食品工业科技
Science and Technology of Food Industry
ISSN 1002-0306,CN 11-1759/TS

《食品工业科技》网络首发论文

题目：基于 Transformer 的零样本食品图像检测
作者：宋静茹，闵巍庆，周鹏飞，饶全瑞，盛国瑞，杨延村，王丽丽，蒋树强
DOI：10.13386/j.issn1002-0306.2024030027
网络首发日期：2024-05-31
引用格式：宋静茹，闵巍庆，周鹏飞，饶全瑞，盛国瑞，杨延村，王丽丽，蒋树强．基于 Transformer 的零样本食品图像检测[J/OL]．食品工业科技.
https://doi.org/10.13386/j.issn1002-0306.2024030027

网络首发时间：2024-05-31 09:33:07
网络首发地址：https://link.cnki.net/urlid/11.1759.TS.20240530.0958.009

作者简介：宋静茹（1999-），女，硕士研究生，研究方向：食品计算和计算机视觉，E-mail：songjingru@m.ldu.edu.cn。
*通讯作者：王丽丽（1978-），女，博士，教授，研究方向：宽带通信和多媒体通信，E-mail：wanglili@ldu.edu.cn。
基金项目：国家自然科学基金项目（61705098；61872170）；山东省自然科学基金项目（ZR2023MF031）；

基于 Transformer 的零样本食品图像检测

宋静茹[1]，闵巍庆[2,3]，周鹏飞[2,3]，饶全瑞[1]，盛国瑞[1]，杨延村[1]，王丽丽[1,*]，蒋树强[2,3]

（1.鲁东大学信息与电气工程学院，山东烟台　264025；
2.中国科学院计算技术研究所，北京　100190；
3.智能信息处理重点实验室，北京　100190）

摘　要：食品检测作为食品计算的一项基本任务，能够对输入的食品图像进行定位和识别，在智慧食堂结算和饮食健康管理等食品应用领域发挥着至关重要的作用。然而在实际场景下，食品类别会不断更新，基于固定类别训练的食品检测器很难对未见过的食品类别进行精准的检测。为了解决这一问题，本文提出了一种零样本食品图像检测方法。首先，构建了一个基于 Transformer 的食品基元生成器，其中每个基元都包含与食品类别相关的细粒度属性，根据食品的特性，可以有选择地组装这些基元，以合成未见类特征。其次，为了给未见类的视觉特征更多约束，本文提出了一个视觉特征解纠缠的增强组件，将食品图像的视觉特征分解为语义相关特征和语义不相关特征，以此能更好地将食品类别的语义知识转移到其视觉特征。所提出的方法在 ZSFooD 和 UEC-FOOD256 两个食品数据集上进行了大量实验和消融研究，在零样本检测（Zero-Shot Detection，ZSD）设置下，未见类别取得了最优的平均精度，分别达到了 4.9%和 24.1%，在广义零样本检测（Generalized Zero-Shot Detection，GZSD）的设置下，可见类和未见类的调和平均值（HM，Harmonic Mean）分别达到了 5.8%和 22%，证明了所提出方法的有效性。
关键词：食品图像检测；零样本学习；生成式模型；Transformer；深度学习

Zero-Shot Food Image Detection Based on Transformer

SONG Jing-ru[1], MIN Wei-qing[2,3], ZHOU Peng-fei[2,3], RAO Quan-rui[1], SHENG Guo-rui[1], YANG Yan-cun[1], WANG Li-li[1,*], JIANG Shu-qiang[2,3]

(1. School of Information and Electrical Engineering, Ludong University, Yantai 264025, China;
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
3.Key Lab of Intelligent Information Processing, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: As a fundamental task in food computing, food detection plays a crucial role in locating and identifying food items from input images, particularly in applications such as intelligent canteen settlement and dietary health management. However, food categories are constantly updating in practical scenarios, making it difficult for food detectors trained on fixed categories to accurately detect previously unseen food categories. To address this issue, this paper proposes a zero-shot food image detection method. Firstly, a Transformer-based food primitive

generator is constructed, where each primitive contains fine-grained attributes relevant to food categories. These primitives can be selectively assembled based on the food characteristics to synthesize new food features. Secondly, an enhancement component of visual feature disentanglement is proposed in order to impose more constraints on the visual features of unseen food categories. The visual features of food images are decomposed into semantically related features and semantically unrelated features, thereby better transferring semantic knowledge of food categories to their visual features. The proposed method is extensively evaluated on the ZSFooD and UEC-FOOD256 datasets through numerous experiments and ablation studies. Under the zero-shot detection (ZSD) setting, optimal average precision on unseen classes reaches 4.9% and 24.1%, respectively, demonstrating the effectiveness of the proposed approach. Under the Generalized Zero-Shot Detection (GZSD) setting, the harmonic mean of visible and unseen classes reaches 5.8% and 22%, respectively, further validating the effectiveness of the proposed method.

Key words: food image detection; zero-shot learning; generative models; Transformer; deep learning

中图分类号：S126　　文献标志码：A

DOI：10.13386/j.issn1002-0306.2024030027

在当今互联网、物联网和社交网络快速发展的环境下，大量的多媒体食品数据如洪水般涌现，这些数据蕴含着巨大的应用潜力和社会价值[1]。食品图像检测[2]作为食品计算的一项基本任务，在食品推荐、膳食评估等方面被广泛应用[3-5]，其中，深度学习的方法在提高食品图像检测性能上发挥着巨大的作用[6-7]。在食品图像检测技术的发展过程中，常见的做法是利用通用目标检测框架，如图 1 所示。例如，Sun 等[8]提出了一个基于 YOLOv2 的移动应用程序，用于实时检测食品图像；Guilar 等[9]提出了一种在食堂和饭店环境下检测和预测托盘中菜品类别的方法。这些方法利用深度学习有效地解决了多个识别对象的问题，使得系统能够更全面地理解和解释食品图像，提高了食品图像检测的综合性能。然而，在现实场景中，新的食品类别不断涌现，其中一些可能是跨文化的合成食品或新型饮食趋势所带来的。面对这样的不可控因素，仅能识别有限固定类别的食品图像检测器在处理新类别时往往表现出不佳的结果。此外，由于深度学习方法是数据驱动的，其依赖于大量标记的训练样本，这也意味着需要大量的人力和时间的投入[10]。

