arXiv:2304.07387v1 [cs.MM] 14 Apr 2023

基于加权对抗学习的跨域食物图像到食谱检索

Bin Zhu[1], Chong-Wah Ngo[2],Jingjing Chen[3],和 Wing-Kwong Chan[1]

[1]香港城市大学[2]新加坡管理
大学[3]复旦大学

摘要

食物图像到食谱检索旨在学习一个嵌入空间，将食谱中的丰富语义与食物图像中的视觉内容联系起来，以实现跨模态检索。现有研究工作在学习这一空间时，假设所有图像-食谱训练样本对都属于同一种菜系。因此，尽管文献中报告了出色的性能，这种空间仍无法迁移以检索其他菜系的食谱。本文旨在通过跨域食物图像到食谱检索解决这一问题：利用源域（一个菜系）中丰富的图像-食谱样本对，使嵌入空间能够泛化到目标域（另一种菜系），即使目标域中没有可与食谱配对用于训练的图像。基于不同源样本的重要性应有所不同这一想法，本文提出了两种用于跨域食物图像到食谱检索的新机制，即源数据选择器和加权跨模态对抗学习。前者旨在选择与目标数据相似的源样本，并筛除具有显著差异的样本用于训练。后者能够为与目标数据更相似的源样本赋予较高权重，并为具有显著差异的样本赋予较低权重，从而在跨模态学习和对抗学习中抑制这些样本。权重根据预训练源模型提取的食谱特征计算得出。在三种不同菜系（川菜、粤菜和和食）上的实验表明，所提方法在所有迁移任务中均取得了最先进的性能。

1. 引言

众所周知，深度学习的发展在很大程度上依赖于充足的标注数据。图像分类等领域的成功案例包括 [14, 16], 目标检测 [28, 13] 和语义分割 [20]。然而，为 ImageNet 等大规模数据集提供标注既耗时又成本高昂。为缓解这一问题，领域自适应 [24] 旨在将利用现成的大量标注数据（源域）训练的模型所学到的知识迁移出来，以训练一个能够在一组相关数据（目标域）上良好泛化的新模型。由于新模型需要处理不同的数据（目标域），将模型从源域迁移到目标域很容易导致性能下降。

本文研究跨域食物图像到食谱检索问题。具体而言，在训练阶段，目标域仅包含食谱。这种设定的动机有两方面。一方面，获取充足的跨模态学习配对数据并非总是易事，例如，在社交媒体（如食谱分享网站）发展较不成熟的国家、某些地区的少数菜系，以及数据收集中的知识产权问题。另一方面，分享网站上的大多数食谱都没有关联食物图像。以 Recipe 1M [30]数据集为例，只有 33% 的食谱至少包含一张图像。本文旨在利用拥有大量图像—食谱配对数据的源域，学习食谱的嵌入表示，使其与对应食谱的食物图像相近。请注意，训练期间无法获取这些图像。测试时，给定一张查询食物图像，任务是检索该图像对应的食谱。如 [41], 所述，通过检索查询图像对应的食谱，可以直接提取食材以进行营养估计。此外，与分类方法 [2, 4, 23, 5]不同，后者需要为每张图像标注食物和食材标签，本方法无需任何标注工作。

现有的跨模态食谱检索工作 [30, 6,3, 40, 33, 10, 38, 29, 25, 35] 假设训练集和测试集的数据分布相同。相反，本文处理的是源数据和目标数据来自不同域的问题。在跨



图 1. 以川菜为源域、日本料理为目标域的示例。所提出的方法为与目标数据更相似的源样本赋予较高权重，并为那些具有鲜明特征的源样本赋予较低权重，以抑制其影响。每张食物图像下方均标有菜品标题。

域食物图像到食谱检索中，这指的是源域和目标域中的菜肴具有不同的菜系。不同菜系之间的域差异可能源于不同的食材组成、切割和烹饪方法、饮食文化等。例如，墨西哥菜系由于经常使用“辣椒”而往往较为辛辣。而法国菜系不那么辛辣，法国菜肴中的常见食材是“奶酪”和“松露”。因此，该问题颇具挑战性，其原因在于域差异（即不同菜系）、源域和目标域之间使用多种模态，以及目标域中没有可利用的配对信息。

在现有文献中，大多数研究关注单模态领域自适应 [32, 21, 36, 11, 31, 37,7, 34, 8]。弥合域差异的方法包括度量测量 [32, 21, 36]或对抗学习 [12, 11, 31, 37, 7, 34, 8]。近年来，少数研究 [27, 17, 41] 从跨模态视角探索域迁移，将模态拓展至文本、音频、图像等。这些研究的侧重点各不相同，包括利用带标签的源域配对数据预测目标域跨模态配对数据的类别标签 [27], 、利用单一模态的带标签源域数据为多种模态的目标数据添加标签 [17], , 以及在食物领域中利用完整配对的源域数据对部分目标数据进行配对 [41]。由于食谱包含丰富的文本信息，包括标题、食材列表和描述烹饪过程的烹饪步骤，因此使用食谱所取得的改进比仅使用图像更为显著 [41]。因此，本文聚焦于目标域仅包含食谱的情形。此外， [41]使用域分类器对齐源域和目标域分布，却未考虑不同菜系内部的独特属性（例如食材和烹饪风格），因此具有鲜明特征的源样本可能在训练过程中引入噪声，从而限制迁移性能。

本文通过提出数据选择和重加权方案来解决这一局限。具体而言，我们认为应根据源域数据与目标数据的相似度，为源域数据分配不同权重以使数据分布与目标域对齐。其直觉在于，不同样本所体现的特定领域（菜系）属性程度各不相同。因此，应降低独特源样本的权重，而更相似的样本则应在特征对齐中占据更重要的地位。为此，所提出的源批次选择器（SBS）会挑选与目标数据更相似的样本，组成用于训练的源批次。此外，所提出的加权对抗跨模态学习分别通过加权三元组损失和加权对抗损失，对跨模态学习和对抗学习进行重加权；这些权重根据预训练源模型提取的源域和目标域特征计算得出。

