基于成分引导的多模态自蒸馏食品图像分割

侯素娟[1]， 孙月娟[1]， 闵巍庆[2*]， 王瑞平[2]， 蒋树强[2]
([1]山东师范大学 济南 250358
[2]中国科学院计算技术研究所 北京 100190)

摘要 目的：随着计算机视觉技术的发展，精确地识别并分割食品图像中的不同成分区域，对于食品营养分析和促进饮食健康管理至关重要。然而，当前图像分割模型多依赖于单一图像输入，这一做法在处理视觉差异较小的食品图像时，往往难以捕捉到细微的区分特征，从而影响分割精度。本文旨在解决单一模态在分割任务中的不足，利用文本信息为模型提供更加丰富的上、下文信息，采用自蒸馏技术，引导模型对食品图像的有效分割。方法：提出一种基于成分信息引导的多模态自蒸馏分割模型。该模型采用对比语言文本预训练模型(CLIP)捕捉成分信息，再与图像知识有效融合，结合扩散模型在稠密预测方面的优势，实现对食品图像的精准分割。结果：在基准数据集 FoodSeg103 上验证，所提模型的评估指标 mIoU 达到 47.93%，超越了当前最优的 FoodSAM 模型 1.51 个百分点。在基准数据集 UEC-FoodPIX Complete 上，模型的评估指标 mIoU 达到 75.13%，比 FoodSAM 模型高 8.99 个百分点。结论：所提出的多模态自蒸馏网络在食品图像分割任务中表现出色，验证了成分信息对分割任务的有效指导作用，提升了分割精度，为食品图像分析提供了新的解决方案。

关键词 食品图像；图像分割；多模态；自蒸馏
文章编号 1009-7848(2024)11-0010-12 **DOI**：10.16429/j.1009-7848.2024.11.002

食品是保障人类生存、生长发育和健康的重要物质基础。随着生活水平的提高，公众对健康管理和食品质量的要求越来越高，食品计算相关的研究也因此不断涌现[1]。许多与食品相关的应用，如食品的自动识别[2]、营养成分分析[3]等，都是在完整分割食品菜品目标的基础上完成的。作为食品计算的核心任务，食品图像分割是指利用计算机视觉技术，从 1 张图像中分离出食品区域的过程，这一过程不仅包括对食品项的准确识别，还涉及高分辨率下的像素级定位[4]。精准的食品分割对实现食品种类的细粒度识别、膳食健康评估以及开发健康相关的应用具有重要研究价值。然而，食品图像具有以下特点：1)同一食材在不同加工状态、不同烹饪方式下呈现的形态多样性(图 1a)：由于食品外观具有多样性并缺乏刚性结构，同一食品成分在不同加工状态或不同烹煮方式处理后呈现出形态多样性，特别是对于那些形状不规则、对比度较低、缺乏明显颜色或纹理特征的食品项，这种情况更为显著；2)食品间的前、后和上、下重叠(图 1b)：食品间因堆叠或摆放而导致的重叠；3)胡萝卜同类之间存在较大的差异性，桃子和芒果不同类之间存在相似性(图 1c)：较高的类内差异性和类间相似性。上述特性给食品图像的精准分割带来较大的挑战。

近年来，机器视觉与图像处理技术的快速发展，极大地推动了食品图像分割技术的进步[5]。例如，Long 等[6]首次通过引入全卷积神经网络(FCN)，实现了基于深度学习技术的食品图像语义分割。Chen 等[7]利用扩张卷积对 FCN 进行改进，在特征提取阶段扩大了特征的感受野。为了有效利用上、下文信息，Zhao 等[8]设计一种 PSPNet 网络，在提取特征过程中使用不同的池化层来聚合上、下文信息。此外，Wang 等[9]设计了一种非本地网络，对特征图中的每个像素对之间的关系进行建模。Huang 等[10]在上述非本地网络基础上采用交叉注意力，节约了网络层的计算成本。随着 Transformer 技术的发展[11-12]，Zheng 等[13]和 Wang 等[14]先后将 Vision Transformer 技术应用于食品图像的语义分割，更有效利用各区域的上、下文信息，丰富了全局信息，取得了较好的分割效果。随

收稿日期：2024-10-29
基金项目：国家重点研发计划项目(2023YFF1105104)；国家自然科学基金面上项目(62072289,62372278)；北京市自然科学基金项目(JQ24021)
第一作者：侯素娟，女，博士，教授
通信作者：闵巍庆 E-mail：minweiqing@ict.ac.cn

(a)　(b)　(c)

图 1　食品图像分割面临的挑战

Fig.1　Challenges in food image segmentation

着大模型在各个垂直领域的广泛应用，Wu 等[15]首次构建了一种食品图像预训练模型 ReLeM，通过将菜谱信息融合到视觉表征中，降低了类内差异。随着生成式技术的流行，Jaswanthi 等[16]设计一种用于食品图像分割的混合方法，利用 GAN 模型[17]，为食物图像生成建议掩码，然后基于 CNN 的识别模型进行区域分类。Lan 等[18]通过在现有的分割模型中嵌入 SAM[19]，能够利用提示工程生成高质量的掩模。然而，上述食品图像分割模型主要以图像作为单一的输入，忽略了与图像相关的周边信息（比如文本信息），这些可以作为辅助信息促进分割效果。