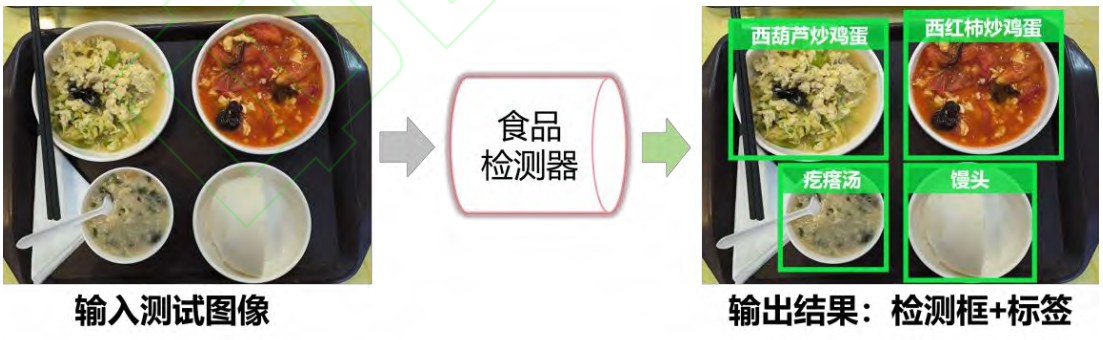
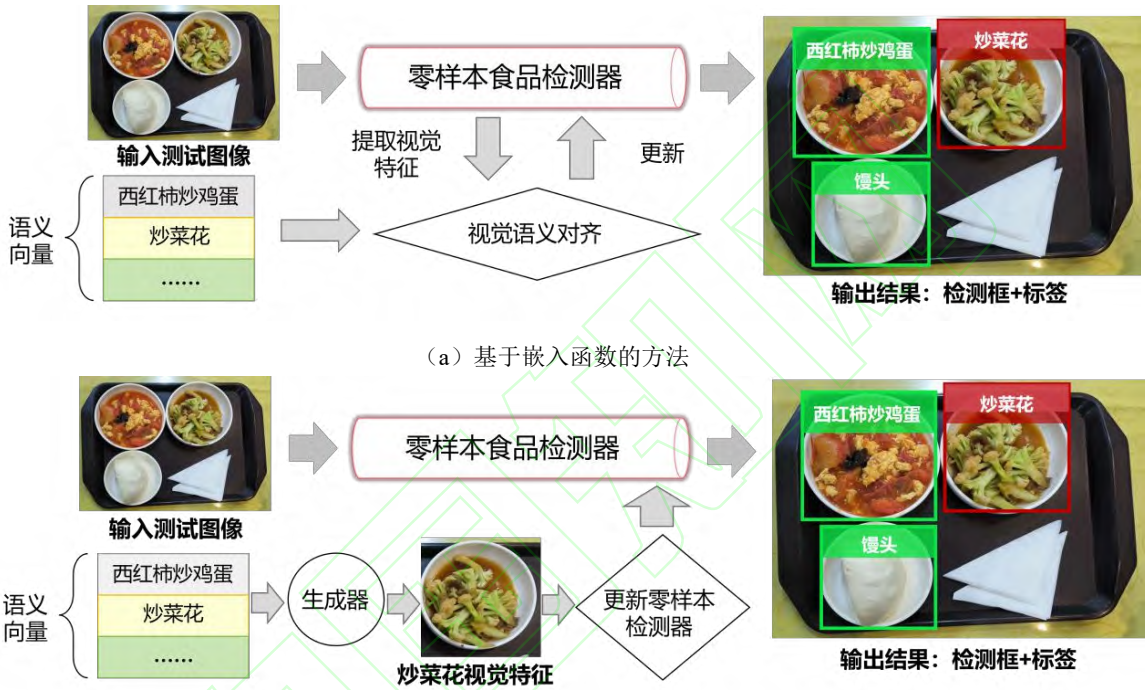


图 1　食品图像检测任务框架图

Fig. 1　Framework diagram of food image detection task

为了解决这些问题，零样本检测（Zero-Shot Detection，ZSD）方法[11]应运而生，现有的 ZSD 方法利用从大规模语料库[12-13][11]中的无监督学习中获得的语义向量作为辅助信息，这样获得的语义向量之间会因其各类属性的不同存在着不同的远近关系。根据语义向量的使用方式，零样本检测方法可以分为两组，即基于嵌入函数的方法[14-16]和基于生成模型[17-18]的方法，如图 2 所示。前者主要是学习一种视觉空间与语义空间的映射函数，将视觉特征映射到语义空间中，然后在语义空间中执行最近邻搜索来预测未见的对象类别。ConSE[14]提出了一种利用类别标签嵌入向量的凸组合将图像映射到语义

嵌入空间的方法；BLC[15]提出了一种利用背景感知表示来提高零样本检测性能的方法，并引入了一个新的评价指标——广义零样本检测（Generalized Zero-Shot Detection，GZSD），旨在评估模型在同时检测可见和未见类别的能力；CZSD[16]结合了两个语义引导的对比学习子网，以优化视觉数据结构。然而，由于映射函数完全是基于训练数据提供的可见类别来学习的，因此当在测试中处理视觉特征时，模型将显著偏向于可见类别。后者基于生成模型的方法通常利用生成模型，例如对抗生成网络（Generative Adversarial Network, GAN）[19]和变分自编码器（Variational Autoencoder, VAE）[20]，使用对应类别的语义嵌入合成视觉特征，然后合成的视觉特征就可以用于训练未见类的检测器。SU[17]通过具有多样性调节的 GAN 模型合成看不见的特征，并应用到零样本检测任务中；RRFS[18]设计了一个鲁棒的区域特征合成器，用于生成鲁棒的多样性特征。