2. 相关工作

领域自适应旨在解决从源域向目标域进行知识迁移的问题 [24]。由于源域和目标域之间的数据分布存在域差异，迁移可能会对性能产生不利影响。文献中的大多数研究都试图通过缩小特征空间中源数据与目标数据之间的差距，来减轻域偏移的影响。一方面，可以直接使用最大均值差异（MMD）等度量来计算域差异 [32, 21, 36]。 [32] 中采用结合 MMD 的自适应层来学习域不变的深度特征。 [21] 在此基础上提出 MMD 的多核变体，通过选择最优核来提高特征的可迁移性。在 [36], 中，加权 MMD 为源域引入类别特定权重，以解决跨域类别偏差问题。

另一方面，其他一些研究则重点通过使用域分类器 [11, 31, 37, 7, 34, 8]以对抗方式学习领域不变特征。领域对抗神经网络（DANN） [11] 采用域分类器来区分输入来自源域还是目标域。特征提取器通过梯度反转层与域分类器相连，该层会在反向传播过程中反转梯度。在 [31],

训练分为两个阶段：第一阶段仅使用源域数据首先学习源模型，然后将判别器与预训练源模型结合，用于通过对抗适应学习目标模型。文献 [37, 7]还探索了基于加权的对抗方法。再加权对抗适应网络（RAAN） [7] 通过最小化地球移动距离，对源域标签分布进行重新加权，使其匹配目标标签分布，从而通过对抗学习实现分类器自适应。在 [37], 中，额外的域分类器被用于计算源样本的权重。随后，将这些权重与源域数据和目标数据一同输入第二个域分类器进行对抗训练。在 [8], 中，设计了逐渐消失的桥接机制，以在源域和目标域之间生成一个中间域，从而增强对抗训练的平衡性。渐进式对抗网络（PAN） [34] 被提出，将课程学习 [1] 与对抗学习相结合，以实现细粒度领域自适应。总体而言，本文属于基于加权的对抗方法。不同于 [37, 7], ，我们关注的是跨域食物图像到食谱的检索，而非视觉领域自适应。此外，本文中的权重由预训练源模型生成，而不是通过额外的判别器 [37]或最小化地球移动距离 [7]获得。我们还提出了源批次选择器，用于筛选源样本，并提出了专门针对跨模态对抗学习的加权方案。

现有的大多数领域自适应方法只考虑单一模态。跨模态迁移鲜有研究 [27, 17, 41]。 [17] 重点研究如何将知识从单一模态的源域迁移到跨模态目标域。例如，源域是一组带标签的图像，而目标域可以是无标签的文本、视频或音频。本文提出一种模态对抗混合迁移网络，通过同时在任务专属层中采用最大均值差异，并利用域分类器进行对抗学习，实现不同模态特征的语义对齐。在 [27],中，源域和目标数据都包含图像-音频对，且只有源样本带有类别标签。其目标是利用从源域学到的知识为目标域中的配对数据添加标签。利用从不同模态提取的特征计算注意力分数，再将这些特征融合以进行分类。该方法采用三个域分类器，分别对模态、注意力分数和融合特征进行判别，从而实现特征对齐。

与本文最相关的工作是跨领域跨模态食物迁移（CCFT） [41], ，该研究探讨了食物领域中的跨模态迁移。CCFT 考虑在不同模态之间，以及从完整数据到部分数据进行知识迁移。例如，源域包含用于训练的图像和食谱，而目标域在训练时仅包含食谱，甚至只有菜肴标题。尽管这一问题场景很有趣，CCFT 仅利用食谱和图像特征的域分类器，学习域不变特征，以实现跨模态和不完整数据迁移。由于不同菜系具有各自独特的属性，源域中的部分样本可能无助于迁移，反而会增加训练噪声。通过对源域和目标域的所有特征进行同等对齐，CCFT 的迁移性能受到限制。本文针对这一局限，提出数据选择和重新加权机制，在训练时筛除具有显著差异的源样本，并根据每个源样本与目标域的相似度赋予不同权重，以用于跨模态对抗学习。近期的工作 [39]通过交换源域和目标域之间的食谱章节，探索了混合增强在跨语言食谱适配中的应用。由于在不使用机器翻译的情况下，语言差异仍然极难克服， [39], 因此我们专注于同一语言中的食谱适配。

3. 方法

借助拥有大量训练数据的源域，本文旨在实现目标域上的图像到食谱检索。本文关注的场景是目标域中仅有食谱。据报道， [41] 由于食谱提供了信息丰富的文本描述，因此，相较于图像，食谱更有助于提升领域迁移性能。单独的食物图像若未与食谱配对，所能达到的性能有限。将源域记为 $\mathcal{D}^s = \{(r_i^s,\ v_i^s)\}_{i=1}^{N^s}$，其中包含 N^s 食谱-图像对； r_i^s 表示源食谱， v_i^s 表示与之配对的菜品图片。同时，目标域包含一组食谱，记为 $\mathcal{D}^t = \{(r_j^t)\}_{j=1}^{N^t}$，其中 r_j^t 表示目标食谱。上标 s 和 t 分别表示源域和目标域。值得注意的是，假设源域和目标域中的食谱均完整；换言之，每份食谱都包含三个部分（即标题、食材列表和烹饪步骤）。

图2展示了所提出的加权跨模态对抗学习模型的网络架构。基于跨域食物图像到食谱检索中不应将所有样本一视同仁这一理念，我们设计了两种关键机制：其一是源批次选择器（SBS，第3.1节），它能够挑选出与目标域更相似的源样本用于训练；其二是加权跨模态对抗学习（第3.2节），通过为选定的源批次数据分配不同权重，使模型不仅能在跨模态学习中，还能在对抗学习中对齐源域和目标域分布。

