IG-CMAN）来实现食品图像识别，该方法能够以粗粒度到细粒度、基于类别信息和食材信息的多任务方式从食品图像中找到多个食品图像的区域。在类别信息监督的注意力子网络（Category-supervised Attention Sub-Network, CASN)中，该方法通过空间转换网络（Spatial Transformer Network, STN）生成去除复杂背景信息的初始区域。然后以该局部注意力区域为基础，将 STN 和长短期记忆网络（Long Short Term Memory, LSTM）结合起来，从以下几个层级的食材监督的注意力子网络（Ingredient-supervised Attention SubNetwork, IASN）中依次发现具

有细粒度的多样化注意力区域。为了验证方法的有效性，本文还额外构建了 ISIA Food-200 数据集，其中包含 Wikipedia 列表中的 200 种食品、大约 200,000 张食品图像和 319 种食材。接下来会详细介绍模型的设计和实现以及实验结果。

4.2 ISIA Food-200 数据集构建

构建 ISIA Food-200 数据集，本文进行了以下四步：

（1）**构建食品名字列表。**首先根据 Wikipedia 中“按食材分类的食品清单”来构建初始食品列表，然后使用深度优先算法搜索食品网站链接来构建更准确的食品名字列表,最后通过人工删除重复的食品类别和进行同义词重组以获得最终的食品名字列表。

（2）收集食品图像。根据食品名字列表，以列表中的食品名字为查询搜索各种搜索引擎（百度、谷歌等）中的候选食品图像。本文添加“食品”和“菜肴”等关键字来扩展搜索词以确保搜索到的图像是食品图像。然后将关键字搜索到的食品图像进行重组。因为从不同搜索引擎获得的食品图像进行组合，可能会包含重复的图像。因此本文进行了图片去重。

（3）**清理食品图像。**本文通过两步来清理数据集：1）自动清理。使用程序自动去除尺寸小、程序无法读取和像素很低的食品图像，然后基于现有食品图像数据集和 ImageNet 上的图像构建二分类模型来自动地删除不是食品的图像。2）手工清理。基于自动清洗完后的图片，本文进行了人工清理并标注，删除那些不合理不符合预期的食品图像。

（4）**整理数据集**。通过人工清洗和人工标注以后，有些食品种类的图片数量很少，因此本文将这些食品种类名字翻译成不同语言，在不同搜索引擎上再次爬取图片。然后通过第三步清洗，最终本文获得了 200 种食品、大约 200,000 张食品图像。

4.3 模型设计和实现

如图 4.1 所示,整个模型框架由 2 个主要部分组成,分别是 CASN 和 IASN。CASN 可以去掉食品图像的复杂背景信息，从而在全图中找到食品主体。基于这个食品主体，IASN 可以找到食材对应的区域。对于每一个子网络都包含 STN 和

LSTM 及全连接层。每层循环的 STN 用于找到注意力区域，而来自不同循环的 LSTM 堆叠在一起可以建模这些局部区域的顺序依赖性，同时为下一循环的 STN 生成转换参数。最后，整个框架采用多任务学习的模式，使用一个类别损失函数和食材损失函数以端到端的方式优化整个网络。

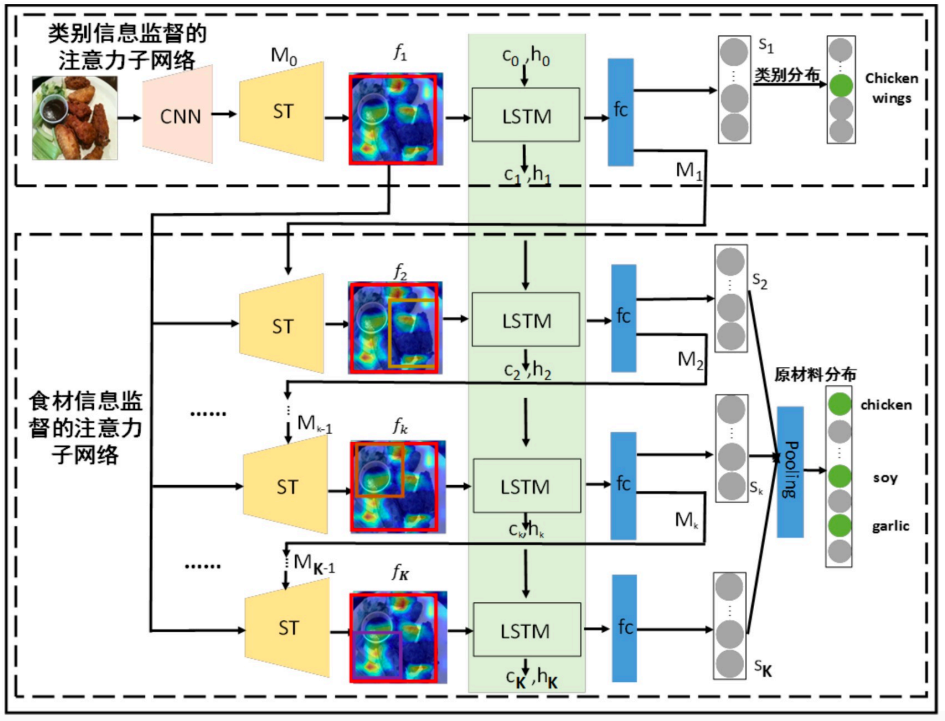


图 4. 1 IG-CMAN 模型框架

Figure 4.1 Overview of proposed Ingredient-Guided Cascaded Multi-Attention Network

4. 3. 1 类别信息监督的注意力子网络

如图 4. 1 所示，CASN 是由传统的 CNN、STN 和 LSTM 组成。STN 可以将其输入映射到空间，转换为给定大小的输出映射，这个输出映射对应于输入映射的一个子区域。食品图像输入到传统的 CNN 中，本文提取最后一层卷积层的特征图f_0，经过 STN 可以得到原图的一个区域的特征图f_1，当我们的目标是定位注意力区域时，本文将变换矩阵M限制为仅涉及裁剪，平移和缩放。所以变换矩阵M的定义为:

$$M = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix}$$

其中s_x, s_y表示平移，t_x, t_y表示缩放。

LSTM 和下面循环的 LSTM 构建成堆叠的 LSTM，用来建模局部区域的顺

序依赖性。经过 STN，可以获得 LSTM 的输入x_1。

$$f_1 = ST(f_0, M_0), \qquad x_1 = relu(W_{fx} f_1 + b_x) \tag{4.1}$$

其中$ST()$是空间转换函数，$relu()$表示修正线性函数（Rectified Linear Function），W_{fx}和b_x表示的是函数参数。

然后通过 LSTM，可以获得隐含表示h_1和状态c_1。基于这个输出h_1，LSTM 不仅可以预测类别得分，而且输出下一循环的转换矩阵M_1：

$$\begin{aligned} z_1 &= relu(W_{hz} h_1 + b_z) \\ s_1 &= W_{zs} z_1 + b_s \\ M_1 &= W_{zm} z_1 + b_m \end{aligned} \tag{4.2}$$