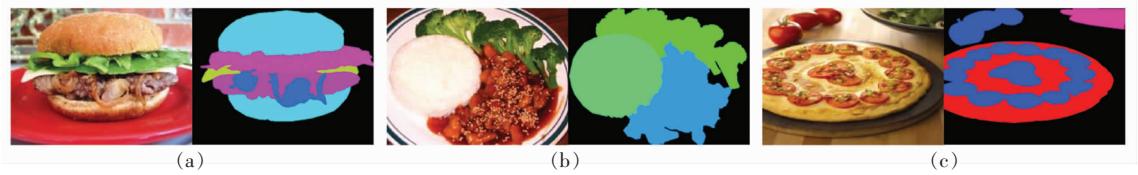
本文提出一种多源信息融合的自蒸馏网络。该网络基于扩散模型在稠密预测任务方面的优势[20]，使用 CLIP 预训练模型捕捉与图像对应的成分信息所蕴含的知识，并将其与图像知识有效融合。通过设计一种自蒸馏模型，探究成分信息在推理阶段对食品识别的指导作用，提高了食品图像分割的准确性。

1　材料与方法

1.1　材料

利用下述公开可用的食品图像分割数据集来评估所提分割模型的性能。

1.1.1　FoodSeg103[15]　该数据集来自食谱数据集 Recipe1M[21]。从 Recipe1M 数据集中统计食材类别频率排名，保留排名前 124 的类别，经进一步精炼后确定为 103 个食材类别，最后从 Recipe1M 中挑选图像，要求每张图像包含至少 2 种且不超过 16 种食材。这些食材应可见并易于标注。最终得到 7 118 张用于掩码标注的图像，部分例图如图 2 所示。



(a)　(b)　(c)

注：左侧图片为源图，右侧图片为分割图。

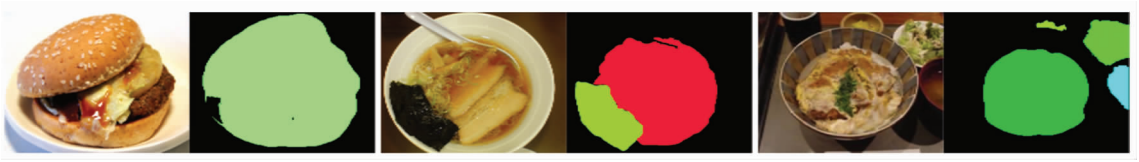
图 2　FoodSeg103 例图

Fig.2　FoodSeg103 examples

1.1.2　UEC-FoodPIX Complete[22]　该数据集包含 102 种菜品类别，共计 9 000 个训练图像和 1 000 个测试图像，通过对 UEC-FoodPix 数据集进行手工标注而来。分割掩码使用 GrabCut 半自动获取，由人工注释者根据一组预定义的规则进一步改进而来[23]。部分示意图如图 3 所示。

1.2　评估指标

假设 N 为类别的总数量，TP_i（True positive）表示被正确分类为类 i 的像素数量，FP_i（False positive）表示被错误分类为类 i 的像素数量（也称作假阳性），FN_i（False negative）是被错误分类为其它类别的像素数量（也称作假阴性），它们的真实



(a) (b) (c)

注:左侧图片为源图,右侧图片为分割图。

图 3 UEC-FoodPIX Complete 例图

Fig.3 UEC-FoodPIX Complet examples

标签是类 i。采用语义分割任务中评估模型性能的常用评估指标:平均交并比(Mean intersection over union, mIoU)和类别平均准确率(Mean accuracy, mAcc)来评估模型像素级分类的准确性。分别定义如下:

$$\mathrm{mIoU}(\%)=\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i+\mathrm{FN}_i}\times 100 \tag{1}$$

$$\mathrm{mAcc}(\%)=\frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FN}_i}\times 100 \tag{2}$$

1.3 设备与试验设置

试验设备使用 Linux 操作系统、英伟达显卡 Tesla V100 进行试验,并以用于密集视觉预测的扩散模型(DDP)[20]作为基准方法。

在模型训练过程中:对 FoodSeg103 数据集选取 4 983 张训练图像、29 530 张成分掩码图,2 135 张测试图像、12 567 张成分掩码,图像的大小裁剪为 512 像素×1 024 像素,参数 batchsize 设为 2。对 UEC-FoodPIX Complete 数据集选取 9 000 张训练图像、1 000 张测试图像,图像的大小剪裁为 320 像素×320 像素,参数 batchsize 设为 8。试验过程中使用 ConvNeXt-L[24]作为骨干网络进行特征的初步提取。

1.4 方法模型

为了探索文本模态对于分割任务的有效性,本文提出一种成分信息引导的多模态自蒸馏分割模型,模型整体架构如图 4 所示,主要包括 2 个部分:教师模型和学生模型。教师模型在训练阶段以图像和成分信息为输入,目标是实现文本和图像等多模态特征的对齐和信息融合,确保图像编码部分更全面地捕捉食品图像的语义信息。学生模型是以自蒸馏的方式在教师模型的指导下,利用 KL 散度量化自身模型与教师模型之间的信息损耗,接收来自教师模型中多模态信息的知识引导,实现模型的不断优化,并应用于测试阶段。