（a）基于嵌入函数的方法

（b）基于生成模型的方法

图 2　零样本食品图像检测两种主流方法对比

Fig. 2　Comparison of two mainstream methods for zero-shot food image detection

虽然上述的生成式方法在减轻偏见问题方面取得了一定成效，但这类生成器忽略了语义信息对图像特征的指导作用，在这种情况下容易导致生成低质量的合成特征[21-22]。此外，食品图像检测任务在其特性上与其他类型的目标检测任务显著不同，因此在处理食品图像检测时，面临着一系列独特的挑战[23-25]。这些挑战主要体现在两个方面，一方面是食品的细粒度特征，即同种食品之间在视觉上难以区分，不同食品之间具有细微的区别和相似性，如图 3 所示，属于同一种食品类别的炸鸡排，但在视觉上存在多种样式，炸藕盒和炸茄盒、水煮鱼和水煮肉片等食品之间具有很相似的视觉特征。另一方面是食品的复杂属性，如成分、烹饪方式、口味等，如图 4 所示，口水鸡和口水鸭由众多复杂属性组成，它们在共享相似的视觉模式下，由于这些复杂特性，会导致仅使用类别嵌入来区分它们变为异常困难。

为了应对上述挑战，本文以 Transformer[26]框架为基础，利用丰富的基元信息来构建食品的视觉特征表示，并将生成的视觉特征分为与食品语义嵌入向量相关的和不相关的部分，以增强生成器表达的多样性，特别是在食品丰富的细粒度属性方面，使得不同食品类别的特征合成更加可靠。

图 3　食品的细粒度特征

Fig. 3　Fine-grained features of food

图 4　食品的复杂属性

Fig. 4　Complex attributes of food

1　材料与方法

1.1　材料

本文选择 UEC-FOOD256[27]和 ZSFooD[28]作为数据源。如表 1 所示，UEC-FOOD256 包含 20452 张训练图像和 5735 张测试图像，共 256 个类别，ZSFooD 包含 10463 张训练图像和 10140 张测试图像，共 228 个类别。由于零样本检测任务的特殊性，在原有数据集基础上，还需要对数据的食品类别进行划分。对于 UEC-FOOD256，将食品类别划分为 205 个可见类和 51 个未见类，在 ZSFooD 中，将类别划分为 184 个可见类和 44 个未见类，图 5 展示了所使用数据的一些实例图像。

表 1　ZSFooD 和 UEC-FOOD256 数据集的统计数据

Table 1　The statistical data for the ZSFooD and UEC-FOOD256 datasets

数据集	类别			图像		
	可见类	未见类	总计	训练	测试	总计
ZSFooD	184	44	228	10463	10140	20603
UEC-FOOD256	205	51	256	20452	5732	26184

a.ZSFooD　　b.UEC-FOOD256

图 5　ZSFooD 和 UEC-FOOD256 的一些实例图像

Fig. 5　Some sampled images from ZSFooD and UEC-FOOD256

拆分创建训练集和测试集的过程：为了将类别划分为可见（训练）类和未可见（测试）类，本文首先使用语义向量之间的余弦相似度作为度量将它们聚类成 K 个簇[11]，然后每个簇中随机选择 80%的类，并将它们分配给可见类集，将每个集群中剩余的 20%的类分配给测试集。

构建语义向量：在训练阶段获取可见类和未见类的语义向量，其作用是建立可见类和未见类之间的联系，在语义向量空间中，视觉上相似的类别位于靠近的位置。任何类别的语义向量都可以手动或自动生成。手动生成的语义向量由于需要人工注释成本较大，所以自动生成的语义向量被视为是一种更优方案。本文中提到的 ZSFooD 和 UEC-FOOD256 数据集，选择使用 CLIP[29]大规模预训练的语言模型来获取食品类别的语义向量，当以无监督方式生成时，这样的嵌入可能会包含一些噪声，但相较于手动创建的向量，它提供了更大的灵活性和可扩展性。

a.学习可见类（绿色）　　b.检测不可见类（红色）

图 6　零样本食品图像检测任务示意图

Fig. 6　Zero-shot food image detection task diagram

1.2 模型工作原理

本文的研究旨在通过使用训练集中有标注的食品图像学习零样本检测器，进而可以检测出训练集中未曾出现的食品类别。首先对零样本食品图像检测任务进行定义，如图 6 所示：零样本食品图像检测任务的目的是在训练集T_S上学习一个具有零样本检测能力的食品图像检测器，从而能够精准检测出测试集T_U中未见过的食品对象。在零样本食品图像检测中，有两个不相交的食品类别集合，即可见类别Y_S和未见类别Y_U，其中$Ys \cap Yu = \emptyset$。训练集只包含

可见类的食品对象，每个食品图像中的食品对象都具有相应的类标签和边界框坐标。而测试集可以仅包含未见的食品对象，也可以包含可见和未见的食品对象。在训练和测试期间，为食品对象的类别定义其语义向量集为$W = \{W_S, W_U\}$，其中W_S和W_U分别是可见类和未见类的语义向量集。