其中W_{hz}，W_{zs}，W_{zm}，b_s，b_z，b_m是转换参数。

4.3.2 食材信息监督的注意力子网络

基于 CASN 定位到的区域f_1和转换矩阵M_1，在 IASN 中，堆叠的 LSTM 和 STN 以循环迭代的方式协同工作：LSTM 预测来自 STN 定位的局部区域的食材分数，并同时更新 STN 的转换参数以用于下一个注意力区域定位。对于 IASN 中的每个子网络，都将去掉复杂背景信息的局部区域f_1作为基础，并使用更新的转换矩阵M_{k-1}来计算更加细粒度的局部区域：$f_k = ST(f_1, M_{k-1})$。LSTM 将特征映射f_k作为输入，来计算单元和隐藏状态：

$$\begin{aligned} x_k &= relu(W_{fx} f_k + b_x) \\ f_k &= \sigma(W_{xf} x_k + W_{hf} h_{k-1} + b_f) \\ i_k &= \sigma(W_{xi} x_k + W_{hi} h_{k-1} + b_i) \\ g_k &= tanh(W_{xg} x_k + W_{hg} h_{k-1} + b_g) \\ c_k &= f_k \odot c_{k-1} + i_k \odot g_k) \\ o_k &= \sigma(W_{xo} x_k + W_{ho} h_{k-1} + b_o) \\ h_k &= o_k \odot c_k \end{aligned} \tag{4.3}$$

其中$\sigma()$表示的是 Sigmoid 函数，tanh ()代表双曲正切函数（Hyperbolic Tangent Function），⊙表示的是点乘（Pointwise Multiplication）。h_{k-1}和c_{k-1}是前一次迭代

的隐藏状态和存储单元。i_k，f_k，o_k和g_k分别是 k 次循环子网络的输入门，忘记门，输出门和输入调制门的输出。

对于隐藏状态h_k，可以根据以下公式更新M_k：

$$z_k = relu(W_{hz}h_k + b_z)$$
$$M_k = W_{zm}z_k + b_m \tag{4.4}$$

其中M_k是 k+1 次循环的转换矩阵。

4.3.3 多注意力机制网络

CASN 和 IASN 之间相互合作、相互影响，形成自上至下的级联网络。在第一次循环 k = 1 时，CASN 从原始输入特征映射f_0定位一个粗略的区域f_1。然后 IASN 将局部区域f_1作为细粒度区域定位f_k的输入。它们在第 k 次循环表示如下：

$$f_1 = ST(f_0, M_0) \quad k = 1$$
$$f_k = ST(f_1, M_{k-1}) \quad k > 1 \tag{4.5}$$

LSTM 将特征映射f_k作为输入来计算存储单元和隐藏状态。它根据公式 4.3 可以获得第 k 次循环的h_k。给定隐藏状态h_k，本文使用公式 4.4 更新M_k。在第一次循环的时候，通过 STN 直接估算M_0，并在 CASN 中使用食品类别监督信息。基于M_0，本文在 IASN 中更新M_k。

4.3.4 多任务学习

本文最终在一个多任务公式中对本文的方法进行建模，主要通过两类损失进行优化，即类别分类损失L_{cls}和原食材性学习损失L_{ing}，用于生成大尺度的粗略图像区域和多个细粒度较小的图像区域。此外，本文利用另一种类型的损失L_{loc}来控制 STN 的注意力区域定位约束，以保证注意力区域的定位精度，从而产生以下损失函数：

$$L = L_{cls} + \gamma_1 L_{ing} + \gamma_2 L_{loc} \tag{4.6}$$

其中γ_1，γ_2是平衡参数。

在类别分类损失函数中，CASN 采用食品类别标签作为监督信息，指导 STN 定位注意力区域。经过 CASN 中的 LSTM，使用公式 4.2 让输入最终变为s_1，交

叉熵分类损失函数采用如下公式：

$$L_{cls} = -\frac{1}{N}\sum_i \log\Big(P\big((y_i|s_{1i})\big)\Big) \tag{4.7}$$

其中N是训练样本的数量，s_{1i}是第 i 个样本的特征表示，y_i是相应的食物类别。

在食材属性学习损失函数中，经过 IASN，可以获得每个循环的最终特征表示s_k，如下所示：

$$z_k = relu(W_{hz}h_k + b_z)$$

$$s_k = W_{zs}z_k + b_s \tag{4.8}$$

其中s_k是每个循环中食材的得分分布。在 IASN 中，我们可以获得每个循环的食材得分分布$\{s_2, \dots, s_k, \dots, s_K\}$, K表示的是循环次数，其中$s_K = \{s_K^1, \dots, s_K^v, \dots, s_K^V\}$，$V$表示有多少食材标签。对于食材得分分布本文使用最大池化，然后会得到最终的得分$\mathrm{s} = \{s^1, \dots, s^v, \dots, s^V\}$，$s^v$的计算过程如下：

$$s^v = \max(s_2^v, \dots, s_K^v), v = 1,2,3, \dots, V. \tag{4.9}$$

接着可以得到预测概率的向量p_i

$$p_i^v = \frac{\exp(s_i^v)}{\sum_{m=1}^{V}\exp(s_i^m)},\ \mathrm{v} = 1,2,\dots,\mathrm{V}. \tag{4.10}$$

最终食材属性学习损失函数表示为：

$$L_{ing} = \frac{1}{N}\sum_i\sum_v(p_i^v - P_i^v)^2 \tag{4.11}$$

其中$P_i = {}^{q_i}/_{||q_i||_1}$是第 i 个样本的真实标签概率向量，$q_i = \{q_i^1, \dots, q_i^2, \dots, q_i^V\}$，$q_i^v$是 0-1 表示的向量。

注意力区域局部定位损失函数类似于文献[74]，为了使 STN 成功地定位多样化的多尺度图像区域，本文还采用了以下三种类型的损失，包括锚（Anchor）约束，尺度约束和正约束。

对于锚约束，该约束使得注意力区域分散在食物图像中的不同语义区域上。它被表述为:

$$\Gamma_A = \frac{1}{2}\{(t_x^k - c_x^k)^2 + (t_y^k - c_y^k)^2\} \tag{4.12}$$

其中(c_x^k, c_y^k)为第 k 个锚点，t_x^k、t_y^k为 k 循环子网络的水平和垂直平移（$k \geq 2$）。

对于尺度约束，该约束用于将定位的注意力区域限定到一定范围内，可以表示为：

$$\Gamma_S = (\max(|s_x| - a, 0))^2 + (\max(|s_y| - a, 0))^2 \tag{4.13}$$

其中a是一个临界值，在不同的子网络中，a是不同的。在 CASN 中，因为它用于定位粗略图像区域，所以a相对大一点。相反，在 IASN 中a应该很小。这是因为它用于定位具有较小比例的细粒度图像区域。