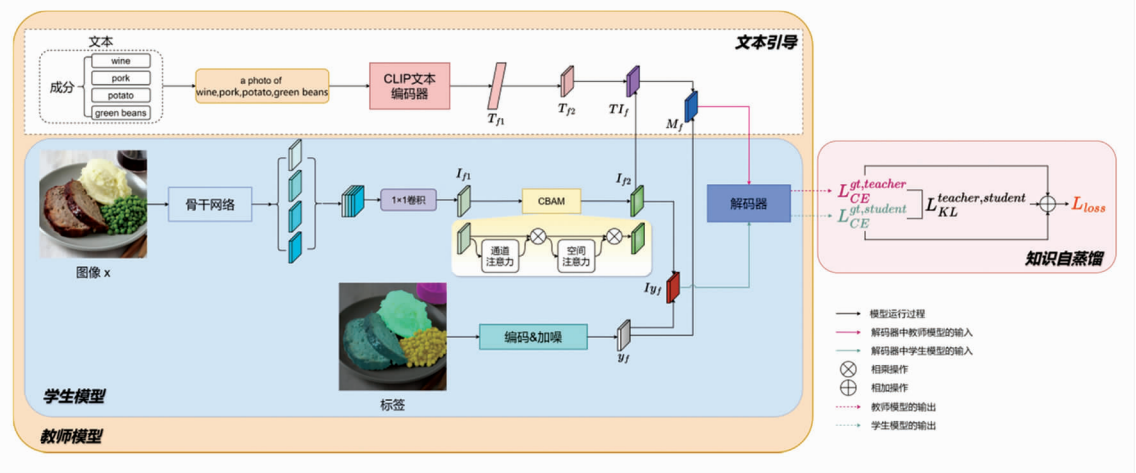


图 4 模型整体架构

Fig.4 The overall architecture of the model

1.4.1 教师模型 教师模型在训练时的输入是图像及对应的食品成分信息,处理过程如下:以图像 x 作为输入,经过骨干网络(Backbone)进行特征提取,生成 4 个不同分辨率的多尺度特征,随后通过

1×1 卷积操作进行聚合。聚合后的特征 I_{f1} 依次经过通道注意力和空间注意力的特征处理后，生成分辨率为 $256\times\frac{h}{4}\times\frac{w}{4}$ 的图像特征 I_{f2}。对于文本信息，以图像 x 对应的食品成分信息，如“wine”“pork”“potato”“green beans”等作为输入，首先形成“a photo of wine，pork，potato，green beans”的提示，后经 CLIP 文本编码器[25]处理，生成大小为 batchsize×768 的文本特征 T_{f1}。再将文本特征 T_{f1} 通过升维与图像特征 I_{f2} 进行特征对齐，并与图像特征进行融合，得到融合后的特征 TI_f。通过将高斯噪声加入到编码的真实值（Groundtruth，gt）中，构造带噪映射的加噪特征 y_f。再将融合后的特征 TI_f 与加噪特征 y_f 进行特征连接，生成最终的融合特征 M_f。进一步，将 M_f 作为解码器（堆叠 6 层可变形注意力模块）[26]输入，最后进行逐像素地分类，给出模型的预测输出结果 $\mathrm{Pre}_{\mathrm{teacher}}$。当 gt 为食品图像语义分割的真实值时，采用预测值与真实值的交叉熵损失更新教师模型，交叉熵损失定义如下：

$$L_{\mathrm{CE}}^{\mathrm{gt,teacher}}=L_{\mathrm{CE}}(\mathrm{gt},\mathrm{Pre}_{\mathrm{teacher}}) \tag{3}$$

在训练过程中，模型会通过将高斯噪声加入到编码的真实值中来，构造带噪映射 y_f。在推理过程中，噪声映射 y_f 从高斯分布中利用去噪扩散隐式模型（DDIM）[27]更新规则随机采样并迭代改进，以获得期望的预测结果。

1.4.2　学生模型　学生模型旨在通过自蒸馏技术，从教师模型中学习多模态知识，从而能够在单一图像模态输入时获得文本信息的有效指导，实现文本引导的像素级语义分割。

当以单模态图像为输入，由学生模型进行推理时，得到预测图 $\mathrm{Pre}_{\mathrm{student}}$。同样使用交叉熵（Cross entropy，CE）损失来指导学生模型，学生模型的损失如下：

$$L_{\mathrm{CE}}^{\mathrm{gt,student}}=L_{\mathrm{CE}}(\mathrm{gt},\mathrm{Pre}_{\mathrm{student}}) \tag{4}$$

1.4.3　基于自蒸馏的文本引导　借鉴知识蒸馏技术[28]，本文设计一种自蒸馏方式，实现教师模型的多模态信息对学生模型的指导，从而实现文本信息对学生模型进行图像分割时的有效引导。

在文本引导过程中，将食品图像及其对应的成分信息转化为文本提示，如“a photo of wine，pork，potato，green beans”，以捕捉图像的成分细节。这些提示经过 CLIP 文本编码器处理，生成维度为 batchsize×768 的文本特征向量。这些文本特征通过线性升维操作与图像特征对齐，以确保二者在同一特征空间中进行有效融合，融合后的特征不仅包含视觉信息，还融合了文本语义信息。