零样本食品图像检测模型的具体过程如图 7 所示，主要分为两部分：目标检测网络和特征生成模块。在本文中，采用 Faster-RCNN 模型[30]作为目标检测模块，以 ResNet-101[31]为骨干网络。首先，使用可见类的食品图像以及其对应的真实标注框来训练 Faster-RCNN 模型，一旦模型训练完成，便可以使用它来提取可见类食品图像的区域特征A_r。其次，训练特征生成模块来学习食品的语义向量和视觉特征之间的映射，其训练目的是使用学习的生成器来为未见的食品类别生成视觉特征A_g。有了这些合成的看不见的区域特征及其相应的类标签，就可以为未见的食品类别训练未见的分类器。最后，更新 Faster-RCNN 模型中的分类器，形成了一个能够定位和识别未见类别的零样本食品图像检测器。

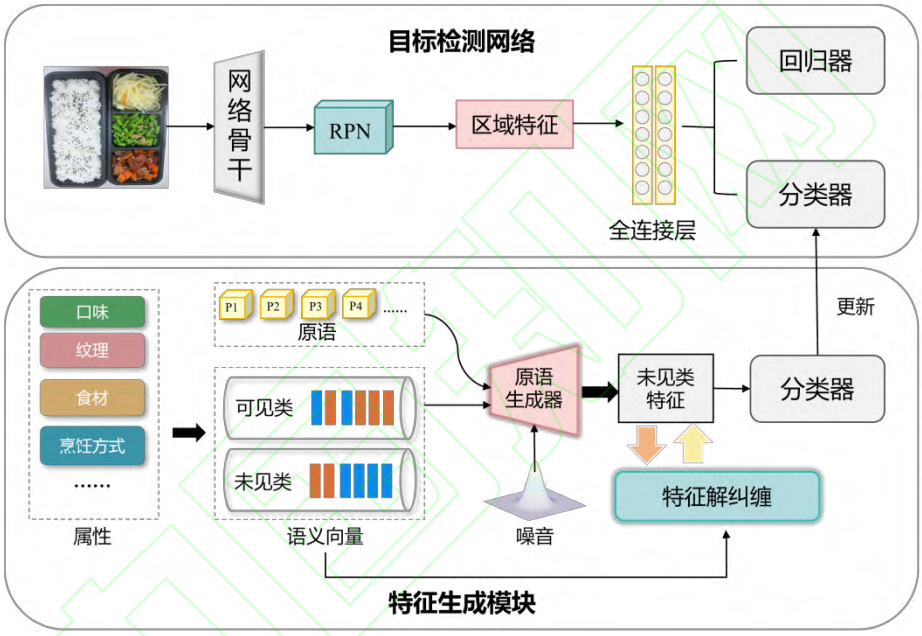


图 7　零样本食品图像检测总体框架

Fig. 7　The overall framework of zero-shot food image detection

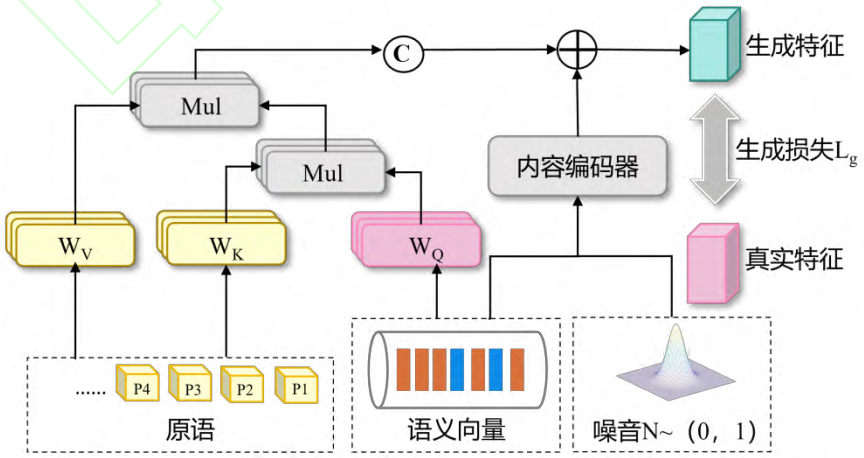


图 8　基于 Transformer 的基元生成器

Fig. 8　Transformer-based primitive generator

1.3 实验方法

1.3.1 实验环境 所提出模型的训练和测试均是基于 Pytorch 深度学习框架完成的。硬件环境 CPU 采用 Inter®Core(TM) i9-9820X，主频 3.30 GHz，2 张 NVIDIA RTX2080Ti GPU，内存 128 GB，显存 22 GB。软件环境采用 Ubuntu18.04，Python3.6 的编程环境。

1.3.2 实验细节 本文的目标检测模块采用的是以 ResNet-101 为骨干网络的 Faster-RCNN 模型，在生成器框架中， Transformer 的层数、损失权重 λ（公式 8）、温度 τ、σ 分别设置为 3、0.002、0.1、{2、5、10、20、40、60}。编码器E_R和E_U都是包含隐藏层，LeakyReLU 激活和 Dropout 的多层感知机（MLP）。E_D是由两个堆叠的单个 MLP 层构成。E_R、E_U和E_D的参数使用的为 Adam 优化器，初始学习率设定为2×10^{-4}。

1.3.3 基元生成器的构建 如图 8 所示，本文使用 Transformer 架构来构建基元生成器。首先，随机初始化一组可学习基元，表示为$P=\{p_i\}_{i=1}^{N}$，其中$p_i \in R^{d_k}$，d_k是通道数，N 是基元数量。这些基元包含了与食品类别相关的细粒度属性，例如成分、口味、颜色等。这些基元类型的不同组合构成了食品类别的不同表示。首先，对这些基元进行自注意力机制[23]，以此来构造基元属性之间的关系图，自注意力的计算过程如公式 1 所示。接下来，使用两个不同的线性层W_K和W_V来处理基元P，以获得交叉注意的键 Key 和值 Value，分别表示为 K 和 V。然后，将语义嵌入作为查询 Query，表示为Q，最后Q、K、V三者执行交叉注意力：

$$\mathrm{selfAttention}(Q,K,V)=\mathrm{softmax}\left(\frac{Q\cdot K^{T}}{\sqrt{d_k}}\right)\cdot V \tag{1}$$

$$A_g = w_1\left(\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V+W+N\right) \tag{2}$$