对于正约束，此约束用于使注意区域不被镜像（Mirrored），可以表示为：

$$\Gamma_P = \max(0, \beta - s_x) + \max(0, \beta - s_y) \tag{4.14}$$

其中β是设定的阈值。

最终，注意力区域局部定位损失函数可以表示为：

$$L_{loc} = \Gamma_S + \lambda_1 \Gamma_A + \lambda_2 \Gamma_p \tag{4.15}$$

其中λ_1，λ_2是权重参数。

4.3.5　多尺度联合表示

这个模型一旦训练完毕，就可以在每张食品图像上获得多个从粗粒度到细粒度的注意力区域。在这个模型中，将会存在三种不同类型的区域：完整的食品图像、CASN 所产生的粗粒度区域和 IASN 产生的几个细粒度区域。本文针对每种类型的区域训练一个 CNN 模型。基于这些训练的 CNN 模型，本文从完整图像、粗粒度区域和细粒度区域中提取这三种不同粒度的特征：$\{F_0, F_1, \ldots, F_K\}$，其中$F_0$表示完整图像的视觉特征，K 为区域的总数。本文分别对每个特征进行归一化，然后将它们串联起来作为最终的特征表示，最后使用 Softmax 分类器进行食品图像的识别。

4.4 实验验证与分析

4.4.1 实验数据

ETH Food-101[13]是一个包含101种食物和101,000张图像的数据集。相关研究[54]进一步提供了相应的食材列表。本文的方法是使用食品图像的食材来定位注意力图像区域。因此，本文删除了食品图像那些不能看见的食材。最终可以看见的食材为174种。

VireoFood-172[78]包含172种食品、110,241张食品图像和353种食材。

ISIA Food-200包含200中食品和197,323张食品图像，其中每个类别至少有500张图像。为了验证本文方法的有效性，本文进一步在这个数据集上进行相关实验。类似于文献[78]，该数据集被划分为60%、10%和30%的图像分别用于训练、验证和测试。图4.2展示了一些样例。

图 4. 2 ISIA Food-200 的食品图像样例

Figure 4.2 Some food examples from ISIA Food-200

表4.1展示了三个数据集的相关统计。从表4.1中，我们可以看到ISIA Food-200在食物类别和图像数量上都大于ETH Food-101和VireoFood-172。 通过进一步观察，ISIA Food-200与其他两个数据集之间的共享类别非常少（ETH Food-101只有15个类别，VireoFood-172只有2个类别）。因此，ISIA Food-200与这两个数据集

互补，本文希望ISIA Food-200可以进一步促进食品图像相关领域的发展。

表 4.1　三个不同食品数据集的统计

Table 4.1 The statistics of three different datasets.

数据集	种类	图片量	食材
ETH Food-101[13]	101	101,000	174
VireoFood-172[78]	172	110,241	353
ISIA Food-200	200	197,323	319

4.4.2　评测指标

本文将使用Top-1和Top-5分类准确率作为评价指标。Top-1分类准确率表示测试图像中最有把握的预测标签正确的百分比。Top-5分类准确率表示真实标签在排名前5位的预测标签中的测试图像所占的百分比。因为多项研究工作都使用Top-1和Top-5分类准确率，为了公平对比，本文也将Top-1和Top-5分类准确率作为评价指标。

4.4.3　实现细节

本文将VGG-16最后一层卷积层的特征图（Feature Map）当作STN的输入。所有训练图像均二线性插值到224×224大小。在训练过程中，使用Adam优化器来优化整个网络，输入图片的批量（Batch size）为16，Adam优化器的动量（Momentum）参数设置为0.9和0.999。整个模型的初始学习率设置为0.00001，然后每30个周期除以10。本文选择在验证集上损失函数最小的模型进行测试。因为ETH Food-101没有验证集，所以选择训练损失函数不再改变时的模型用来测试。

对于模型中多任务学习的参数，本文使用标准的反向传播进行优化。分类交叉熵损失函数和食材属性损失函数设置相同的权重，并不需要预先定义。因此，λ_1为1.0。根据以往的经验，将ETH Food-101，VireoFood-172和ISIA Food-200数据集的λ_2分别设置为0.1、0.5和0.5。在公式4.12中，本文在IASN中将局部细粒度图像区域的数量设置为5。因此，除了中心点坐标（0, 0）之外，根据以往经验，还将四个锚点坐标分别设置为（0.4, 0.4）、（0.4, -0.4）、（-0.4, 0.4）和（-0.4，-0.4），即整个图像的左上角、右上角、左下角和右下角，这样能全面的覆盖整张食品图像。在公式4.13中，对于CASN和IASN，将α分别设置为0.9和0.5。在CASN

中，a应该较大，因为CASN是用来定位粗粒度图像区域的。相反，IASN是用来定位细粒度图像区域，因此IASN中的a应该较小。在公式4.14中，根据以往经验，将CASN和IASN中的β分别设置为0.6和0.1。在公式4.15中，当在ETH Food-101数据集上进行训练时，将λ_1和λ_2分别设置为0.01和0.5，当在VireoFood-172数据集上进行训练时，将λ_1和λ_2分别设置为1和1，当在ISIA Food-200数据集上进行训练时，将λ_1和λ_2分别设置0.2和0.2。

模型一旦训练完毕，本文就可以在每张食品图像上获得多个从粗粒度到细粒度的注意力区域。然后分别对CASN和IASN生成的区域和整张食品图像使用DenseNet-161网络来提取特征，本文只需将它们简单的串联来进行特征融合。

作为融合后最终的特征表示，本文采用Softmax分类器进行分类。在实验过程中，本文均采用相同的特征串联方法进行特征融合。

4.4.4 性能对比

本文首先对 IASN 中多个区域的特征融合进行了性能比较。在本文实验中，IASN 定位了几个细粒度的图像区域。本文比较了不同区域的融合结果，并在表 4.2 中展示了对比实验结果。对于 IASN 定位的细粒度区域，本文将细粒度区域 1-C 表示为 C 个连续定位区域。比如细粒度区域 1-1 是第一个细粒度区域，而细粒度区域 1-5 表示所有局部的细粒度区域。如表 4.2 所示，我们可以看到（1）当越来越多的局部区域进行特征融合时，识别性能会逐渐提高。（2）所有区域的特征融合性能是最好的，并且这种最佳结果得益于多个细粒度图像区域的互补性。

接着，本文对模型中不同的组成进行性能比较。表 4.3 展示了模型不同组成的实验结果。我们可以看到，与单独的 CASN 或 IASN 相比，CASN 的粗粒度区域和 IASN 的细粒度区域进行特征融合后可以进一步提高性能。在融合了整张图像中的特征之后，本文的方法在 Top-1 和 Top-5 分类准确率上均达到了最佳性能。因此，可以得出结论：整张图像、CASN 和 IASN 中不同粒度的区域具有互补性，而且来自不同类型的特征融合会生成更加全面、更具有判别性的特征表示。

表 4.2　IASN 中不同区域特征融合在 ETH Food-101 数据集的性能（%）

Table 4.2 Performance comparison on feature fusion from different regions in Ingredient-supervised Attention Sub-Network (IASN) on ETH Food-101 (%)