自蒸馏过程中采用 KL 散度（Kullback-leibler divergence）作为损失函数，以量化教师模型和学生模型之间的信息损耗，从而指导模型优化过程。使用 KL 散度损失可以最小化 2 个分支预测之间的差异。使用离散概率分布的 KL 散度实现模型的自蒸馏：

$$L_{\mathrm{KL}}^{\mathrm{teacher,student}}=\mathrm{Loss}_{\mathrm{KL}}(\mathrm{Pre}_{\mathrm{teacher}}\|\mathrm{Pre}_{\mathrm{student}}) \tag{5}$$

通过使用真实值标签作为监督，在 $L_{\mathrm{CE}}^{\mathrm{gt,teacher}}$ 和 $L_{\mathrm{CE}}^{\mathrm{gt,student}}$ 两者损失下监督教师模型分支和学生模型分支的关系预测，因此，模型的总损失为：

$$L_{\mathrm{loss}}=L_{\mathrm{CE}}^{\mathrm{gt,teacher}}+L_{\mathrm{CE}}^{\mathrm{gt,student}}+L_{\mathrm{KL}}^{\mathrm{teacher,student}} \tag{6}$$

2　结果与分析

本部分给出在 FoodSeg103 和 UECFoodPix Complete2 个数据集上的试验结果。首先，列出本文提出的模型在食物图像分割任务中与其它分割方法的性能对比结果；其次，给出消融研究的试验设置和结果；最后，展示所提模型在分割任务上的部分定性结果。

2.1　与其它分割方法的对比

在 FoodSeg103 数据集上进行试验，对照组包括 15 个分割方法，分别为 FPN[29]、ReLeM-FPN[15]、DeepLabV3+[30]、PEM[31]、CCNet[10]、ReLeM-CCNet[15]、Segformer[32]、Upernet[33]、KNet+UperNet[34]、STPPN[14]、SeTR-Naive[13]、Swin-Transformer[14]、ReLeM-SeTR-Naive[15]、SeTR-MLA[13]、FoodSAM[18]，具体的对比结果如表 1 所示。

同时在 UECFoodPix Complete 数据集上进行试验，对照组包括 12 个分割方法，分别为：FPN[29]、YOLACT[35]、ReLeM-FPN[15]、Upernet[33]、GourmetNet[36]、BayesianDeepLabv3+[37]、CCNet[10]、KNet+UperNet[34]、DeepLabV3 +[30]、FoodSAM [18]、Segformer [32]、Swin -

表 1 在 FoodSeg103 上与其它分割方法的比较

Table 1 Comparison with other segmentation methods on FoodSeg103

组别	方法	mIoU/%	mAcc/%
对照组 1	FPN[29]	27.67	38.36
对照组 2	ReLeM-FPN[15]	29.13	39.99
对照组 3	DeepLabV3+[30]	31.04	42.66
对照组 4	PEM[31]	33.98	45.73
对照组 5	CCNet[10]	35.57	45.45
对照组 6	ReLeM-CCNet[15]	36.35	47.08
对照组 7	Segformer[32]	38.67	48.96
对照组 8	Upernet[33]	39.80	52.37
对照组 9	KNet + UperNet[34]	40.18	52.01
对照组 10	STPPN[14]	40.30	53.98
对照组 11	SeTR-Naive[13]	41.30	52.70
对照组 12	Swin-Transformer[14]	41.60	-
对照组 13	ReLeM-SeTR-Naive[15]	43.90	57.00
对照组 14	SeTR-MLA[13]	45.10	57.44
对照组 15	FoodSAM[18]	46.40	58.27
试验组 1(基准模型)	DDP[20]	47.82	60.49
试验组 2(本文方法)	**DDP-FPN+CLIP+CBAM(DCC)**	**48.46**	**60.22**
试验组 3(本文方法)	**DDP-FPN+CLIP+CBAM+Self-distillation(DCCD)**	**47.93**	**58.65**

注:"-"表示原论文中缺少相关数据。

表 2 在 UECFoodPix Complete 数据集上与其它分割方法的比较

Table 2 Comparison with other segmentation methods on UECFoodPix Complete

组别	方法	mIoU/%	mAcc/%
对照组 1	FPN[29]	53.34	67.21
对照组 2	YOLACT[35]	54.85	-
对照组 3	ReLeM-FPN[15]	57.34	71.36
对照组 4	Upernet[33]	59.35	74.44
对照组 5	GourmetNet[36]	62.88	75.87
对照组 6	BayesianDeepLabv3+[37]	64.21	76.15
对照组 7	CCNet[10]	64.62	77.50
对照组 8	KNet + UperNet[34]	64.88	76.94
对照组 9	DeepLabV3+[29]	65.61	77.56
对照组 10	FoodSAM[18]	66.14	78.01
对照组 11	Segformer[32]	67.5	78.97
对照组 12	Swin-Transformer[14]	67.72	79.13
试验组 1(基准模型)	DDP[20]	74.64	84.59
试验组 2(本文方法)	**DDP-FPN+CLIP+CBAM(DCC)**	**76.62**	**85.63**
试验组 3(本文方法)	**DDP-FPN+CLIP+CBAM+Self-distillation(DCCD)**	**75.13**	**84.65**