3.1. 源批次选择器

源批次选择器旨在从一开始就筛除远离目标数据的“劣质”源样本，

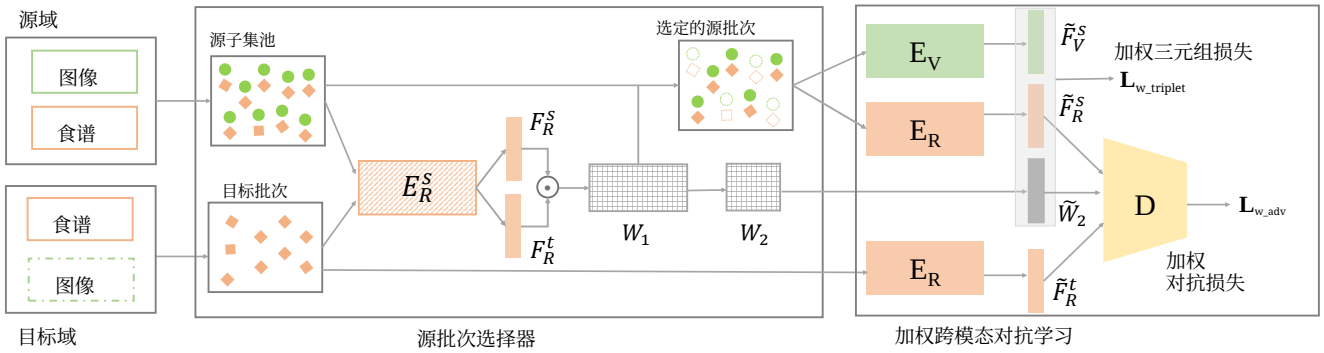

图2。所提出的方法概览。源批次选择器旨在从目标数据中筛选出具有代表性的源样本，而加权跨模态对抗学习则为更相似的样本分配较高权重，以在跨模态学习和对抗学习中实现源域和目标域分布对齐。矩阵 W_1 和 W_2 均由预训练源模型（E_R^s）生成，该模型在训练过程中被冻结。

我们认为，模型应更多关注与目标数据相似的数据，而不是源域中独特或具有区分性的样本。如图2所示，分别从源域数据和目标数据中随机抽取一个包含n^s 个食谱-图像对的源子集池和一个包含 n^t 个食谱的目标批次（令$n^s \geq n^t$）。随后，使用在源域数据上预训练的食谱编码器 E_R^s 提取源域和目标域的食谱特征，分别记为 $F_R^s = \{(f_{r_i^s})\}_{i=1}^{n^s}$ 和$F_R^t = \{(f_{r_j^t})\}_{j=1}^{n^t}$，其中$f_{r_i^s}$ 和 $f_{r_j^t}$分别是 r_i^s 和 r_j^t 的特征。请注意，在整个训练过程中， E_R^s 的权重均保持冻结。接着，计算矩阵 $W_1 \in \mathbb{R}^{n^t \times n^s}$，用于刻画目标批次中的食谱与源子集池中的食谱之间的成对相似度，计算公式如下：

$$W_1(j,i) = \frac{f_{r_j^t} \cdot f_{r_i^s}}{\|f_{r_j^t}\| \|f_{r_i^s}\|}. \tag{1}$$

基于 W_1，通过针对每个目标食谱从源池中收集最相似的前 K 个样本，构建新的源子集，其中 K 为超参数。实际上，当 $n^s \geq 2n^t$ 时，我们将 $K = 2$ 设为 。因此，新子集中的其余样本与至少一个目标样本具有一定相似性，而具有代表性的样本则被排除在训练之外。这样，便可以从新子集中随机选取源批次，使其大小与目标批次相匹配，以便进行后续的加权跨模态对抗学习。此外，由 W_1 推导出矩阵 $W_2 \in \mathbb{R}^{n \times n}$，该矩阵表示所选源批次与目标批次中食谱之间的相似性，其中 n 为批次大小。

3.2. 加权跨模态对抗学习

加权跨模态对抗学习的基本思想是，为与目标数据更相似的源样本分配较高权重，并以较低权重抑制不相似的样本。如图2所示，源域和目标域食谱同时输入一个可学习的共享权重食谱编码器 E_R。由 E_R 提取的源域和目标域食谱特征~ ~分别表示为 $F_R^s = \{(f_{\tilde{r}_i^s})\}_{i=1}^n$和 $\tilde{F}_R^t = \{(\tilde{f}_{\tilde{r}_j^t})\}_{j=1}^n$，其中 $\tilde{f}_{\tilde{r}_i^s}$是 $\tilde{r}_i^s$ 的特征~， $f_{\tilde{r}_j^t}$是 $\tilde{r}_j^t$的特征，而 n 是批次大小。此外，源图像特征由一个可学习的~ ~图像编码器 E_V 提取，记为 $F_V^s = \{(f_{\tilde{v}_i^s})\}_{i=1}^n$，其中~$f_{\tilde{v}_i^s}$是 $\tilde{v}_i^s$的特征。

~为获得用于模型训练的一维权重向量 $W_2 \in \mathbb{R}^n$，对相似度矩阵 W_2 进一步进行如下处理：

$$\begin{aligned}
\tilde{W}_2(i) &\leftarrow \sum_{j=1}^{n} W_2(i,j), \\
\tilde{W}_2(i) &\leftarrow \frac{\tilde{W}_2(i) - \min(\tilde{W}_2)}{\max(\tilde{W}_2) - \min(\tilde{W}_2)}, \\
\tilde{W}_2(i) &\leftarrow \frac{n\tilde{W}_2(i)}{\sum_{i=1}^{n} \tilde{W}_2(i)}.
\end{aligned} \tag{2}$$

对目标批次求和、归一化~并缩放后，权重向量 W_2 表示每个源样本相对于批次中目标数据的重要性。

所提出的方法通过以下目标函数进行训练：

$$\mathcal{L} = \mathcal{L}_{w_triplet}^s + \beta \mathcal{L}_{w_adv} + \lambda \mathcal{L}_{reg}^s, \tag{3}$$