细粒度区域的融合	Top-1 分类准确率	Top-5 分类准确率
细粒度区域 1-1	83.53	96.03
细粒度区域 1-2	86.50	97.17
细粒度区域 1-3	87.17	97.35
细粒度区域 1-4	88.27	97.71
细粒度区域 1-5	**88.94**	**97.87**

表 4.3　模型不同组成在 ETH Food-101 数据集的性能（%）

Table 4.3 Performance comparison for different components of model on ETH Food-101 (%)

不同组成的融合	Top-1 分类准确率	Top-5 分类准确率
CASN	85.41	96.67
IASN	88.94	97.87
CASN+IASN	89.89	98.21
IG-CMAN	**90.37**	**98.42**

最后，本文还与食品识别的最新技术进行了性能比较。如表 4.4 所示，在表中列出了在 ETH Food-101 食品数据集上最新相关方法。在表中还展示了在不同 CNN 上的性能，例如 AlexNet、InceptionV3、ResNet、DenseNet 和 WRN。从表 4.4 中，我们可以看到（1）WRN 的性能优于其他卷积神经网络。(2）WISeR 通过增加带有切片卷积层的切片网络分支来改善 WRN，该网络可以捕获食品图像特定的垂直结构。(3）本文的方法在 Top-1 分类准确率上达到了最好性能，而且比专门为食品识别设计的 WISeR 网络在 Top-1 分类准确性上提高 0.1%。尽管识别性能有所提高，但本文方法并未使用其他数据增强方式，而 WISeR 网络另外使用了各种各样的亮度增强和 AlexNe 式的色彩增强，并且使用了 10-crop 进行测试。

表 4.4　IG-CMAN 在 ETH Food-101 数据集的性能（%）

Table 4.4 Comparison of our model and state-of-the-art methods on ETH Food-101 (%)

方法	Top-1 分类准确率	Top-5 分类准确率
AlexNet-CNN[13]	56.40	-
DCNN-FOOD[27]	70.41	-
DeepFood[79]	77.40	93.70
FCAN[80]	86.50	-
CurriculumNet[81]	87.30	-
Inception V3[29]	88.28	96.88
ResNet-200[31]	88.38	97.85
DenseNet-161[63]	86.94	97.03
WRN[34]	88.72	97.92
WISeR[34]	90.27	**98.71**
IG-CMAN	**90.37**	98.42

本文除了在 ETH Food-101 食品数据集上验证方法的有效性，还在另外 1 个中国菜 VireoFood-172 数据集上做了相关对比实验，来验证本文方法的鲁棒性。本文首先比较 IASN 中不同区域特征融合的实验结果。如表 4.5 所示，我们可以看到 IASN 中所有局部区域的特征融合均实现了最好性能。表 4.6 进一步展示了本文方法在 VireoFood-172 食品数据集上模型不同组成的实验结果。通过三种类型区域的特征融合，本文在 Top-1 和 Top-5 分类准确率上均达到了最好的识别性能。表 4.7 还展示了与其他方法识别准确率对比结果。我们可以看到，本文方法在 Top-1 和 Top-5 分类准确率在这个数据集上都达到了最好性能。与具有相同基础网络的多任务方法[78]相比，Top-1 和 Top-5 分类准确率分别提高了约 3.4%和 1.1%。这种性能的提升主要来自语义注意力区域和多个注意力区域的融合。

为了进一步验证本文方法的有效性，本文还额外构建了 ISIA Food-200 数据集。因此，本文同样在 ISIA Food-200 数据集做了相关对比实验。如表 4.8 所示，我们可以看到 IASN 中所有局部区域的特征融合均实现了最佳性能。表 4.9 进一步展示 ISIA Food-200 上本文方法不同组成的实验结果。同样，本文方法在 Top-1 和 Top-5 分类准确率上都达到了最佳性能。表 4.10 还展示了与其基线识别准确率的对比结果。由于 ISIA Food-200 是新提出的食品数据集，因此本文在不同 CNN 上进行了基线实验。从表 4.10 中可以看出，本文方法在 Top-1 和 Top-5 分类准确率都达到了最佳性能。实验结果再次验证了本文提出方法的有效性。

表 4.5　IASN 中不同区域特征融合在 VireoFood-172 数据集的性能（%）

Table 4.5 Performance comparison on feature fusion from different regions in Ingredient-supervised Attention Sub-Network (IASN) on VireoFood-172 (%)

细粒度区域的融合	Top-1 分类准确率	Top-5 分类准确率
细粒度区域 1-1	82.35	95.35
细粒度区域 1-2	85.96	96.92
细粒度区域 1-3	87.47	97.46
细粒度区域 1-4	88.83	97.91
细粒度区域 1-5	**89.43**	**98.06**

表 4.6　模型不同组成在 VireoFood-172 数据集的性能（%）

Table 4.6 Performance comparison for different components of model on VireoFood-172 (%)

不同组成的融合	Top-1 分类准确率	Top-5 分类准确率
CASN	87.39	97.15
IASN	89.43	98.06
CASN+IASN	90.34	98.31
IG-CMAN	**90.63**	**98.40**

表 4.7　IG-CMAN 在 VireoFood-172 数据集的性能（%）

Table 4.7 Comparison of our model and state-of-the-art methods on VireoFood-172 (%)

方法	Top-1 分类准确率	Top-5 分类准确率
AlexNet	64.91	85.32
VGG-16	80.41	94.59
DenseNet-161	86.93	97.17
MultiTaskDCNN(VGG-16)[78]	82.06	95.88
MultiTaskDCNN(DenseNet-161)[78]	87.21	97.29
IG-CMAN	**90.63**	**98.40**

表 4.8　IASN 中不同区域特征融合在 ISIA Food-200 数据集的性能（%）

Table 4.8 Performance comparison on feature fusion from different regions in Ingredient-supervised Attention Sub-Network (IASN) on ISIA Food-200 (%)

细粒度区域的融合	Top-1 分类准确率	Top-5 分类准确率
细粒度区域 1-1	58.88	86.18
细粒度区域 1-2	62.09	88.36
细粒度区域 1-3	63.29	89.33
细粒度区域 1-4	64.39	89.92
细粒度区域 1-5	**65.59**	**90.70**

表 4.9　模型不同组成在 ISIA Food-200 数据集的性能（%）

Table 4.9 Performance comparison for different components of model on ISIA Food-200 (%)

不同组成的融合	Top-1 分类准确率	Top-5 分类准确率
CASN	61.13	87.66
IASN	65.59	90.70
CASN+IASN	66.71	91.45
IG-CMAN	**67.47**	**91.75**

表 4.10　IG-CMAN 在 ISIA Food-200 数据集的性能（%）

Table 4.10 Comparison of our model and baselines on ISIA Food-200 (%)

方法	Top-1 分类准确率	Top-5 分类准确率
AlexNet	49.34	79.30
VGG-16	59.05	86.53
ResNet-152	61.07	87.87
DenseNet-161	62.62	88.28
IG-CMAN	**67.47**	**91.75**