其中，A_g表示合成的食品特征，w_1是线性层，W表示语义嵌入，N表示具有固定高斯分布的随机样本。不同于通过多个线性层处理语义嵌入来生成特征，通过加权组装这些丰富的基元来合成视觉特征，这样能够提供更多样和更丰富的表示。此外，对于语义空间中共享的一些相似性的食品相关类别，基元提供了一种明确的方式来表达这种相似性。例如，西红柿炒鸡蛋和西红柿鸡蛋汤都具有西红柿和鸡蛋这两种成分属性，因此与西红柿和鸡蛋相关的基元对西红柿炒鸡蛋和西红柿鸡蛋汤的语义嵌入查询Q表现出较高的响应。利用这些描述细粒度属性的基元，可以很容易地构造不同的类别表示，并将可见类的知识转移到未见类。

本文定义基元生成器的生成损失L_G，以此来减少两个概率分布之间的最大平均差异：

$$L_G=\sum_{a,\hat{a}\in A_r}k(a,\hat{a})+\sum_{a',\hat{a}'\in A_g}k(a',\hat{a}')-2\sum_{a\in A_r}\sum_{a'\in A_g}k(a,a') \tag{3}$$

$$k(a,a')=\exp\left(-\frac{1}{2\sigma^{2}}\|a-a'\|^{2}\right) \tag{4}$$

其中，A_r和A_g分别表示食品类别的真实视觉特征和合成视觉特征。k是一个核函数，其中σ为带宽参数。当一个来自未见类的语义嵌入被输入到训练好的基元生成器中时，就可以得到它对应类别的合成视觉特征。然后，用可见类别的真实类特征和未见类别的合成类特征来重新训练分类器。在整个过程中，基元极大地增强了生成器的表达多样性和有效性，从而更好地解决了偏见问题。

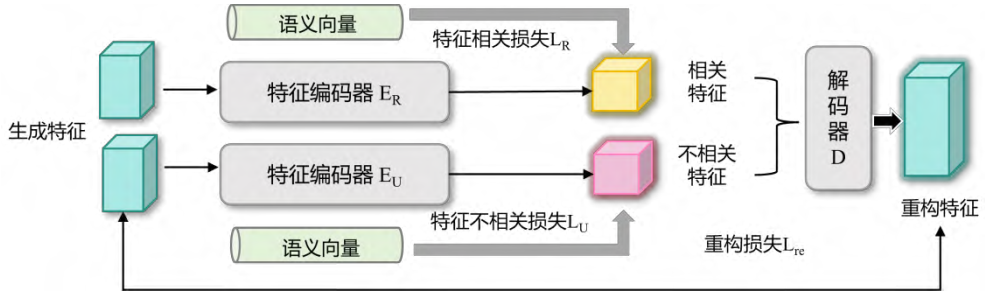


图 9　特征解纠缠组件

Fig. 9　Feature disentanglement component

1.3.4　特征解纠缠组件　不同的食品类别之间有着各自的联系和区别。以红烧肉、红烧排骨和清蒸鱼为例，显然红烧肉和红烧排骨之间的关系比红烧肉和清蒸鱼之间的关系更密切。语义空间中的类关系是强大的先验知识，而特定于类别的特征生成并没有显式地利用这种关系。本文用语义嵌入建立了这样的关系，并探索将这些知识转移到视觉空间，根据类别关系进行语义视觉对齐[32-33]。通过考虑这种关系，对未见类别的特征生成有更多的约束，以拉或推它们与可见类别的距离。

食品图像的视觉特征并不完全与其语义表示一致，而是包含更丰富的信息，包括语义相关和语义无关的视觉特征。语义无关的特征可能具有很强的视觉线索，有助于分类，但与语义表征的相关性较低。直接将语义嵌入与原始视觉特征对齐会混淆生成器，降低其对未见类别的泛化能力。为了解决这个问题，本文将视觉特征分为语义相关和语义无关的视觉特征。给定一个特征a_i，其中，$a_i \in A$是来自主干或生成器的视觉特征，如图 9 所示，特征解纠缠组件旨在学习如何解纠缠和重建a_i本身。使用编码器E_R来提取语义相关的特征，即$\dot{a}_i = E_R(a_i)$。然后，计算语义相关特征$\dot{a}_i$和语义嵌入$W = \{W_1, \cdots, W_{S+U}\}$的相关性得分。使用交叉熵损失来训练$E_R$，以赋予语义相关的特征$\dot{a}_i$有区别的语义知识，即

$$L_R = -\sum_i \sum_k l([\dot{a}_i] = k) \log \frac{\exp(\dot{a}_i W_k / \tau)}{\sum_k \exp(\dot{a}_i W_k / \tau)} \tag{5}$$

其中$[\dot{a}_i]$是$\dot{a}_i$的真实框，$l(\cdot)$是一个指示函数，如果条件为真，输出为 1，否则输出 0，τ是温度参数。使用另一个编码器E_U来提取语义无关的特征，记为$\ddot{a}_i = E_U(a_i)$，假设语义无关的特征具有正态分布$N(0, 1)$，使用 KL 散度来约束分布范围：

$$L_U = \sum_i D_{KL}[\ddot{a}_i \| N(0,1)] \tag{6}$$

其中，$D_{KL}[p \| q] = \int p(z) \log \frac{p(z)}{q(z)}$，这使得每个类都有自己独立且不同的特征组件。为了推动网络提取更具代表性的语义相关特征并保留视觉特征信息，在l_1损失下使用解码器D_e重建特征：

$$L_{Re} = \sum_i \left\| a_i - D_e(\dot{a}_i, \ddot{a}_i) \right\|_1 \tag{7}$$

然后，进行语义相关视觉空间和语义空间之间的关系对齐。使用 KL 散度使任意两个语义相关特$\dot{a}_i$和$\dot{a}_j$的相似度达到它们对应的语义嵌入$w_{[\dot{a}_i]}$和$w_{[\dot{a}_j]}$的相似度，即

$$L_{Alig}=D_{KL}\left[\frac{\dot{a}_i\dot{a}_j}{\|\dot{a}_i\|\|\dot{a}_j\|}\right]/\tau\|\left[\frac{w_{[\dot{a}_i]}w_{[\dot{a}_j]}}{\|w_{[\dot{a}_i]}\|\|w_{[\dot{a}_j]}\|}/\tau\right] \tag{8}$$