注:"-"表示原论文中缺少相关数据。

Transformer[14],具体的对比结果如表 2 所示。

从表 1 对比结果可看出,本文的模型在 FoodSeg103 数据集上的 mIoU 指标达到了目前最好(SOTA)性能,比当前公开的最好的 FoodSAM 模型高 1.51 个百分点, 比基准模型 DDP 高 0.11 个百分点。类似地,从表 2 对比结果所示,本文的模型在 UECFoodPix Complete 数据集上 2 个评估指标均达到 SOTA 性能,其中 mIoU 达到 75.13%,比当前公开的最好的 FoodSAM 模型高 8.99 个百分点,比基准模型 DDP 高 0.49 个百分点。

2.2 消融试验

为验证模型中各关键模块在分割过程中的作用,在 FoodSeg103 基准数据集上对这些模块进行消融试验的验证,具体如下:

2.2.1 DDP-FPN 在基准模型 DDP 的基础上去掉 FPN 模块。通过试验发现,DDP-FPN 模型在 FoodSeg103 数据集上比 DDP 的性能高 0.34 个百分点,在 UECFoodPix Complete 数据集上高 0.87 个百分点。这可能是由于 FPN 模块通常用于融合来自不同层次的特征图,而由于食品数据集的特点,以及骨干网络 ConvNeXt-L 在特征提取能力上已足够强大,复杂的融合策略可能存在特征之间的冗余,反而不能带来性能提升,因此后续试验都是在 DDP-FPN 模型基础上进行验证。

2.2.2 CLIP 文本处理模块 CLIP 是一种文本-图像对的预训练多模态模型,该模型由 2 部分组成:文本编码器和图像编码器。文本编码器中使用屏蔽自注意力,以保留使用预先训练的语言模型进行初始化或添加语言建模作为辅助目标的能力。CLIP 的文本编码器通过使用模板“A photo of {object}”解决了仅用单个单词做提示词引起的单词歧义性和训练与推理数据不一致性的问题。该机制对于食品图像分割数据集中为每个食品提供单词类标签的语义标签非常友好,允许食品图像中以单词型的成分文本信息以句子的形式作为输入,使多模态增强食品图像分割的方法成为可能。如表 3 所示,通过比较使用 CLIP 前、后在语义分割中的有效性,发现使用 CLIP 后模型取得更好的性能。证明 CLIP 文本编码器通过提取食品图像中成分的文本信息,有效辅助食品图像编码部分,更全面地捕捉食品图像的的语义信息,提升模型更强的信息对齐和融合能力。

2.2.2.1 CBAM(Convolutional block attention module)模块 该模块结合了通道注意力机制(Channel attention mechanism)和空间注意力机制(Spatial attention mechanism),用于增强卷积神经网络在特征提取过程中的表征能力。如表 3 所示,本文研究了模型加入 CBAM 模块对性能的影响,发现 CBAM 模块对于模型在多数指标上发挥正向作用。这可能是因为对于食品图像编码部分提取的图像特征,通过 CBAM 模块来提高模型对通道特征和空间位置的关注度,能够帮助神经网络更好地学习和利用特征信息,从而达到提升模型性能的效果。

据表 3 结果可知,本文提出的方法在 FoodSeg103 和 UECFoodPix Complete 数据集上 mIoU 指标均有不同程度的性能提升,反映了本文方法有助于提高对图像的整体分割质量。然而,mAcc 指标在 UECFoodPix Complete 数据集上提高了 1.04 个百分点,在 FoodSeg103 上相较于基准方法没有提升,这可能是因为 mAcc 是统计所有分类的平均准确率,由于 FoodSeg103 中的数据存在类别重叠的情况较多,特定类别的分类准确性比较低,因此影响整体的分类准确率。

表 3 关键模块的消融试验

Table 3 Ablation tests of key modules

关键模块			FoodSeg103 数据集		UECFoodPix Complete 数据集	
DDP-FPN	CLIP	CBAM	mIoU/%	mAcc/%	mIoU/%	mAcc/%
			47.82	60.49	74.64	84.59
√			48.16	60.29	75.51	83.89
√	√		48.33	60.34	76.07	85.33
√		√	48.31	59.96	75.87	86.20
√	√	√	48.46	60.22	76.62	85.63

注:“√”表示试验中加入该模块。

2.2.2.2 多模态融合方式 在引入食品成分这一文本信息后,试验验证了如何将文本特征和图像特征进行有效融合,共设计 2 种融合方式,即:concat 和 add。前者直接进行特征连接,后者在通道维度上进行特征相加,试验结果如表 4 所示。对于文本特征和图像特征的多模态融合方式,发现特征相加比特征连接的效果更好。这是因为在通道维度上进行特征相加,相当于对特征图的每个