4.4.5　定性分析和可视化

为了更深入地了解这些实验结果，本文对模型的注意力区域进行了定性分析以及注意力区域的可视化。本文首先展示了 IASN 中的细粒度局部区域。如图 4.3 所示，展示了一些 IASN 中细粒度区域的样例，在图中每个区域图像的下面，本文展示了基于 IASN 中每个细粒度区域排列前 3 的食材概率分布。我们可以观察到，这些局部区域对相应的食品类别是具有判别性的。另外，许多局部区域对应的语义食材都具有可解释性。比如，在图 4.3 中的第一行中，我们可以观察到许

多局部区域分别对应于其概率分布最高的食材。第一个局部区域是鸡肉，则食材鸡肉是预测的食材概率分布中最高的概率。第三个局部区域是大豆，预测的食材大豆则具有最高的概率。因此，本文所提出的 IASN 能够找到食材对应的区域，而且这些区域具有判别性。图 4.4 进一步展示了更多的粗粒度和细粒度局部区域的图像示例。从图 4.4 中可看出基于 CASN 的粗粒度区域能够去除一些食品图像复杂的背景信息，而基于 IASN 的细粒度区域能够定位到食品图像食材对应的区域。从粗粒度到细粒度区域都具有判别性，而且二者更具有互补性，因此能提高识别性能。

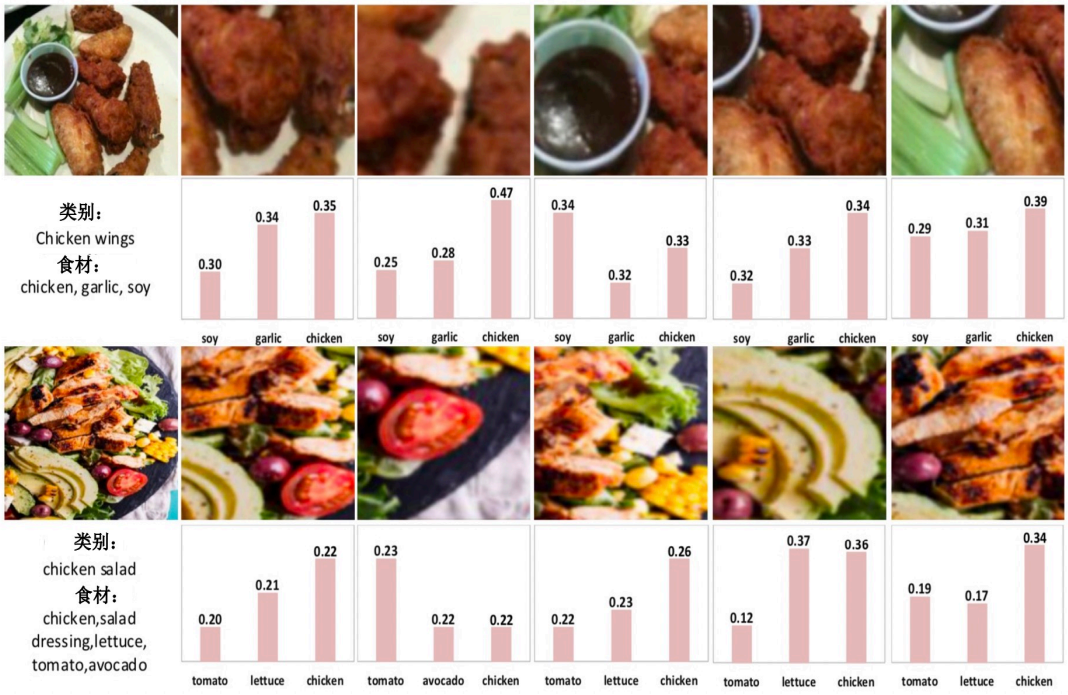


图 4. 3 IASN 中食材概率分布前 3 的局部图像区域

Figure 4.3 Localized image regions with probability distribution on top-3 ingredients from some food examples in IASN (Ingredient-supervised Attention Sub-Network).

图 4. 4 IG-CMAN 定位到的局部区域样例

Figure 4.4 Localized regions from some samples in the proposed model.

4. 4. 6 讨论

在本文的方法中，该方法能够使用食品图像特有的食材信息来明确地指导网络发现多样化的不同粒度的注意力区域。而且，许多局部区域对应于语义食材。因此，本文所提方法的局部区域具有可解释性。但是，在食品图像中并非总是如此，比如混合的食材没有明确的分布，食材太小以及食材的空间结构发生了变化（如图 4.2 所示）。在这些情况下，本文的方法无法定位食材的语义区域。即使这些局部区域无法解释，但是它们仍然可以提供互补的视觉信息。因此，仍然有助于食品图像的识别。

另一方面，本文在 IASN 中预定义了五个区域，因此 IASN 只能定位五个细粒度的图像区域。如表 4.2，表 4.5 和表 4.8 所示，我们可以看到，随着局部细粒度区域数量的增加，识别性能也不断提高。因此，可以推出随着局部区域数量的增加，性能可能会提高。但是，随着区域数量的增加，网络训练需要更多的时间成本和 GPU 资源。因此，需要进一步研究设置多少个区域才会达到最佳性能。如何在网络复杂性和局部区域数量之间取得平衡也是值得探索的问题。

4.5　小结

在本文中，提出了一种 IG-CMAN 方法来进行食品图像识别。通过类别信息监督和食材信息监督将 STN 和 LSTM 结合使用，它能够顺序定位到不同尺度图像上的不同注意力区域。该方法首先从 CASN 中生成初始粗粒度的注意力区域，该区域能够去除原始图像一些复杂的背景信息。基于这个初始粗粒度的注意力区域，IASN 能够找到食材对应的细粒度区域。最后，将这些区域的特征融合成最终的特征表示，用 Softmax 分类器进行食品图像识别。实验结果表明，来自粗粒度区域和细粒度区域的融合特征具有互补性，而且更具判别性。此外，本文还构建了一个新的食品数据集 ISIA Food-200，它与现有的食品图像数据集非常互补，例如 ETH Food-101 和 VireoFood-172。在这两个主流的食品图像数据集和新构建的 ISIA Food-200 上综合实验结果表明，本文的方法实现了最好的识别性能。性能改进得益于多语义注意力区域的融合。

第 5 章　总结与展望

5.1　总结

如今，社交网络、移动网络和物联网等各种网络的快速发展，用户可以轻松地通过这些网络共享食物图像，食谱，烹饪视频或记录食物日记，从而形成大规模的食物数据集。这些食物数据意味着丰富的知识，因此可以为与食物相关的研究提供巨大的机会，本文基于食品图像特有的食材信息对食品图像识别进行了相关研究，具体取得的工作进展如下：

（1）本文提出了一种基于多尺度多视角特征融合（MSMVFA）方法来实现食品图像识别，其中多视角意味着不同类型的特征。考虑到食物通常不会表现出明显的空间排列方式，因此针对每种类型的特征采用了多尺度融合方法。最粗粒度的尺度是整张食品图像，因此保留了全局空间布局，而更细粒度的尺度可以捕获到食物图像的更多局部细粒度细节。因此，这样的融合特征对于食品图像几何变形更加鲁棒。基于每种特征类型的多尺度表示，MSMVFA 可以进一步将高层语义特征，中层属性特征和深层视觉特征融合成统一的特征表示。这三种类型的特征可以从不同的粒度来表示食品图像。因此，融合的特征可以最大可能地捕获其语义信息。