其中，$[\dot{a}_i]$是$\dot{a}_i$的真实框，τ是温度参数。可见类$\dot{a}_i^s$是真实的特征或合成特征，而未见类$\dot{a}_i^u$仅来自生成器的合成特征。本文中有两种对齐，组内对齐和组间对齐，在等式（8）中有不同的侧重点。当$\dot{a}_i$和$\dot{a}_j$来自同一组时，例如，$\dot{a}_i^s$和$\dot{a}_j^s$来自已知类别，它是组内对齐，这有助于通过关系作为约束提取更好的类表示。当它们来自不同的组时，例如，$\dot{a}_i^s$来自可见类别组和$\dot{a}_i^u$来自未见类别组，这是组间对齐，旨在将已知类别和未知类别的关系进行知识迁移。组间对齐对已知类别和未知类别的真实的特征和合成特征之间的关系给出了约束，它大大提高了模型对未知类别的适应性和泛化能力，整个基元生成器的损失如下：

$$L_T=L_G+\lambda\left(L_R+L_U+L_{Re}+L_{Alig}\right) \tag{9}$$

其中，λ是控制语义解纠缠组件的权重，一旦生成器训练完成后，它就可以为看不见的食品类别生成特征，最后与可见类的真实特征一起训练新的分类器。

2 结果与分析

2.1 评估方案

对于 UEC-FOOD256 和 ZSFooD，实验中使用的交并比（Intersection over Union，IoU）默认为 0.5。在目标检测中，IoU 即为计算预测边界框与真实边界的重叠程度，重叠程度越高，说明越接近真实框。IoU 计算的是“预测的边框”和“真实的边框”的交集和并集的比值，其计算公式如下：

$$IoU=\frac{S_b}{S_a} \tag{10}$$

其中，S_b表示预测边界框与真实边界的重叠区域，$S_a=S_1+S_2$，S_1和S_2分别表示预测区域和真实区域的面积，如图 10 所示。

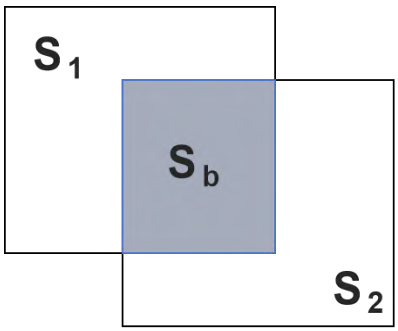


图 10　IoU 示意图

Fig. 10　IoU diagram

为衡量训练模型的性能，实验使用了平均精度（Mean Average Precision，mAP）和召回率@100 作为模型的评价指标。mAP 是通过对不同类别的 AP 值进行平均计算得到的，它反映了模型在多个类别上的整体性能。每个类别的 AP 值，是通过构建精准率Precision与召回

率Recall组成的 PR 曲线来计算的，即该曲线下方的区域与坐标轴之间的面积。

精准率的定义是：在所有模型预测为正例的情况中，实际上正确地被识别为正例的比例。其计算公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

召回率的定义是：在所有真实为正例的情况中，被模型正确预测出的比例。它可以被视为正确识别的目标数量占所有真实正例的比例。其计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

其中，真正例TP是正确预测的正例数量，假正例FP是错误预测为正例的负例数量，假负例FN是未能正确预测的正例数量，它们均来自混淆矩阵，如表 2 所示。

此外，本文还报告了广义零样本检测（Generalized Zero-Shot Detection，GZSD）设置下方法的性能。不同于 ZSD，GZSD 设置下测试图像包含可见类和未见类。可见类和未见类的调和平均值（HM，Harmonic Mean）是用于评价 GZSD 设置下检测性能的主要指标，其计算方式如下所示：

$$\text{HM} = \frac{2 \times \text{P}_{\text{seen}} \times \text{P}_{\text{unseen}}}{\text{P}_{\text{seen}} + \text{P}_{\text{unseen}}} \tag{13}$$

其中， P_{seen}和P_{unseen}分别表示在可见类和未见类上计算得到的评估指标。

表 2　混淆矩阵示意图

Table 2　Confusion matrix diagram

真实情况	预测结果	
	正例	负例
正例	TP（真正例）	FN（假负例）
负例	FP（假正例）	TN（真负例）

表 3　各类模型在 ZSFooD 上的性能比较（%）

Table 3　Performance comparison of various models on the ZSFooD （%）

指标	模型名称	ZSD	GZSD		
			S	U	HM
Recall@100	ConSE	39.7	58.0	38.1	46.4
	BLC	41.2	55.3	40.5	46.8
	CZSD	48.0	86.1	44.8	58.9
	SU	45.3	82.3	44.1	57.4
	RRFS	48.8	86.6	47.6	61.4
	Ours	**51.7**	**86.8**	**48.1**	**62.0**
mAP	ConSE	0.8	54.3	0.7	1.4
	BLC	1.1	51.1	0.9	1.8
	CZSD	4.0	81.2	2.1	4.1
	SU	3.9	79.1	2.3	4.5