通道进行加权融合，这种融合方式更加有效地结合了文本特征和图像特征的信息。直接进行特征连接只是简单地将文本特征和图像特征拼接在一起，可能存在冗余或不必要的信息。

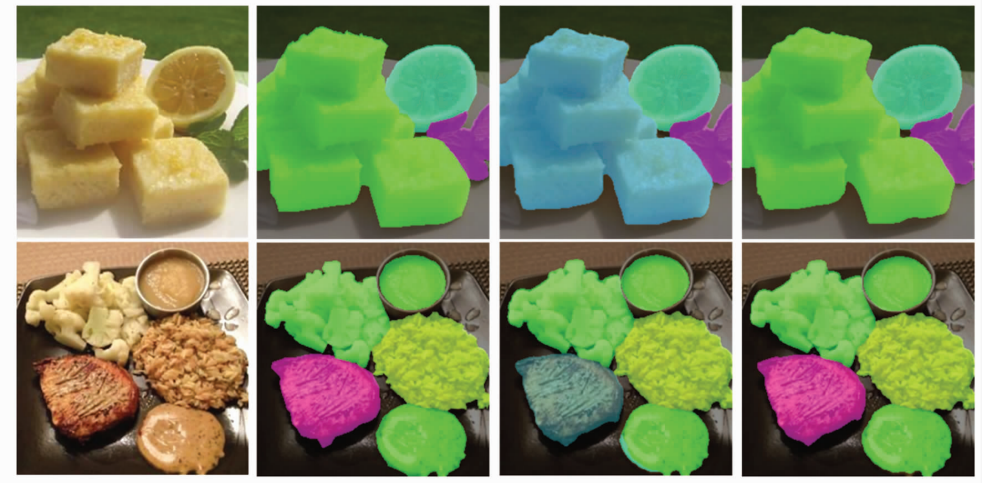
表 4 多模态融合方式的消融研究

Table 4 Ablation tests of multimodal fusion modalities

多模态融合方式	FoodSeg103 数据集		UECFoodPix Complete 数据集	
	mIoU/%	mAcc/%	mIoU/%	mAcc/%
concat	48.0	59.40	75.74	85.95
add	48.33	60.34	76.07	85.33

2.3 定性分析

图 5 展示模型中 CLIP 文本处理模块的前、后可视化差异，分别对比了 DDP-FPN 方法和 DDP-FPN + CLIP 方法的试验结果，发现加入 CLIP 后的模型比 DDP-FPN 方法在食品成分语义信息识别上实现了更好的性能。这说明引入 CLIP 文本处理模块，对食品成分的文本信息进行有效的特征提取，并引导模型在食品图像上，实现了语义分割的性能提升。

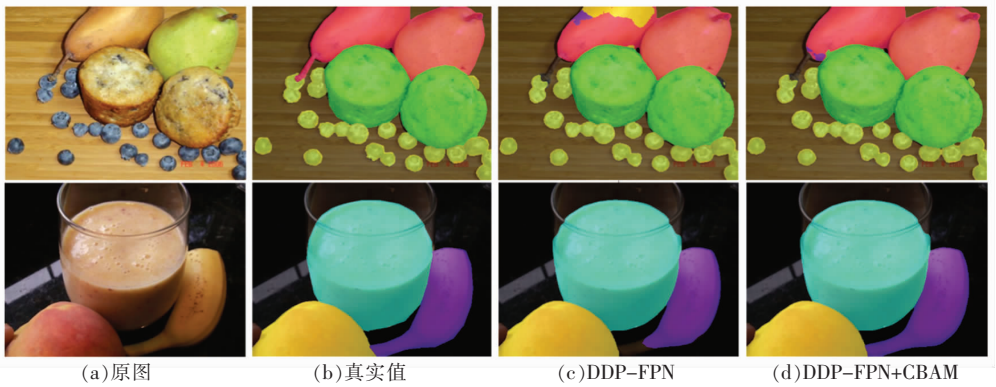


(a)原图 (b)真实值 (c)DDP-FPN (d)DDP-FPN+CLIP

图 5 模型加入 CLIP 前、后的可视化对比

Fig.5 Visual comparison of the model before and after adding CLIP

图 6 展示 DDP-FPN 模型在加入 CBAM 注意力模块前、后的可视化对比。通过对比发现，加入 CBAM 模块后模型在分割精度上更准确，说明 CBAM 模块通过其通道注意力机制和空间注意力机制，增强了卷积神经网络在食品图像特征提取过程中的表征能力。