（2）本文提出了一种基于食材指导的级联多注意力网络（IG-CMAN）来实现食品图像识别，该方法能够以粗粒度到细粒度、基于类别信息和食材信息的多任务方式从食品图像中找到多个食品图像的区域。在 CASN 中，该方法通过 STN 生成去除复杂背景信息的初始区域。然后以该局部注意力区域为基础，将 STN 和 LSTM 结合起来，从以下几个层级的 IASN 中依次发现具有细粒度的多注意力区域。实验结果表明这种具有不同粒度的注意力区域是有互补性的，能最大可能地的提高识别性能。

（3）本文构建了一个包含食材的新食品图像数据集。要构建此数据集，本文首先根据 Wikipedia 的“按食材分类的食品列表”来建立食品类别的词汇表。然后，将食品名称用作查询，以检索不同图像搜索引擎（例如 Google 和 Bing）的候选食品图像来提高图像视觉的多样性。接着，通过人工标注删除了不相关的具有噪音的食品图像。最终共获得 197,323 张食品图像和 319 种可见的食材标注。

5.2 展望

虽然本文基于食材信息的食品图像识别研究取得一定的成果，但是在当前数据爆炸的时代下需要我们亟待解决和深入研究不断出现的新问题。展望未来的工作，本文认为可以从以下三个方面继续探索和研究:

（1）**构建大规模标准食品数据集。**像 ImageNet 用于计算机视觉中的一般物体一样，ImageNet 这样级别的食品图像数据集是开发高级算法的关键，比如基于内容的大规模食品图像搜索，分类和理解算法等。构建大规模的食品数据集，一种可行的方法是将社交媒体上的食品图像抓取与众包平台 AMT 的手动标注相结合。另外，应该考虑到食品图像的地理分布，例如不同的美食，以覆盖整个世界。每个地区都有自己的特色美食和菜肴，没有食物专家来掌握所有菜肴。因此，大规模食品数据集的建设也需要全世界研究人员的共同努力。

（2）**鲁棒的大规模食品识别系统。**基于视觉的食品系统对于各种实际应用（例如膳食评估和管理系统）非常重要。首要任务是开发一个大型的，鲁棒的食品识别系统。近年来，CNN 等深度学习方法及其变体为我们提供实现此目标的绝佳机会。与使用传统的人工特征相比，深度学习的优势在于可以从多层体系结构的原始图像像素中逐步自动地学习更多抽象特征。有些研究工作正朝着这个方向努力。例如，文献[34]提出了一种切片卷积网络来捕获食品垂直结构，并结合 CNN 的视觉特征以实现最好性能。本文作者相信还有其他特殊的食品结构和特性可供探索。如果设计的深度模型可以从不同方面捕获食品图像的结构，则性能将得到进一步改善。而且，构建的大规模标准食品数据集对于推进食品识别系统的开发也至关重要。根据维基百科统计，全球有 8,000 多种菜肴，但是与大量的菜式相比，食材的数量是有限的。因此，一种替代解决方案是食材识别。一些研究工作已经根据食品图像的食材列表进行了多标签食材预测。食材识别也可能是一种解决方案，来提供一种自动机制来识别图像以简化对营养习惯的监控，从而实现更准确的饮食评估。

（3）**面向食品的多模态知识图构建和推理。**可以使用复杂的数据分析技术来开发与食品相关的大量数据，发现其共同模式和新的知识。但是，异构模式进行更复杂的面向食品检索，问题解答和推理，一种有效的解决方法是构建丰富的上下文多模态知识图。在自然语言处理中，已经展示了一些有价值的研究

工作，比如语义网络技术（例如推理机制）已用于糖尿病饮食护理。关于三连体网络之间的视觉关系研究已经出现在计算机视觉领域，包括视觉关系的检测和场景图的生成。这些技术有助于构建可视化网络。其他研究工作试图建立一个大规模的多模态知识库系统来支持视觉查询，并被证明是一种构建面向食品多模态知识图的有效方法。这样的多模态知识图对来自各种异构数据源的食品数据是很有用的。而且还可以基于知识图进行推理，通过推理引擎支持复杂查询、问题解答和多模态对话。

参考文献

[1] Fitzmaurice C, Allen C, Barber R M. A systematic analysis for the global burden of disease study[J]. JAMA Oncol, 2017, 3(4): 524-548.

[2] Khanna S K. Food and Culture: A Reader, by Carole Counihanand Penny Van Esterik: New York: Routledge, 608 pp[J]. 2009, (2009): 157-159

[3] Sajadmanesh S, Jafarzadeh S, Ossia S A, et al. Kissing cuisines: Exploring worldwide culinary habits on the web[C]. Proceedings of the International Conference on World Wide Web Companion. 2017: 1013-1021.

[4] Chung J, Chung J, Oh W, et al. A glasses-type wearable device for monitoring the patterns of food intake and facial activity[J]. Scientific Reports, 2017, 7: 41690.

[5] Nestle M, Wing R, Birch L, et al. Behavioral and Social Influences on Food Choice[J]. 1998，Nutrition Reviews(5):5.

[6] Sørensen L B, Møller P, Flint A, et al. Effect of sensory perception of foods on appetite and food intake: a review of studies on humans[J]. International Journal of Obesity, 2003, 27(10): 1152-1166.

[7] Pauly D. A simple method for estimating the food consumption of fish populations from growth data and food conversion experiments[J]. 1986.

[8] Chen Z, Tao Y. Food safety inspection using “from presence to classification” object-detection model[J]. Pattern Recognition, 2001, 34(12): 2331-2338.

[9] Harris M . Comment on Vayda's review of good to eat: Riddles of food and culture[J]. Human Ecology, 1987, 15(4):511-517.

[10] Mouritsen O G, Edwards-Stuart R, Ahn Y Y, et al. Data-driven methods for the study of food perception, preparation, consumption, and culture[J]. Frontiers in ICT, 2017, 4: 15.

[11] Sajadmanesh S, Jafarzadeh S, Ossia S A, et al. Kissing cuisines: Exploring worldwide culinary habits on the web[C]. Proceedings of the International Conference on World Wide Web Companion. 2017: 1013-1021.

[12] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[13] Bossard L, Guillaumin M, Van Gool L. Food-101–mining discriminative components with random forests[C]. European conference on computer vision. Springer, Cham, 2014: 446-461.

[14] Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]. Proceedings of the Conference on Computer Vision. 2017: 5209-5217.

[15] Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2017: 4438-4446.

[16] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[17] Wu W, Yang J. Fast food recognition from videos of eating for calorie estimation[C]. Proceedings of the International Conference on Multimedia and Expo. 2009: 1210-1213.

[18] Anthimopoulos M M, Gianola L, Scarnato L, 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.

[19] Yang S, Chen M, Pomerleau D, et al. Food recognition using statistics of pairwise local features[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2010: 2249-2256.