其中， $\mathcal{L}_{w\ triplet}^s$ 是加权三元组损失， $\mathcal{L}_{w\ adv_}$是加权对抗损失， $\mathcal{L}_{reg}$ 是正则化损失。 β 和 λ 是权衡超参数。上标 s 表示这些损失仅应用于源域。

表1. 按中位数排名（MedR）衡量的跨域食物图像到食谱检索性能对比。请注意，“w/o w triplet”和“w/o w adv”分别表示使用未重新加权的原始三元组损失和对抗损失。预言机模型通过使用源域和目标域的全部配对数据进行监督训练，展示了性能上限。最后一列是所有迁移任务的平均值。“C”“Y”和“W”分别代表“川菜”“粤菜”和“和食”。

方法	C→W（1K）	Y→W（1K）	C→Y（1K）	C→Y（5K）	Y→C（1K）	Y→C（5K）	Avg.
随机方法	500	500	500	2500	500	2500	1,166.7
仅源域 [41]	33.6	19.2	4.8	19.6	12.5	54.7	24.1
IWAN [37]	33.0	18.8	4.7	18.5	12.2	54.4	23.6
CCFT [41]	11.7	15.7	4.2	16.6	8.8	38.6	15.9
去除 SBS 的所提方法	11.5	14.3	4.2	16.5	8.6	37.5	15.4
去除加权三元组损失的所提方法	11.3	13.7	4.1	16.4	8.6	36.1	15.0
去除加权对抗损失的所提方法	11.2	13.3	4.1	16.3	8.3	35.9	14.9
所提出的方法	**10.6**	**13.1**	**4.0**	**16.1**	**7.9**	**35.2**	**14.5**
Oracle（上界）	4.9	8.0	2.7	8.7	2.0	6.7	5.5

与近期跨模态食谱检索领域的研究 [10, 33, 40] 类似，我们采用三元组损失，使正食谱-图像对彼此靠近，并使负样本彼此远离。不同之处在于，我们在为每个样本计算三元组损失时对源域数据重新加权，以缩小源域和目标域之间的差距。加权三元组损失的形式化定义如下：

$$\begin{aligned}\mathcal{L}^s_{w_triplet} =& \tilde{W}_2[d(\tilde{f}_{\tilde{r}^s_a}, \tilde{f}_{\tilde{v}^s_p}) - d(\tilde{f}_{\tilde{r}^s_a}, \tilde{f}_{\tilde{v}^s_n}) + \alpha]_+ \\ &+ \tilde{W}_2[d(\tilde{f}_{\tilde{v}^s_a}, \tilde{f}_{\tilde{r}^s_p}) - d(\tilde{f}_{\tilde{v}^s_a}, \tilde{f}_{\tilde{r}^s_n}) + \alpha]_+,\end{aligned} \tag{4}$$

其中，$d(\cdot, \cdot)$表示余弦相似度距离，下标 a、p 和 n 分别表示锚点、正样本和负样本，超参数 α 为间隔。$\mathcal{L}^s_{w\ triplet}$ 包含两项，因为锚点可以是食谱特征，也可以是图像特征。

为对齐源域和目标域分布，我们采用域判别器 D 。以源域 $\tilde{F}^s_R$ 和目标域 $\tilde{F}^t_R$ 的食谱特征为输入，判别器与食谱编码器 E_R 进行对抗训练，以预测域标签。具体而言， E_R 经过训练以生成无法与 D 区分的食谱特征，而D 则最大化正确预测的概率。加权对抗损失定义如下：

$$\begin{aligned}\mathcal{L}_{w_adv} =& \tilde{W}_2\mathbb{E}_{\tilde{F}^s_R \sim p^s_R}[\log D(\tilde{F}^s_R)] \\ &+ \tilde{W}_2\mathbb{E}_{\tilde{F}^t_R \sim p^t_R}[\log(1 - D(\tilde{F}^t_R)].\end{aligned} \tag{5}$$

此外，还使用了 [41], 一个正则项 $\mathcal{L}^s_{reg}$ ，由两部分组成：多标签分类损失和重构损失。前者由图像特征预测食材时产生的误差计算得出，后者则来自真实图像与根据食谱特征重构的图像之间的差异。

4. 实验

4.1. 实验设置

数据集。 实验在三个数据集上进行，分别对应三个菜系，即四川（“川菜”）、粤菜（“粤菜”）和日本料理（“和食”） [41]。其中，川菜数据集包含 42,797 道食谱和 155,750 张图像，粤菜数据集包含 27,256 道食谱和 119,758 张图像，和食数据集包含 9,626 道食谱和 48,485 张图像。请注意，所有食谱均为中文。 [41]还提供了涵盖这三个菜系的 1,635 种常见食材列表。

评估指标。 Following [41], 在跨模态食谱检索 [30]设置下评估性能。给定一张食物图像作为查询，任务是检索相关食谱。采用中位排名（MedR）和前 K 名召回率（R@K）来评估迁移性能。MedR 是所有测试查询的真阳性排名位置的中位数。R@K 表示对应食谱排名位于前 K 个返回结果中的查询图像所占的百分比。因此，较低的 MedR 和较高的 R@K 表明迁移模型性能更好。请注意，实验重复进行 10 次，并报告 MedR 和 R@K 的平均结果。

基线方法。 我们将所提出的方法与两种最先进的方法 [37, 41] 以及两种基线方法——随机方法和仅源域方法进行比较。随机基线方法是根据查询图像对食谱进行随机排序。仅源域基线方法 [41]指仅使用源域数据训练标准的跨模态食谱检索模型，并直接将其应用于目标域进行测试。此外，我们还比较了一种基于加权的对抗域适应方法 [37] 和一种基于对抗的跨域食物图像到食谱检索方法 [41]。重要性加权对抗网络（IWAN） [37] 通过双判别器学习域不变特征，其中一个用于生成权重，另一个用于区分源域和目标域。一种

