计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(1):141−151 [doi: 10.15888/j.cnki.csa.010087] [CSTR: 32024.14.csa.010087]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CLIP 的无监督域适应图像分类①

丁华玲, 杨　欢

(华南师范大学 人工智能学院, 广州 510631)
通信作者: 杨　欢, E-mail: 1312324353@qq.com

摘　要: 无监督域适应 (unsupervised domain adaptation, UDA) 旨在将源域中训练好的模型应用于仅有未标记数据的目标域. 当前的无监督域适应方法主要通过统计差异最小化或对抗学习来对齐源域和目标域特征空间, 从而学习域不变特征. 然而, 这些约束可能导致语义特征结构的扭曲和类可辨别性的丧失. 针对上述问题, 本文提出一种名为 DAMPL 的方法. 该方法利用 CLIP 模型注入文本描述信息, 深入挖掘图像语义内容, 采用针对领域特性的提示学习范式, 有效保留不同域的特有信息, 避免了信息丢失. 此外, 通过语义引导机制对目标域的伪标签进行校正, 以缩小域间差异, 增强模型的泛化能力. 最后还引入互信息最大化损失 (mutual information maximization loss, IML), 以保留目标域的特征可区分性. 最终 DAMPL 方法在 Office-Home、miniDomainNet 和 VisDA-2017 数据集上分别达到 83.8%、79.7%、89.8% 的分类准确率, 展现了最佳的性能.

关键词: 无监督域适应; CLIP 模型; 提示学习; 伪标签; 互信息最大化损失

引用格式: 丁华玲,杨欢.基于 CLIP 的无监督域适应图像分类.计算机系统应用,2026,35(1):141–151. http://www.c-s-a.org.cn/1003-3254/10087.html

Unsupervised Domain Adaptation Image Classification Based on CLIP

DING Hua-Ling, YANG Huan

(School of Artificial Intelligence, South China Normal University, Guangzhou 510631, China)

Abstract: Unsupervised domain adaptation (UDA) aims to apply a trained model in the source domain to the target domain with only unlabeled data. Current UDA approaches learn domain-invariant features by aligning the source domain and target domain feature spaces via statistical difference minimization or adversarial learning. However, these constraints may result in the distortion of semantic feature structures and loss of class discriminability. To this end, this study proposes a new method called DAMPL. This method utilizes the CLIP model to inject textual descriptive information to deeply mine the semantic content of the image, and adopts a prompt learning paradigm for domain characteristics to effectively retain information specific to different domains, thus avoiding information loss. Additionally, the pseudo-labelling of the target domains are corrected via a semantic bootstrapping mechanism to reduce the inter-domain differences and enhance the generalization ability of the model. Finally, mutual information maximization loss (IML) is also introduced to preserve the feature distinguishability of the target domains. The final DAMPL method demonstrates optimal performance by achieving 83.8%, 79.7%, and 89.8% classification accuracy on the Office-Home, miniDomainNet, and VisDA-2017 datasets, respectively.

Key words: unsupervised domain adaptation (UDA); CLIP model; prompt learning; pseudo-labelling; mutual information maximization loss (IML)

① 收稿时间: 2025-05-16; 修改时间: 2025-07-10, 2025-09-19; 采用时间: 2025-10-09; csa 在线出版时间: 2025-12-01
CNKI 网络首发时间: 2025-12-02

近年来, 随着深度学习技术的飞速发展, 各种计算机视觉任务如图像分类、目标检测等均取得了令人瞩目的成果. 然而, 这些成就的背后, 我们仍面临着数据标注成本高昂、标注质量参差不齐以及模型泛化能力不足等问题. 为了解决上述问题, 无监督域适应 (unsupervised domain adaptation, UDA)[1,2]应运而生, 它旨在减少对标注数据的依赖, 通过学习在不同域之间的特征表示, 使得能够在没有或仅有少量标注的目标域上实现有效的知识迁移.

当前的方法主要是通过最小化统计差异[3−5]或对抗性训练[6,7]来学习域不变的特征表示, 以实现源域与目标域之间的分布偏移最小化. 这种方法使得在源域上训练的分类器能够直接应用于目标域数据. 然而, 这种域对齐过程往往伴随着语义信息的丢失, 例如, 特征表示可能因与灰度图像对齐而失去颜色信息, 进而损害了类别的可分辨性, 因为领域特定知识在分类中扮演着重要角色.

与此同时, 大型视觉语言模型 (vision-language model, VLM) 的崛起为解决无监督域适应中的这一难题提供了新的路径. 以 CLIP (contrastive language-image pre-training)[8]为例, 这一可扩展的对比预训练模型通过联合学习图像和文本特征, 利用庞大的图像-文本对数据集, 展现了其显著的语义丰富性和零样本泛化能力.