[20] Joutou T, Yanai K. A food image recognition system with multiple kernel learning[C]. Proceedings of the International Conference on Image Processing. 2009: 285-288.

[21] Hoashi H, Joutou T, Yanai K. Image recognition of 85 food categories by feature fusion[C]. Proceedings of the International Symposium on Multimedia. 2010: 296-301.

[22] Nguyen D T, Zong Z, Ogunbona P O, et al. Food image classification using local appearance and global structural information[J]. Neurocomputing, 2014, 140: 242-251.

[23] Martinel N, Piciarelli C, Micheloni C, et al. A structured committee for food recognition[C]. Proceedings of the International Conference on Computer Vision Workshops. 2015: 92-100.

[24] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Proceedings of the International Conference on Advances in Neural Information Processing Systems. 2012: 1097-1105.

[25] Kagaya H, Aizawa K, Ogawa M. Food detection and recognition using convolutional neural network[C]. Proceedings of the International Conference on Multimedia. 2014: 1085-1088.

[26] Kawano Y, Yanai K. Food image recognition with deep convolutional features[C]. Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication. 2014: 589-593.

[27] Yanai K, Kawano Y. Food image recognition using deep convolutional network with pre-training and fine-tuning[C]. Proceedings of the International Conference on Multimedia & Expo Workshops. 2015: 1-6.

[28] Wu H, Merler M, Uceda-Sosa R, et al. Learning to make better mistakes: Semantics-aware visual food recognition[C]. Proceedings of the International Conference on Multimedia. 2016: 172-176.

[29] Hassannejad H, Matrella G, Ciampolini P, et al. Food image recognition using very deep convolutional networks[C]. Proceedings of the International Workshop on Multimedia Assisted Dietary Management. 2016: 41-49.

[30] Ming Z Y, Chen J, Cao Y, et al. Food photo recognition for dietary tracking: System and experiment[C]. Proceedings of the International Conference on Multimedia Modeling. 2018: 129-141.

[31] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[32] Pandey P, Deepthi A, Mandal B, et al. FoodNet: Recognizing foods using ensemble of deep networks[J]. IEEE Signal Processing Letters, 2017, 24(12): 1758-1762.

[33] McAllister P, Zheng H, Bond R, 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.

[34] Martinel N, Foresti G L, Micheloni C. Wide-slice residual networks for food recognition[C]. Proceedings of the International Conference on Applications of Computer Vision. 2018: 567-576.

[35] Zagoruyko S, Komodakis N. Wide residual networks[J]. arXiv preprint arXiv:1605.07146, 2016.

[36] Wang X, Kumar D, Thome N, et al. Recipe recognition with large multimodal food dataset[C]. Proceedings of the International Conference on Multimedia & Expo Workshops. 2015: 1-6.

[37] Zhang M M. Identifying the Cuisine of a Plate of Food[J]. Univ. of California at San Diego, La Jolla, CA, USA, Tech. Rep. CSE, 2011, 190.

[38] Su H, Lin T W, Li C T, et al. Automatic recipe cuisine classification by ingredients[C]. Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: adjunct publication. 2014: 565-570.

[39] 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, 2016, 19(5): 1100-1113.

[40] Xiao G, Wu Q, Chen H, et al. A Deep Transfer Learning Solution for Automating Food Material Procurement using Electronic Scales[J]. IEEE Transactions on Industrial Informatics, 2019.

[41] Druck G. Recipe attribute prediction using review text as supervision[C]. Cooking with Computers 2013, IJCAI workshop. 2013.

[42] Matsuda Y, Hoashi H, Yanai K. Recognition of multiple-food images by detecting candidate regions[C]. Proceedings of the International Conference on Multimedia and Expo. 2012: 25-30.

[43] Matsuda Y, Yanai K. Multiple-food recognition considering co-occurrence employing manifold ranking[C]. Proceedings of the International Conference on Pattern Recognition. 2012: 2017-2020.

[44] Ragusa F, Tomaselli V, Furnari A, et al. Food vs non-food classification[C]. Proceedings of the International Workshop on Multimedia Assisted Dietary Management. 2016: 77-81.

[45] Singla A, Yuan L, Ebrahimi T. Food/non-food image classification and food categorization using pre-trained googlenet model[C]. Proceedings of the International Workshop on Multimedia Assisted Dietary Management. 2016: 3-11.

[46] Aguilar E, Bolaños M, Radeva P. Exploring food detection using CNNs[C]. Proceedings of the International Conference on Computer Aided Systems Theory. Springer, Cham, 2017: 339-347.

[47] Bolaños M, Radeva P. Simultaneous food localization and recognition[C]. Proceedings of the International Conference on Pattern Recognition. 2016: 3140-3145.

[48] Aguilar E, Remeseiro B, Bolaños M, et al. Grab, pay, and eat: semantic food detection for smart restaurants[J]. IEEE Transactions on Multimedia, 2018, 20(12): 3266-3275.

[49] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[50] Anthimopoulos M, Dehais J, Diem P, et al. Segmentation and recognition of multi-food meal images for carbohydrate counting[C]. Proceedings of the International Conference on BioInformatics and BioEngineering. 2013: 1-4.

[51] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[52] Chen H, Xu J, Xiao G, et al. Fast auto-clean CNN model for online prediction of food materials[J]. Journal of Parallel and Distributed Computing, 2018, 117: 218-227.

[53] Pan L, Pouyanfar S, Chen H, et al. Deepfood: Automatic multi-class classification of food ingredients using deep learning[C]. Proceedings of the Conference on Collaboration and Internet Computing. 2017: 181-189.

[54] Bolaños M, Ferrà A, Radeva P. Food ingredients recognition through multi-label learning[C]. Proceedings of the International Conference on Image Analysis and Processing. 2017: 394-402.

[55] Zhang X, Zhou F, Lin Y, et al. Embedding label structures for fine-grained feature representation[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2016: 1114-1123.

[56] Zhang X J, Lu Y F, Zhang S H. Multi-task learning for food identification and analysis with deep convolutional neural networks[J]. Journal of Computer Science and Technology, 2016, 31(3): 489-500.

[57] Zhou F, Lin Y. Fine-grained image classification by exploring bipartite-graph labels[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2016: 1124-1133.

[58] Zhu F, Bosch M, Schap T R, et al. Segmentation assisted food classification for dietary assessment[C]. Computational Imaging IX. International Society for Optics and Photonics, 2011, 7873: 78730B.

[59] Kong F, Tan J. Dietcam: Regular shape food recognition with a camera phone[C]. Proceedings of the International Conference on Body Sensor Networks. 2011: 127-132.

[60] Oliveira L, Costa V, Neves G, et al. A mobile, lightweight, poll-based food identification system[J]. Pattern Recognition, 2014, 47(5): 1941-1952.

[61] Ravì D, Lo B, Yang G Z. Real-time food intake classification and energy expenditure estimation on a mobile device[C]. Proceedings of the International Conference on Wearable and Implantable Body Sensor Networks. 2015: 1-6.