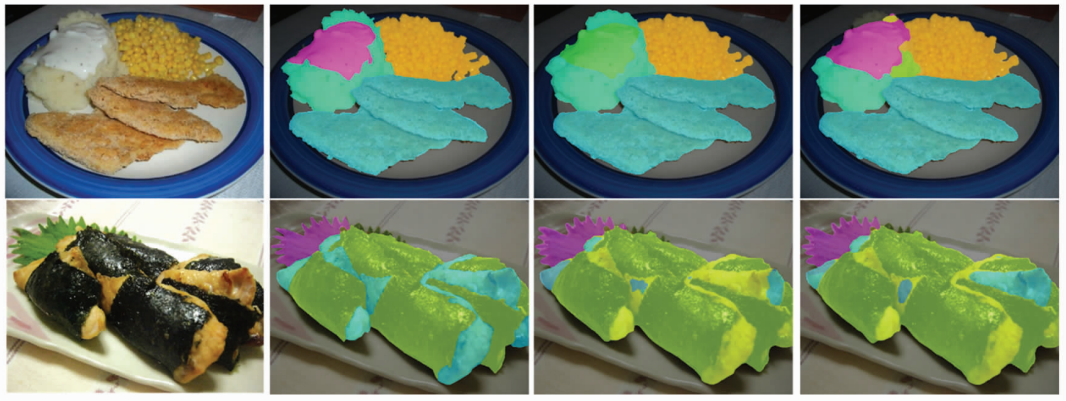


(a)原图 (b)真实值 (c)DDP-FPN (d)DDP-FPN+CBAM

图 6 模型加入 CBAM 前、后的可视化对比

Fig.6 Visual comparison of the model before and after adding CBAM

图 7 展示文本特征和图像特征不同的融合方式的试验效果,对比了 2 种多模态融合方式。结果发现,add 融合方式下在多数情况可以正确识别出食品图像中包含的成分类别，相较于 concat 融合方式有较大提升。图 7b 示例中,虽然 2 种融合方式均正确识别出食品图像中所包含的成分类别，但是 add 的融合方式相较于 concat 的融合方式在分割精度上更准确，这说明在 add 多模态融合方式上,模型可以更加有效地借助 CLIP 处理文本特征，实现文本特征信息对食品图像语义分割更显著的指导。



(a)原图　(b)真实值　(c)concat　(d)add

图 7　不同多模态融合方式的可视化结果

Fig.7　Visual results of different multimodal fusion methods

同时，分别在 FoodSeg103 数据集和 UEC-FoodPix Complete 数据集上可视化基准方法 DDP 与本文模型 DCC 和 DCCD 的比较,对比结果如图 8、图 9 所示。

通过图 8a~8e 的对比结果可知，本文的模型 DCC 可以正确识别出成分类别,即使蒸馏后,DCCD 模型相较于原始模型 DDP,无论是在正确识别成分信息还是在语义分割精度上都有一定提升。图 8d~8e 的对比结果显示,当 DDP 和本文的模型 DCC、DCCD 都可正确识别成分的类别时，本文的模型 DCC、DCCD 在分割精度上性能更优。

图 9a~9e 的对比结果显示，在文本信息的指

(a)输入 (b)真实值 (c)DDP(基准模型) (d)DCC (e)DCCD

图 8 FoodSeg103 数据集可视化结果

Fig.8 FoodSeg103 data set visualization results

(a)输入 (b)真实值 (c)DDP(基准模型) (d)DCC (e)DCCD

图 9 UECFoodPix Complete 数据集可视化结果

Fig.9 UECFoodPix Complete data set visualization results

导下,本文模型 DCC、DCCD 能够正确识别菜品类别，提高了模型在进行语义分割时识别语义信息的性能，且对于一张图片中含有单菜品图像或多菜品图像均有效。类似地,图 9d~9e 的对比结果显示，当 DDP 和本文模型 DCC、DCCD 都正确地识别菜品的类别时，本文模型的分割性能有明显提高,分割结果更精细。

以上可视化结果显示，加入成分的文本信息作为指导，能够为模型提供一种较为强大的语义信息识别能力，从而提高其在语义分割方面的性能。

3　结论

本工作以食品图像为研究对象，设计了一种成分信息引导的多模态自蒸馏网络，通过结合 CLIP 预训练模型捕捉的成分信息知识,整合图像特征,利用扩散模型的稠密预测优势,实现了多模态网络对食品图像的像素级语义分割。通过在 2 个公开可用的食品图像基准数据集中进行综合评估，发现本文模型在实际应用中显示出良好的应用性能,即对食品图像的语义分割,在性能上超越了现有的方法。消融试验结果验证了引入文本辅助信息和自蒸馏机制在食品图像分割中的有效性和可行性。本工作验证了成分信息在引导食物分割中的潜力。然而，模型存在的一个潜在局限性是:由于当前用于食品图像分割的数据集较少,因此模型对不同食材的表现可能受到数据集规模和样本多样性的限制。尤其是对于一些罕见或极端情况的食材,模型的泛化能力需要进一步提高。

参　考　文　献

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

[2] MIN W Q, WANG Z L, LIU Y X, et al. Large scale visual food recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9932-9949.

[3] SHAO W J, MIN W Q, HOU S J, et al. Vision-based food nutrition estimation via RGB-D fusion network[J]. Food Chemistry, 2023, 424: 136309.

[4] WU X W, YU S C, LIM E P, et al. OVFoodSeg: Elevating open-vocabulary food image segmentation via image-informed textual representation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 4144-4153.

[5] WANG W, MIN W Q, LI T H, et al. A review on vision-based analysis for automatic dietary assessment[J]. Trends in Food Science & Technology, 2022, 122: 223-237.

[6] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440.

[7] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs [C]// International Conference on Learning Representations. San Diego: ICLR, 2015: 1-14.

[8] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6230-6239.

[9] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7794-7803.

[10] HUANG Z L, WANG X G, HUANG L C, et al. Ccnet: Criss-cross attention for semantic segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 603-612.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000-6010.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations. [S.l.]: ICLR, 2021: 1-22.

[13] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-

sequence perspective with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 6877-6886.

[14] WANG Q K, DONG X X, WANG R M, et al. Swin transformer based pyramid pooling network for food segmentation[C]// IEEE 2nd International Conference on Software Engineering and Artificial Intelligence. Xiamen: IEEE, 2022: 64-68.