用于生成权重，另一种用于区分源域和目标域。源模型预先进行预训练，在目标模型与第二个判别器之间进行对抗训练时，其参数不会更新。据我们所知，CCFT [41] 是唯一专门针对跨域食物图像到食谱检索的现有工作。当目标域中仅有食谱时，CCFT采用一个域分类器，对抗性地对齐源域和目标域之间的食谱特征。此外，我们还给出了预言机模型的性能作为参考上界；该模型使用源域和目标域中的所有配对数据进行训练。

实现细节。 图像编码器的骨干网络 E_V 基于 ResNet-50 [14] ，并使用在 ImageNet [9]上预训练的权重进行初始化。ResNet-50 的最后一层被替换为一个输出维度为 1024 的全连接层。与图像特征相同，食谱特征的维度也设为 1024。词嵌入通过 word2vec [22] 获得，维度设为 300。与 [30, 41], 食谱编码器 E_R 类似，首先分别提取标题、食材和烹饪步骤的原始特征。标题和食材编码采用双向长短期记忆网络 [15], 而考虑到烹饪步骤文本较长，则使用层次长短期记忆网络处理。随后，将这三种特征拼接起来，并输入全连接层进行特征变换，以输出食谱特征。判别器D 是一个域分类器，由一个三层感知器构成，用于输出源域和目标域的概率。

所有实验均使用 Adam优化器 [18]进行训练，批次大小为 32。源子集池的大小根据迁移性能选取，详见第 4.4 节。初始学习率设为 0.0001，并持续训练直至模型收敛。公式 3 中的权衡超参数 β 和λ 分别设为 0.01 和 0.002。同时，公式 4 中的间隔设为 α=0.3。所有模型均使用 Pytorch [26] 框架实现。

4.2. 预训练源模型的有效性

由于预训练源模型在样本选择中发挥着关键作用，我们首先分析其有效性。图 3 展示了以源域食谱（粤菜）为查询，利用预训练源模型提取的食谱特征检索目标域（川菜）食谱的示例。请注意，食物图像仅用于可视化，并未参与检索。如果目标食谱中的食材也出现在查询食谱中（第 3_{rd} 行和第 6_{th} 行），则会加下划线并以红色突出显示。从第一个查询（第 2 至 4 行）可以看出，检索结果排名前五的食谱与查询食谱共享一些主要食材。例如，所有食谱中都有“米饭”，“美味南瓜饭”“蘑菇鸡肉焖饭”和“黄焖鸡米饭”中都有“蘑菇”，而“排骨焖饭”和“排骨焖饭”中都有“猪排骨”。与第二个查询（第 5 至 7 行）类似，前四个食谱都包含主要食材“五花肉”，而最后一个食谱“面筋烧白菜”则与查询一样含有“卷心菜”。此外，从视觉角度观察，由于烹饪和切割方法不同，不同域的食材，甚至同一域中不同菜肴的食材，在形状、质地和颜色上都存在较大差异。尽管如此，它们的高层语义仍然对齐，预训练源模型也能够挖掘其中的潜在信息。

如表1所示，Y→C 迁移的预训练源模型在 1K 和 5K 测试集上取得的中位数排名分别为 12.5 和 54.7。尽管模型并非完美，但根据上述分析，它确实为源域和目标域之间的相关性提供了有用的线索和证据。此外，由于模型参数没有更新，因此在训练过程中保持稳定。因此，我们认为该预训练源模型可靠且合理，可用于衡量源食谱与目标域食谱之间的相似度。

4.3 性能比较

定量结果。表1展示了所提方法与基线方法在中位数排名方面的性能比较。首先，所提方法在所有迁移任务中均优于所有基线方法。其次，不同迁移任务的改进幅度差异较大。与跨域食物图像到食谱检索领域的开创性工作 CCFT [41] 相比，在 C→W、Y→W 和 Y→C 上，中位数排名有了大幅提升；而在 C→Y 上，改进相对较小，在 1K 和 5K 测试规模下仅分别提升了 0.2 和 0.5 个排名。尽管如此，中位数排名平均提升了 8.8%，从 15.9 降至 14.5。第三，所提方法显著优于先进的基于加权的对抗方法 IWAN [37]。与仅源域模型相比，IWAN 仅取得了微弱的提升，这表明其在跨域食物图像到食谱检索中的适用性有限。

图4还展示了 R@K 的表现，并对比了仅源域、CCFT 和所提出的方法。仅使用未配对的食谱时，所提出的方法和 CCFT 的表现均显著优于仅源域模型，且提升幅度很大。同时，所提出的方法在所有迁移任务中、针对所有 R@K 指标的表现都优于 CCFT。值得注意的是，与 R@1 相比，R@10 和 R@50 的提升更为显著。换言之，排名提升主要出现在中间和靠后的位置，而非最前列。例如，在 Y→W 迁移任务中（图4（b）），R@10 和 R@50 分别

	源样本	预训练源模型检索到的前 5 个目标样本				
标题	排骨土豆炖饭	美味南瓜饭	蘑菇鸡焖饭	排骨焖饭	黄焖鸡肉米饭	排骨焖饭
食材	米饭；猪排骨；土豆；胡萝卜；蘑菇；豌豆……	南瓜；米饭；猪肉；蘑菇；虾……	鸡肉；土豆；洋葱；米饭；蘑菇；糖……	猪排骨；米饭；胡萝卜；香菇……	鸡肉；胡椒；土豆；大蒜；蘑菇；米饭……	青豆；香菇；米饭；猪排骨；胡萝卜；大蒜……
菜品图片						
标题	什锦五花肉泡菜	五花肉泡菜豆腐锅	麻辣火锅	铁板杏鲍菇	回锅肉配生姜	炖白菜配面筋
食材	五花肉；泡菜；卷心菜；豆腐；洋葱；豆芽；生姜……	五花肉；泡菜；豆腐；豆芽；金针菇；小白菜……	五花肉；洋葱；豆芽；金针菇；香肠；卷心菜……	杏鲍菇；猪肉；五花肉；大蒜；芹菜……	胡椒；五花肉；生姜；大蒜；豆瓣酱……	卷心菜；面筋；盐；葱……
菜品图片						