受到这一发现的启发, 我们提出了一种改进的 UDA 解决方案, 即 DAMPL (基于 CLIP 的无监督领域自适应方法). 首先, 通过引入 CLIP 模型, 为纯视觉数据注入了丰富的文本描述信息, 这有助于模型更深入地挖掘图像背后的语义内容, 从而在分类任务中实现更精准的识别. 同时使用了一套针对不同领域特性的提示学习机制, 这种机制能够有效地适应和保留不同域之间特有的信息, 避免了传统方法中因域对齐而造成的信息丢失问题, 不仅提高了类别的可分辨性, 还确保了模型的适应性和泛化能力. 为了进一步提高伪标签的可靠性, 我们采用了基于语义引导的伪标签调整策略, 这种方法通过利用语义信息来指导伪标签的生成, 从而使得标签更加准确, 减少了错误标签对模型训练的负面影响. 此外, 最后还引入了互信息最大化损失, 通过增强源域和目标域特征表示的共享性和域不变性, 保留关键语义信息, 并减少过拟合风险.

综上所述, 本研究的主要贡献如下.

(1) 提出了 DAMPL 方法, 通过结合 CLIP 模型, 成功地为视觉数据注入了丰富的文本描述信息, 增强了模型对图像语义内容的理解, 从而提高分类任务的识别精度.

(2) 引入了一套针对不同领域特性的提示学习机制, 该机制有效适应并保留了域间特有信息, 避免了传统域对齐方法中的信息丢失问题, 同时提升了类别的可分辨性和模型的泛化能力.

(3) 使用了基于语言引导的伪标签调整策略, 通过自训练生成更加精确和可靠的伪标签, 减少了错误标签对模型训练的影响. 此外, 还集成了互信息最大化损失到训练过程中, 通过增强源域和目标域特征表示的共享性和域不变性, 有效保留了关键语义信息, 并降低了过拟合风险, 双域协同优化提高了模型在无监督域适应任务中的鲁棒性和适应性.

(4) 最终 DAMPL 方法在 Office-Home、 miniDomainNet 和 VisDA-2017 数据集上分别达到 83.8%、79.7%、89.8% 的分类准确率, 展现了出色性能.

1 相关工作

1.1 无监督域适应

无监督域适应旨在解决源域与目标域之间的分布差异问题, 使得在源域上训练的模型能够适应未标记的目标域. 无监督域适应的发展经历了从传统统计匹配[9−11]到深度对抗学习[12−14]、自监督学习与大模型结合[15−18]的演进.

早期工作通过浅层模型 (如 SVM、核方法等) 匹配源域和目标域的统计特征. Pan 等人[9]提出的 TCA 通过核方法学习跨域共享子空间. Sun 等人[10]提出的 CORAL 通过协方差对齐源域与目标域特征分布, 开创了浅层特征适配的先河. 后续深度学习方法, 如 Long 等人[11]首次将最大均值差异 (MMD) 度量引入深度网络, 通过多核 MMD 对齐网络各层次的特征分布. 然而, 这类方法仅能捕捉低阶统计特征, 难以处理复杂非线性分布偏移.

受生成对抗网络 (GAN)[12]的启发, Ganin 等人[2]提出了基于对抗训练的域适应框架. 该方法通过引入梯度反转层 (GRL) 实现特征提取器与域判别器的对抗训练, 从而学习域不变特征. 这一工作奠定了对抗训练在域适应领域的基础地位, 但其存在两个关键问题: 其一, 对抗优化过程易陷入局部最优, 在 VisDA-2017 等复杂数据集上收敛困难; 其二, 域判别器的二分类任务难以

捕捉细粒度的域间关系. 为了解决上述问题, Long 等人[13]提出了条件对抗域适应方法, 该方法将类别信息融入对抗过程, 通过类感知特征对齐实现了更精细的分布匹配, 提升语义一致性. 然而, 条件对抗方法仍然存在语义保持不足的问题. Liu 等人[14]指出, 过度追求域不变性会导致原始特征判别性下降, 在某些场景下甚至引发“负迁移”现象.

随着大模型时代的到来, 无监督域适应研究也呈现新趋势. 2017 年, Vaswani 等人[15]首次提出 Transformer 架构, 为视觉大模型奠定理论基础. 2018 年, 自监督学习范式 (如 MoCo[16]) 兴起, 推动模型参数突破 1 亿规模. Dosovitskiy 等人[17]提出的 ViT-Large 模型以 87.1% 准确率超越 CNN. Swin Transformer[18]提出层次化窗口注意力, 在 COCO 目标检测任务上 mAP 达到 58.7%.

上述方法虽然在许多场景下表现良好, 但其在纹理变化剧烈的复杂场景下往往出现性能显著下降. 这一局限性主要源于模态单一缺陷. 为此, 研究者们开始将目光转向多模态融合和文本信息利用等新兴方向.

1.2 CLIP 在无监督域适应中的创新应用

CLIP 模型由 OpenAI 在 2021 年提出, 模型结构如图 1 所示, 是一种基于文本-图像对比学习的预训练模型. 这类模型的核心优势在于通过结合场景的语义文本描述, 可以构建更具判别性的特征表示, 不仅能够弥补单一视觉特征的不足, 还能通过语义信息的引入实现对场景更深入的理解, 从而显著提升算法在复杂环境下的适应性.