[15] WU X W, FU X, LIU Y, et al. A large-scale benchmark for food image segmentation[C]// Proceedings of the 29th ACM international Conference on Multimedia. [S.l.]: ACM, 2021: 506-515.

[16] JASWANTHI R, AMRUTHATULASI E, BHAVYASREE C, et al. A hybrid network based on GAN and CNN for food segmentation and calorie estimation[C]// International Conference on Sustainable Computing and Data Communication Systems. Erode: IEEE, 2022: 436-441.

[17] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5967-5976.

[18] LAN X, LYU J Y, JIANG H Y, et al. Foodsam: Any food segmentation[J/OL]. IEEE Transactions on Multimedia, (2023-11-03)[2024-10-29]. https://ieeexplore.ieee.org/document/10306316.

[19] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 4015-4026.

[20] JI Y F, CHEN Z, XIE E Z, et al. Ddp: Diffusion model for dense visual prediction[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 21684-21695.

[21] SALVADOR A, HYNES N, AYTAR Y, et al. Learning cross-modal embeddings for cooking recipes and food images[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3068-3076.

[22] OKAMOTO K, YANAI K. UEC-FoodPIX Complete: A large-scale food image segmentation dataset[C]// Pattern Recognition. ICPR International Workshops and Challenges. [S.l.]: ACM, 2021: 647-659.

[23] EGE T, SHIMODA W, YANAI K. A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice [C]// Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management. Nice: ACM, 2019: 82-87.

[24] LIU Z, MAO H Z, WU C Z, et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11966-11976.

[25] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the 38th International Conference on Machine Learning. [S.l.]: IMLS, 2021: 8748-8763.

[26] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal im age segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1280-1289.

[27] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[C]// International Conference on Learning Representations. [S.l.]: ICLR, 2021: 1-22.

[28] ZHANG L, SU J S, MIN Z J, et al. Exploring self-distillation based relational reasoning training for document-level relation extraction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: ACM, 2023: 13967-13975.

[29] KIRILLOV A, GIRSHICK R, HE K M, et al. Panoptic feature pyramid networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 6392-6401.

[30] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]// Proceedings of the European Conference on Computer Vision. Munich: ACM, 2018: 833-851.

[31] CAVAGNERO N, ROSI G, CUTTANO C, et al. Pem: Prototype-based efficient maskformer for image segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 15804-15813.

[32] XIE E Z, WANG W H, YU Z D, et al. Seg-

Former: Simple and efficient design for semantic segmentation with transformers [C]// Proceedings of the 35th International Conference on Neural Information Processing System. [S.l.]: ACM, 2021: 12077-12090.

[33] XIAO T T, LIU Y C, ZHOU B L, et al. Unified perceptual parsing for scene understanding[C]// Proceedings of the European Conference on Computer Vision. Munich: ACM, 2018: 432-448.

[34] ZHANG W W, PANG J M, CHEN K, et al. K-net: Towards unified image segmentation[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2021: 10326-10338.

[35] BATTINI SONMEZ, E, MEMIS S, ARSLAN B, et al. The segmented UEC Food-100 dataset with benchmark experiment on food detection[J]. Multimedia Systems, 2023, 29(4): 2049-2057.

[36] SHARMA U, ARTACHO B, SAVAKIS A. Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention[J]. Sensors, 2021, 21(22): 7504.

[37] AGUILAR E, NAGARAJAN B, REMESEIRO B, et al. Bayesian deep learning for semantic segmentation of food images[J]. Computers and Electrical Engineering, 2022, 103: 108380.

Ingredient-guided Multimodal Self-distillation for Food Image Segmentation

Hou Sujuan[1], Sun Yuejuan[1], Min Weiqing[2*], Wang Ruiping[2], Jiang Shuqiang[2]

([1]*Shandong Normal University, Jinan 250358*

[2]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190*)

Abstract Objectives: With advancements in computer vision technology, accurately identifying and segmenting various components in food images has become essential for food nutrition analysis and promoting healthier diet management. However, most existing image segmentation models rely solely on a single image input, which often struggles to capture subtle distinguishing features in food images with minimal visual differences, ultimately impacting segmentation accuracy. This paper addressed the limitations of single-modality approaches in segmentation tasks by incorporating text information to provide richer contextual data for the model. Additionally, it leveraged self-distillation techniques to guide the model in effectively segmenting food images. Methods: This paper proposed a multi-modal self-distillation segmentation model guided by ingredient information to improve food image segmentation. The model leveraged the comparative languaged pre-training model (CLIP) to capture ingredient information and fused it with image knowledge. By combining the strengths of the diffusion model in dense prediction, the model achieved accurate segmentation of food images. Results: When evaluated on the benchmark dataset FoodSeg103, the model achieved an mIoU of 47.93%, surpassing the current best-performing FoodSAM model by 1.51%. On the UEC-FoodPIX Complete benchmark dataset, the mIoU reached 75.13%, outperforming the FoodSAM model by 8.99%. Conclusions: The proposed multi-modal self-distillation network demonstrated strong performance in food image segmentation, showcasing the effective role of ingredient information in guiding segmentation tasks. This approach significantly improves segmentation accuracy and presents a promising solution for food image analysis.

Keywords food image; image segmentation; multimodal; self-distillation