图3. 在 Y→C 迁移中，以源域食谱为查询，通过预训练源模型进行特征提取得到的排名前 5 的检索到的目标样本示例。源域食谱中出现的食材以红色下划线标出。

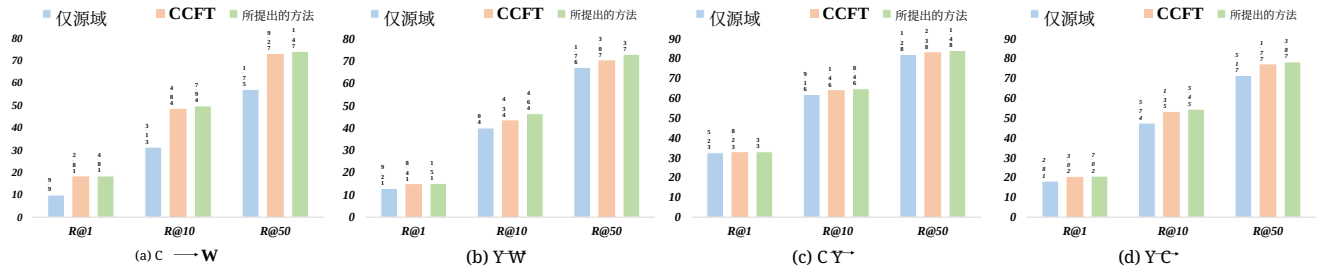

图4. 1K测试集上跨域食物图像到食谱检索性能的 R@K 对比。

提升了 3% 和 2.7%，而 R@1 仅提升了 0.3%。总而言之，在不利用配对信息的情况下，跨域食物图像到食谱检索中，对排名靠前的细粒度食谱进行区分的能力仍然有限。

尽管所提出的方法取得了改进，但与预言机模型相比仍存在明显差距。由于预言机模型是在全监督条件下使用来自源域和目标域的配对数据进行训练的，因此其性能是所有模型的上限。有趣的是，CCFT 与预言机模型之间的差距在不同迁移任务中差异很大，例如，C→W 中预言机模型与 CCFT 之间的差距为 6.8，而 C→Y 中的差距为 1.5。这是因为不同菜系的域偏移和训练数据量各不相同。同样，我们所提方法与 CCFT 之间的间隔也各不相同。

定性结果。图5展示了 Y→W 迁移任务中的两个代表性示例，其中所提出的方法……

证明其排名优于 CCFT。以和食食物图像作为查询，展示两种方法检索出的排名前 3 的食谱及其配对食物图像。真实标注（GT）食谱以边界框突出显示；如果对应食谱中的食材出现在 GT 食谱中，则会以红色下划线标出并突出显示。请注意，模型训练期间从未见过目标图像。在第一个示例中（第 3–5 行），两种方法都能将 GT 食谱排入前 3 名，所提出的方法和 CCFT 的排名分别为 1_{st} 和 3_{rd} 。检索出的前 3 个食谱都与 GT 食谱非常相似，仅在食材组成上有细微差异。与所提出的方法不同，CCFT 无法区分“日本小牛肉饭”和“日式牛肉饭”，因为它们都含有 GT 食谱中没有的“西兰花”。在第二个示例中（第 6–8 行），与 CCFT 相比，所提出的方法的排名提升了 11 位，从第 13 名升至第 2 名。CCFT 检索出的前 3 个食谱与 GT 食谱共享的食材数量明显

查询图像	所提方法				CCFT			
	排名前三的检索食谱及配对图像			真实值排名	排名前三的检索食谱及配对图像			真实值排名
	吉野家 照烧牛肉 牛肉；洋葱；胡萝卜； 黑胡椒…… 	日式牛肉饭 牛肉；洋葱；黄油； 味精…… 	日式小牛肉饭 牛肉；洋葱；胡萝卜； 西兰花；酱油 … 	1	日式小牛肉饭 牛肉；洋葱；胡萝卜； 西兰花；酱油 … 	日式牛肉 Rice 牛肉；洋葱；胡萝卜； 西兰花；鸡蛋；大豆 酱汁…… 	吉野家 照烧牛肉 牛肉；洋葱；胡萝卜； 黑胡椒…… 	3
	蔬菜沙拉 樱桃番茄； 欧芹；蓝莓； 紫甘蓝…… 	豆腐沙拉配 番茄、牛油果 和つゆ 豆腐；洋葱；つゆ； 欧芹；牛油果； 樱桃番茄…… 	花椰菜 米寿司 牛油果；欧芹； つゆ；花椰菜； 胡萝卜；生菜…… 	2	金枪鱼蔬菜 沙拉 樱桃番茄； 腌黄瓜；金枪鱼；橄榄； 生菜；胡椒…… 	清淡沙拉 樱桃番茄； 胡萝卜；生菜； 鸡肉；培根…… 	葡萄柚 牛肉沙拉 生油果；芝麻菜； 牛排；葡萄柚； 	13

图5. 通过 Y→W 迁移，展示了目标域中 CCFT [41] 与所提方法在图像到食谱检索性能上的对比示例。查询图像对应的真实配方以红色边界框标出；真实配方中出现的食材则以红色下划线标出。