	RRFS	4.3	**82.7**	2.7	5.2
	Ours	**4.9**	82.5	**3.0**	**5.8**

注：“S”表示可见类，“U”表示未见类，“HM”表示可见类和未见类的调和平均值，下同。

2.2 实验结果及分析

2.2.1 不同模型的对比实验 在表 3 中，比较了 ZSD 和 GZSD 设置下各类模型在 ZSFooD 数据集的性能，可以观察到，本文的方法在 Recall@100 和 mAP 这两方面优于所有比较的方法。与次优方法 RRFS 相比，本文的方法在 IoU=0.5 时将 ZSD 的 Recall@100 从 48.8%提高到 51.7%，将 mAP 从 4.3%提高到 4.9%。在 GZSD 设置下，本文的方法将“U”的 Recall@100 和 mAP 分别提高了 0.5%、0.3%，达到了最好的性能，在“HM”指标上 Recall@100 和 mAP 分别达到了 62.0%、5.8%最好的性能，这项指标表明本文的方法在可见类和未见类之间保持了良好的平衡，这得益于本文使用的基元生成器和特征解纠缠组件。

在表 4 中，在 UEC-FOOD256 上将本文的方法与最新方法进行了比较。与次优的方法相比，本文的方法在 IoU=0.5 时将 Recall@100 提高了 2.9%，将 mAP 提高了 0.5%。此外，将本文的方法 GZSD 场景下与其他方法进行了比较，可以观察到也实现了显著的性能增益，“U”和“HM”分别从 22.9%和 21.4%提高到 24.5%和 22.0%。

表 4 各类模型在 UEC-FOOD256 数据集的性能比较（%）

Table 4 Performance comparison of various models on the UEC-FOOD256（%）

指标	模型名称	ZSD	GZSD		
			S	U	HM
Recall@100	ConSE	54.4	50.1	38.2	43.3
	BLC	58.9	55.3	43.8	48.9
	CZSD	60.7	**57.6**	45.5	50.8
	SU	61.9	52.5	52.8	52.6
	RRFS	64.8	54.9	55.1	55.0
	Ours	**67.7**	54.8	**57.2**	**55.9**
mAP	ConSE	11.3	19.7	9.0	12.4
	BLC	19.2	20.5	15.2	17.5
	CZSD	22.0	**20.8**	16.2	18.2
	SU	22.4	19.3	20.1	19.7
	RRFS	23.6	20.1	22.9	21.4
	Ours	**24.1**	20.0	**24.5**	**22.0**

表 5 模型在 ZSFOOD 上 mAP 的消融实验（%）

Table 5 The ablation experiment of the model's mAP on ZSFOOD（%）

数据集	方法		ZSD	GZSD		
	基元生成器	语义接纠缠组件		S	U	HM
ZSFOOD	-	-	4.3	**82.7**	2.7	5.2

	√		4.7	82.5	2.9	5.6
	√	√	**4.9**	82.5	**3.0**	**5.8**

表 6　模型在 UEC-FOOD256 上 mAP 的消融实验（%）

Table 6　The ablation experiment of the model's mAP on UEC-FOOD256（%）

数据集	方法		ZSD	GZSD		
	基元生成器	语义接纠缠组件		S	U	HM
UEC-FOOD256	-	-	23.6	**20.1**	22.9	21.4
	√		24.8	20.0	24.1	21.7
	√	√	**25.1**	20.0	**24.5**	**21.9**

2.2.2　消融研究　为了进一步验证本文的方法，在 ZSFooD 和 UEC-FOOD256 两个数据集上对基元生成器和特征解纠缠两个关键模块进行了定量消融分析。表 5 和表 6 分别报告了在 ZSD 和 GZSD 两种设置下，ZSFooD 和 UEC-FOOD256 数据集 mAP 的性能，可以看出结合基元生成器的模型将 ZSD 分别提高了 0.4%、1.2%，“U”分别提高了 0.2%、0.4%，将“HM”在 ZSFooD 和 UEC-FOOD256 上分别提高了 0.4%、0.2%，这表明具有更丰富的基元信息合成器可以考虑到食品类内多样性、类间可区分性等细粒度特征以及食材、烹饪手法、形状、颜色、纹理等丰富的属性信息，从而合成更具有鲁棒性的视觉特征。可以观察到，当加入了特征解纠缠组件时，ZSD 性能在 ZSFooD 上提高到 4.9%，在 UEC-FOOD256 上提高到 25.1%。这些性能增益证明了所构建的特征解缠组件通过建立语义相关的视觉空间可以促进关系对齐，进而表明了语义信息对图像特征具有重要的指导作用，对于 “U”和“HM”，与基线相比，本文的方法在 ZSFooD 和 UEC-FOOD256 分别提高到 3.0%、5.8%和 24.5%、21.9%，这表明所提出的基元生成器和特征解纠缠模块是至关重要的，它们能够很好地桥接语义与视觉空间，并将语义关系应用到视觉特征的学习。

b.UEC-FOOD256

图 11　ZSFood 和 UEC-FOOD256 数据集上一些可视化的检测结果

Fig.11　Some visualized detection results on ZSFooD and UEC-FOOD256 datasets

注：每类数据集的左列为 ZSD 的检测结果，右列为 GZSD 的检测结果，其中，可见类显示在绿色框中，未见类显示在红色框中。

2.2.3　定性结果　为了直观地评估所提出方法所取得的检测成果，在图 11 中展示了 ZSD 和 GZSD 两种设置下，使用 ZSFooD 和 UEC-FOOD256 数据集进行零样本食品图像检测的定性结果。在 ZSD 设置中，模型仅专注于识别检测未见类的食品对象，而在 GZSD 设置下，可以同时检测出可见类和未见类的食品对象。在每个数据集的展示中，分别用第一列和第二列来展示 ZSD 和 GZSD 的检测结果。其中，未见的食品对象用红色框标注，而可见的食品对象则以绿色框表示。本文的方法在不同的设置下准确地检测到了已见和未见的食品对象。此外，该方法还表现出了对不同数据集的广泛适用性，证明了其在实际应用中的巨大潜力。