[62] Kawano Y, Yanai K. Foodcam: A real-time food recognition system on a smartphone[J]. Multimedia Tools and Applications, 2015, 74(14): 5263-5287.

[63] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[64] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[65] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.

[66] Tanno R, Okamoto K, Yanai K. DeepFoodCam: A DCNN-based real-time mobile food recognition system[C]. Proceedings of the International Workshop on Multimedia Assisted Dietary Management. 2016: 89-89.

[67] Pouladzadeh P, Shirmohammadi S. Mobile multi-food recognition using deep learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications. 2017, 13(3s): 1-21.

[68] Bettadapura V, Thomaz E, Parnami A, et al. Leveraging context to support automated food recognition in restaurants[C]. Proceedings of the International Conference on Applications of Computer Vision. 2015: 580-587.

[69] Xu R, Herranz L, Jiang S, et al. Geolocalized modeling for dish recognition[J]. IEEE Transactions on Multimedia, 2015, 17(8): 1187-1199.

[70] Herranz L, Jiang S, Xu R. Modeling restaurant context for food recognition[J]. IEEE Transactions on Multimedia, 2016, 19(2): 430-440.

[71] Wang H, Min W, Li X, et al. Where and what to eat: Simultaneous restaurant and dish recognition from food image[C]. Pacific Rim Conference on Multimedia. Springer, Cham, 2016: 520-528.

[72] Bolaños M, Valdivia M, Radeva P. Where and What Am I Eating? Image-Based Food Menu Recognition[C]. Proceedings of the European Conference on Computer Vision. 2018: 590-605.

[73] Wei Z, Chen J, Ming Z, et al. DietLens-Eout: Large Scale Restaurant Food Photo Recognition[C]. Proceedings of the International Conference on Multimedia Retrieval. 2019: 399-403.

[74] Wang Z, Chen T, Li G, et al. Multi-label image recognition by recurrently discovering attentional regions[C]. Proceedings of the International Conference on Computer Vision. 2017: 464-472.

[78] Chen J, Ngo C W. Deep-based ingredient recognition for cooking recipe retrieval[C]. Proceedings of the International Conference on Multimedia. 2016: 32-41.

[79] Liu C, Cao Y, Luo Y, et al. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment[C]. Proceedings of the International Conference on Smart Homes and Health Telematics. 2016: 37-48

[80] Liu X, Xia T, Wang J, et al. Fully convolutional attention networks for fine-grained recognition[J]. arXiv preprint arXiv:1603.06765, 2016.

[81] Guo S, Huang W, Zhang H, et al. Curriculumnet: Weakly supervised learning from large-scale web images[C]. Proceedings of the European Conference on Computer Vision. 2018: 135-150.
[82] Zhang N, Paluri M, Ranzato M A, et al. Panda: Pose aligned networks for deep attribute modeling[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2014: 1637-1644.
[83] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition. 2015: 3730-3738.
[84] Chen X, Zhu Y, Zhou H, et al. Chinesefoodnet: A large-scale image dataset for chinese food recognition[J]. arXiv preprint arXiv:1705.02743, 2017.
[85] Gong Y, Wang L, Guo R, et al. Multi-scale orderless pooling of deep convolutional activation features[C]. Proceedings of the European Conference on Computer Vision. 2014: 392-407.
[86] Song X, Jiang S, Herranz L. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2721-2735.
[87] Papyan V, Elad M. Multi-scale patch-based image restoration[J]. IEEE Transactions on image processing, 2015, 25(1): 249-261.
[88] Min W, Jiang S, Jain R C. Food Recommendation: Framework, Existing Solutions and Challenges[J]. IEEE Transactions on Multimedia, 2019.
[89]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.
[90] Min W, Jiang S, Liu L, et al. A survey on food computing[J]. ACM Computing Surveys, 2019, 52(5): 1-36.

致 谢

时间匆匆，研究生生涯即将结束。三年前我跟随自己的兴趣爱好从工业工程专业跨到计算机技术专业，三年多的时间，已经完成从对计算机的一知半解到专注于自己的研究方向的蜕变，期间收获的不仅仅的是知识，还有道不尽的友情、师生情。

在三年多的学习生活中感谢二位导师对我的悉心教导。首先要衷心感谢我的导师蒋树强研究员。蒋老师带我进入食品图像研究领域，并不遗余力的为我们创造自由的科研环境。蒋老师严谨的科研作风，谦和的处事态度是我终身学习的榜样。

其次感谢闵巍庆老师在生活和学习中对我的帮助，每当我在生活或者科研上有解不开的难题时，闵老师都能悉心指导和给予帮助，让我又有了新的动力去迎接挑战，科研之路才得以顺利。

感谢实验室师兄弟师姐妹们和同窗好友在生活上和科研上对我的关心和帮助，昔日一起探讨学术的场景历历在目。

最后感谢我的爸妈，对我工作三年后继续读书的理解和支持，感谢他们含辛茹苦的养育。

刘林虎

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：

2010 年 9 月——2014 年 6 月，在郑州航空工业管理学院管理科学与工程学院获得学士学位。

2017 年 9 月——2020 年 6 月，在中国科学院大学人工智能学院攻读硕士学位。

已发表（或正式接受）的学术论文：

[1] Weiqing Min, **Linhu Liu**, Zhengdong Luo, Shuqiang Jiang. Ingredient-Guided Cascaded Multi-Attention Network for Food Recognition. ACM Multimedia 2019: 1331-1339 **(CCF-A)**

[2] Weiqing Min, Shuqiang Jiang, **Linhu Liu**, Yong Rui, Ramesh C. Jain. A Survey on Food Computing. ACM Computing Surveys. 52(5): 92:1-92:36 (2019)

[3] Shuqiang Jiang, Weiqing Min, **Linhu Liu**, Zhengdong Luo. Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition. IEEE Transactions Image Processing 29: 265-276 (2019) **(CCF-A)**

[4] Weiqing Min, Shuhuan Mei, **Linhu Liu**, Yi Wang, Shuqiang Jiang. Multi-Task Deep Relative Attribute Learning for Visual Urban Perception. IEEE Transactions Image Processing 29: 657-669 (2019) **(CCF-A)**

[5] Shuqiang Jiang, Gongwei Chen, Xinhang Song, **Linhu Liu**. Deep Patch Representations with Shared Codebook for Scene Classification. ACM Transactions on Multimedia Computing, Communications, and Applications 15(1s): 5:1-5:17 (2019) **(CCF-B)**

[6] Shuqiang Jiang, Weiqing Min, Yongqiang Lv, **Linhu Liu**. Few-Shot Food Recognition via Multi-View Representation. ACM Transactions on Multimedia Computing, Communications and Applications (2020, Accepted) **(CCF-B)**

[7] 梅舒欢，闵巍庆，**刘林虎**, 等. 基于 Faster R-CNN 的食品图像检索和分类[J]. 南京信息工程大学学报(自然科学版)，2017，9(06):73-79.

申请或已获得的专利：

蒋树强，**刘林虎**，闵巍庆. 一种训练食品图像分类模型的方法及图像分类方法. 201911152246X