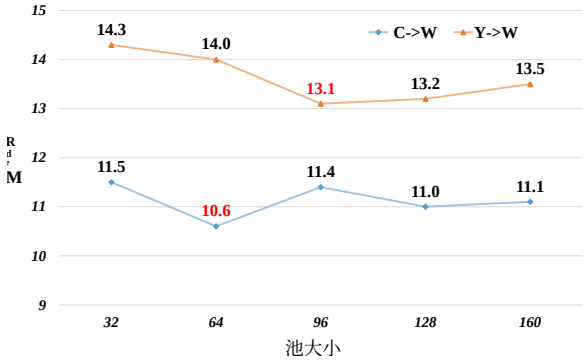

图6. C→W 和 Y→W 中，源子集池大小变化时的中位数排名表现。最优中位数排名以红色标出。

少于所提出的方法。尽管如此，所提出的方法检索出的排名第 1 的食谱“蔬菜沙拉”含有“樱桃番茄”和“欧芹”，排名仍比 GT 食谱靠前一位，这表明要区分食材组成细微差异的食谱颇具挑战性。

4.4. 参数选择与消融研究

参数选择。为研究源池大小对 SBS 的影响，将其设置为训练批次大小的 1 至 5 倍，即 {32, 64, 96, 128, 160}。请注意，当池大小等于 32 时，不采用 SBS。

C→W 和 Y→W 的中位数排名表现如图6所示。显然，池大小越大并不一定会带来更好的性能，不同菜系迁移的最佳池大小也可能不同。如图所示，C→W 和 Y→W 的最佳池大小分别为 64 和 96。此外，池大小对性能有显著影响。以 C→W 为例，如果将池大小设置为 96 而不是 64，中位数排名表现将从 10.6 降至 11.4，这几乎与不使用 SBS 时中位数排名为 11.5 的情况相同。

一般而言，目标样本选取相似源样本的概率与池大小成正比。如果池大小较小，SBS 的效果可能与随机采样没有区别。另一方面，当池大小较大时，同一组相似样本可能会被反复选中。因此，由于训练过程中反复使用相同或相似的源样本，学习效率会受到影响。因此我们推测，池大小的选择是在过滤训练中的无关源样本与保持样本多样性之间进行权衡。适中的池大小不仅能够选出合适的样本以弥合域偏移，还能增加训练过程中的扰动。

消融研究实验分别通过移除 SBS、加权三元组损失和加权对抗学习，测试了所提出模型的三个变体。如表1所示，与 CCFT [41]，相比，如果不使用 SBS，即

所提出的方法（去除 SBS）。因此，预先通过数据选择筛除具有显著特征的源样本，对于加权跨模态对抗学习至关重要。无论是加权三元组损失还是加权对抗损失，若省略其中的权重，所有迁移任务上的性能都会下降。例如，在 C→W 上，去除加权对抗损失的所提方法和去除加权三元组损失的方法的中位数排名分别下降 0.6 和 0.7；而在 C→Y 1K 测试中，这两种变体均仅下降 $0.\bar{1}$。两者相比，在仅采用其中一种时，加权三元组损失带来的增益略高于加权对抗损失。

5. 结论

我们提出了一种基于数据选择和重加权方案的跨域食物图像到食谱检索新方法，该方法在涵盖三种不同菜系的数据集上的所有迁移任务中，均优于现有方法。实证研究表明，使用预训练源模型作为特征提取器来计算源样本的权重是合理的。实验清楚地显示，所提方法能够持续带来性能提升。对于源池大小选择，在过滤掉具有区分性的样本与保持适当的样本多样性之间取得平衡至关重要。通过消融研究，我们发现，过滤数据以排除不相似的源样本，对于后续的加权对抗跨模态学习十分重要。此外，加权三元组损失和加权对抗损失均有助于提升性能。尽管结果令人鼓舞，所提方法在进一步提升排名靠前的检索食谱中细粒度食谱的表现方面仍存在局限，这将是我们未来的研究方向。

参考文献

[1] Yoshua Bengio、J´erˆome Louradour、Ronan Collobert 和 Ja- son Weston。课程学习。载于第*26*届机器学习年度国际会议论文集，第41–48页，2009年。[2] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——利用随机森林挖掘判别性组件。载于欧洲计算机视觉会议，第446–461页。施普林格，2014年。[3] Micael Carvalho、R´emi Cad`ene、David Picard、Laure Soulier、Nicolas Thome 和 Matthieu Cord。烹饪场景中的跨模态检索：学习语义文本-图像嵌入。载于第*41*届*ACM SIGIR*信息检索研究与发展国际会议，第35–44页，2018年。[4] Jingjing Chen 和 Chong-Wah Ngo。基于深度学习的食材识别，用于烹饪食谱检索。载于第*24*届*ACM*多媒体国际会议论文集，第32–41页，2016年。[5] Jingjing Chen、Bin Zhu、Chong-Wah Ngo、Tat-Seng Chua 和 Yu-Gang Jiang。关于多任务和区域级用于食材识别的深度学习。*IEEE*图像处理汇刊，30:1514–1526，2020年。[6] Jing-Jing Chen、Chong-Wah Ngo、Fu-Li Feng 和 Tat-Seng Chua。深入理解烹饪过程，以实现跨模态食谱检索。载于第*26*届*ACM*多媒体国际会议论文集，第1020–1028页，2018年。[7] Qingchao Chen、Yang Liu、Zhaowen Wang、Ian Wassell 和 Kevin Chetty。用于无监督领域自适应的重加权对抗适应网络。载于*IEEE*计算机视觉与模式识别会议论文集，第7976–7985页，2018年。[8] Shuhao Cui、Shuhui Wang、Junbao Zhuo、Chi Su、Qingming Huang 和 Qi Tian。用于对抗领域自适应的渐隐桥。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12455–12464页，2020年。[9] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层次化图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。[10] Han Fu、Rui Wu、Chenghao Liu 和 Jianling Sun。MCEN：利用潜变量模型弥合烹饪食谱与菜品图片之间的跨模态差异。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14570–14580页，2020年。[11] Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pascal Germain、Hugo Larochelle、Franc¸ois Laviolette、Mario Marchand 和 Victor Lempitsky。神经网络的领域对抗训练。机器学习研究杂志，17(1):2096–2030，2016年。[12] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。载于神经信息处理系统进展，第2672 2680页，2014 – 年。[13] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Girshick。Mask R-CNN。载于*IEEE*国际计算机视觉会议论文集，第2961–2969页，2017年。[14] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[15] Sepp Hochreiter 和 J¨urgen Schmidhuber。长短期记忆网络。神经计算，9(8):1735–1780，1997年。[16] Jie Hu、Li Shen 和 Gang Sun。压缩与激励网络。载于*IEEE*计算机视觉与模式识别会议论文集，第7132–7141页，2018年。[17] Xin Huang、Yuxin Peng 和 Mingkuan Yuan。MHTN：用于跨模态检索的模态对抗混合迁移网络。*IEEE*控制论汇刊，2018年。[18] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于第*3*届国际学习表征会议（*ICLR 2015*），2015年。[19] Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton。深度学习。《自然》，521(7553):436–444，2015年。[20] Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。载于