3　结论

本文提出了一种基于 Transformer 的零样本食品图像检测方法，以解决实际场景中食品

a.ZSFooD

类别更新、细粒度特征以及复杂属性所带来的挑战。所提出的模型使用了基元生成器来合成未见类特征，并通过特征解纠缠的增强组件将食品类别的语义知识转移到其视觉特征，从而能够在处理食品类别周期性更新时显示出更高的适应性。本文模型在 ZSFooD 和 UEC-FOOD256 两个公开数据集进行实验，与基线模型相比，未见类别的平均精度在 ZSFooD 数据集上从 4.3%提升至 4.9%，在 UEC-FOOD256 数据集上从 23.6%提升至 24.1%，分别取得了最优的结果，验证了该方法在解决零样本食品图像检测问题方面的有效性和鲁棒性。在消融研究中，进一步分析了所提出的关键模块在 ZSFooD 和 UEC-FOOD256 上的有效性，这为食品零样本检测任务提供了新的思路和解决方案。

参考文献：

[1] 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.

[2] LU Y, STATHOPOULOU T, VASILOGLOU M F, et al. An artificial intelligence-based system to assess nutrient intake for hospitalised patients[J]. IEEE Transactions on Multimedia (TMM), 2020, 23: 1136-1147.

[3] MIN W Q, JIANG S Q, JAIN R. Food recommendation: Framework, existing solutions, and challenges[J]. IEEE Transactions on Multimedia, 2019, 22(10): 2659-2671.

[4] UMMADISINGU A, TAKAHASHI K, FUKAYA N. Cluttered food grasping with adaptive fingers and synthetic-data trained object detection[C]//2022 International Conference on Robotics and Automation. IEEE, 2022: 8290-8297.

[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] ASLAN S, CIOCCA G, MAZZINI D, et al. Benchmarking algorithms for food localization and semantic segmentation[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(12): 2827-2847.

[7] EGE T, YANAI K. Simultaneous estimation of food categories and calories with multi-task CNN[C]//2017 fifteenth IAPR international conference on machine vision applications. IEEE, 2017: 198-201.

[8] SUN J, RADECKA K, ZILIC Z. Foodtracker: A real-time food detection mobile application by deep convolutional neural

networks[J]. arXiv preprint arXiv:1909.05994, 2019.

[9] 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.

[10] 李楠. 基于生成模型的零样本图像分类研究与实现[D].成都: 电子科技大学, 2023. [LI Nan. Research and implementation of zero-shot image classification based on generative models[D]. Chengdu: University of Electronic Science and Technology of China, 2023.]

[11] BANSAL A, SIKKA K, SHARMA G, et al. Zero-shot object detection[C]//Proceedings of the European conference on computer vision, 2018: 384-400.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]//1st International Conference on Learning Representations, 2013.

[13] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing, 2014: 1532-1543.

[14] NOROUZI M, MIKOLOV T, BENGIO S, et al. Zero-shot learning by convex combination of semantic embeddings[J]. arXiv preprint arXiv:1312.5650, 2013.

[15] ZHENG Y, HUANG R R, HAN C Q, et al. Background learnable cascade for zero-shot object detection[C]//Proceedings of the Asian Conference on Computer Vision, 2020.

[16] YAN C X, CHANG X J, LUO M N, et al. Semantics-guided contrastive network for zero-shot object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[17] HAYAT N, HAYAT M, RAHMAN S, et al. Synthesizing the unseen for zero-shot object detection[C]//Proceedings of the Asian Conference on Computer Vision. 2020.

[18] HUANG P L, HAN J W, CHENG D C, et al. Robust region feature synthesizer for zero-shot object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 7622-7631.

[19] CRESWELL A, WHITE T, Dumoulin V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[20] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[21] WU X, YU S, LIM E P, et al. OVFoodSeg: Elevating open-vocabulary food image segmentation via image-informed textual representation[J]. arXiv preprint arXiv:2404.01409, 2024.

[22] LI G, LI Y, LIU J, et al. ESE-GAN: Zero-shot food image classification based on low dimensional embedding of visual features[J]. IEEE Transactions on Multimedia, 2024.

[23] RAMDANI A, VIRGONO A, SETIANINGSIH C. Food detection with image processing using convolutional neural network (cnn) method[C]//Proceedings of the IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). 2020: 91-96.

[24] HE S T, DING H H, JIANG W. Primitive generation and semantic-related alignment for universal zero-shot segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 11238-11247.

[25] SHIMODA W, YANAI K. Webly-supervised food detection with foodness proposal[J]. IEICE Transactions on Information and Systems, 2019, 102(7): 1230-1239.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30: 5998—6008.

[27] KAWANO Y, YANAI K. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation[C]//Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13. Springer International Publishing, 2015: 3-17.

[28] ZHOU P F, MIN W Q, Zhang Y, et al. SeeDS: Semantic separable diffusion synthesizer for zero-shot food detection[C]//Proceedings of the 31st ACM International Conference on Multimedia, 2023: 8157-8166.

[29] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision

[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[30] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:770-778.

[32] DAI J, HU X, LI M, et al. The multi-learning for food analyses in computer vision: a survey[J]. Multimedia Tools and Applications, 2023, 82(17): 25615-25650.

[33] LAWRYNOWICZ A, WRÓBLEWSKA A, KALISKA A, et al. Fine-grained and complex food entity recognition benchmark for ingredient substitution[C]//Proceedings of the 12th Knowledge Capture Conference 2023. 2023: 25-29.