*IEEE*计算机视觉与模式识别会议论文集，第3431–3440页，2015年。[21] Mingsheng Long、Yue Cao、Jianmin Wang 和 Michael Jordan。通过深度适应网络学习可迁移特征。载于国际机器学习会议，第97–105页。机器学习研究论文集，2015年。[22] Tomas Mikolov、Kai Chen、Greg Corrado 和 Jeffrey Dean。在向量空间中高效估计词表示。载于第*1*届国际学习表征会议（*ICLR 2013*），2013年。[23] Weiqing Min、Linhu Liu、Zhengdong Luo 和 Shuqiang Jiang。用于食物识别的食材引导级联多重注意力网络。载于第*27*届*ACM*国际多媒体会议论文集，第1331 1339页，2019 – 年。[24] Sinno Jialin Pan 和 Qiang Yang。迁移学习综述。*IEEE*知识与数据工程汇刊，22(10):1345–1359，2009年。[25] Dim P Papadopoulos、Enrique Mora、Nadiia Chepurko、Kuan Wei Huang、Ferda Ofli 和 Antonio Torralba。学习食物图像和烹饪食谱的程序表示。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第16559–16569页，2022年。[26] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Ze ming Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式、高性能深度学习库。载于神经信息处理系统进展，第8026–8037页，2019年。[27] Fan Qi、Xiaoshan Yang 和 Changsheng Xu。多模态领域自适应的统一框架。载于第*26*届*ACM*国际多媒体会议论文集，第429–437页，2018年。[28] Joseph Redmon、Santosh Divvala、Ross Girshick 和 Ali Farhadi。YOLO（你只看一次）：统一的实时目标检测。载于*IEEE*计算机视觉与模式识别会议论文集，第779–788页，2016年。[29] Amaia Salvador、Erhan Gundogdu、Loris Bazzani 和 Michael Donoser。利用层次化Transformer和自监督学习改进跨模态食谱检索。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15475–15484页，2021年。[30] Amaia Salvador、Nicholas Hynes、Yusuf Aytar、Javier Marin、Ferda Ofli、Ingmar Weber 和 Antonio Torralba。学习烹饪食谱与食物图像的跨模态嵌入。载于*IEEE*计算机视觉与模式识别会议论文集，第3020–3028页，2017年。[31] Eric Tzeng、Judy Hoffman、Kate Saenko 和 Trevor Darrell。对抗判别式领域自适应。载于*IEEE*计算机视觉与模式识别会议论文集，第7167–7176页，2017年。[32] Eric Tzeng、Judy Hoffman、Ning Zhang、Kate Saenko 和 Trevor Darrell。深度领域混淆：最大化领域不变性。*arXiv*预印本 *arXiv:1412.3474*，2014年。[33] Hao Wang、Doyen Sahoo、Chenghao Liu、Ee-peng Lim 和 Steven CH Hoi。利用对抗网络学习烹饪食谱与食物图像的跨模态嵌入。载于*IEEE*计算机视觉与模式识别会议论文集，第 11572–11581 页，2019年。[34] Sinan Wang、Xinyang Chen、Yunbo Wang、Mingsheng Long 和 Jianmin Wang。用于细粒度领域自适应的渐进式对抗网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 9213–9222 页，2020年。[35] Zhongwei Xie、Lin Li、Luo Zhong、Jianquan Liu 和 Ling Liu。事件密集型文本与图像之间的跨模态检索。载于*2022*年国际多媒体检索会议论文集，第 229–238 页，2022年。[36] Hongliang Yan、Yukang Ding、Peihua Li、Qilong Wang、Yong Xu 和 Wangmeng Zuo。注意类别权重偏差：用于无监督领域自适应的加权最大均值差异。载于*IEEE*计算机视觉与模式识别会议论文集，第 2272–2281 页，2017年。[37] Jing Zhang、Zewei Ding、Wanqing Li 和 Philip Ogunbona。用于部分领域自适应的重要性加权对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 8156–8164 页，2018年。[38] Bin Zhu、Chong-Wah Ngo 和 Wing-Kwong Chan。基于网络食谱-图像对进行食物识别：问题、基线方法与性能。*IEEE*多媒体汇刊，24：1175–1185，2021年。[39] Bin Zhu、Chong-Wah Ngo、Ji ngjing Chen 和 Wing-Kwong Chan。结合混合增强的食谱检索跨语言自适应。载于*2022*年国际多媒体检索会议论文集，第 258–267 页，2022年。[40] Bin Zhu、Chong-Wah Ngo、Jingjing Chen 和 Yanbin Hao。R2gan：基于生成对抗网络的跨模态食谱检索。载于*IEEE*计算机视觉与模式识别会议论文集，第 11477–11486 页，2019年。[41] Bin Zhu、Chong-Wah Ngo 和 Jing-jing Chen。跨领域跨模态食物迁移。载于第*28*届*ACM*国际多媒体会议论文集，第 3762–3770 页，2020年。