CLIP 在无监督域适应 (UDA) 中的应用主要分为两类: 其一, 冻结 CLIP 图像编码器, 仅微调适配层; 其二, 通过文本提示生成伪标签. 早期研究直接采用手工设计的文本提示 (如“a photo of [CLASS]”) 进行零样本推理. 解决了快速部署问题, 避免昂贵的微调过程, 但仍存在提示敏感性高, 性能波动等问题. 针对上述问题, CoOp[19]首次提出可学习的上下文令牌, 将人工设计提示中的固定词汇替换为可优化向量, 解决了提示设计的主观性和不稳定性, 但仍存在学习到的上下文缺乏领域适应性的问题. CoCoOp[20]在 CoOp 基础上引入两层网络生成实例相关的动态提示, 使得在细粒度分类任务上 mAP 提升 9.7%, 但导致计算成本增加 30% (需额外前向传播). DAPL[21]针对域适应问题首次引入领域特定的可学习令牌, 轻量化的同时有效提升了分类性能.

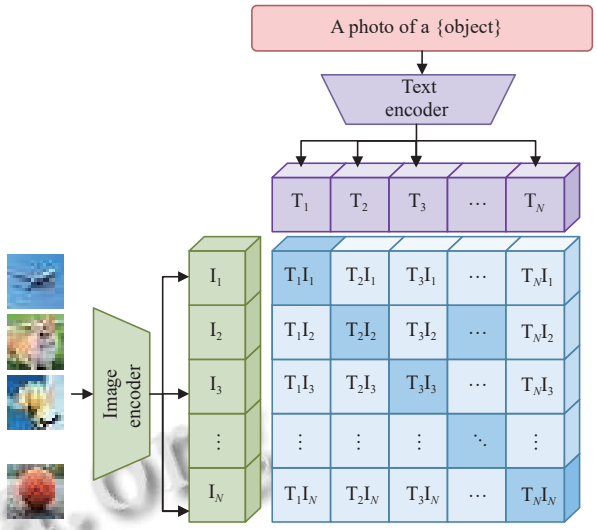

图 1 CLIP 模型结构图

尽管上述方法取得显著进展, 但仍受到文本模态主导预测, 视觉特征利用率不足, 且伪标签噪声累积等问题的影响. 如何更好地同步双域知识、协同优化成为问题关键.

1.3 伪标签技术

伪标签 (pseudo-labelling, PL) 技术通过将未标记数据的高置信度预测结果作为监督信号, 已成为无监督域适应的核心方法之一. 该技术最早由 Lee[22]在半监督学习中提出, 后被广泛应用于域适应领域, 形成两类典型应用范式.

第 1 类方法侧重于预测一致性约束[23,24]. 如 Tarvainen 等人[23]提出的 Mean Teacher 框架, 通过教师模型生成目标域的伪标签, 强制学生模型保持预测一致性. Xie 等人[24]进一步引入 MixMatch 策略, 在数据增强空间实施伪标签一致性正则化. 这类方法可使分类准确率显著提升, 但其有效性依赖于强数据增强策略的设计.

第 2 类方法聚焦于特征分布对齐[25,26]. Zhang 等人[25]将伪标签集成至对抗训练框架, 通过判别器引导特征空间的对齐. Chen 等人[26]提出动态伪标签加权机制, 在特征适配过程中实现类条件分布匹配.

然而, 现有方法普遍存在两个关键局限: 其一, 隐含假设源域与目标域的条件分布相似性, 这在现实世界中可能不成立; 其二, 传统伪标签仅依赖视觉模态的置信度阈值, 忽略了跨模态语义一致性验证.

2 方法

2.1 模型架构

针对上述问题, 本文提出的方法整体架构如图 2 所

示, 包括一个特征提取器$\mathcal{F}(\cdot)$和一个分类器$\mathcal{G}(\cdot)$, $\mathcal{E}_{\mathcal{I}}$和$\mathcal{E}_{\mathcal{T}}$为引入的预训练模型 CLIP 的图像编码器和文本编码器. 给定一组有标签的源域样本$D_s=\{(x_i^s,y_i^s)\}_{i=1}^{N_s}$, 一组没有标签的目标域样本$D_t=\{(x_i^t)\}_{i=1}^{N_t}$. 首先, 我们引入了特定领域的提示学习范式将输入数据的标签与提示模板结合生成相应的文本描述, 再利用 CLIP 模型进行零样本推理, 生成相应的伪标签. 考虑到生成的伪标签可能包含噪声和错误等问题, 采用了一种语义引导的伪标签校正策略, 从而提高伪标签的准确性和泛化性能. 此外, 我们还引入了互信息最大化损失, 通过增强源域和目标域特征表示的共享性和域不变性, 保留关键语义信息, 并减少过拟合风险.

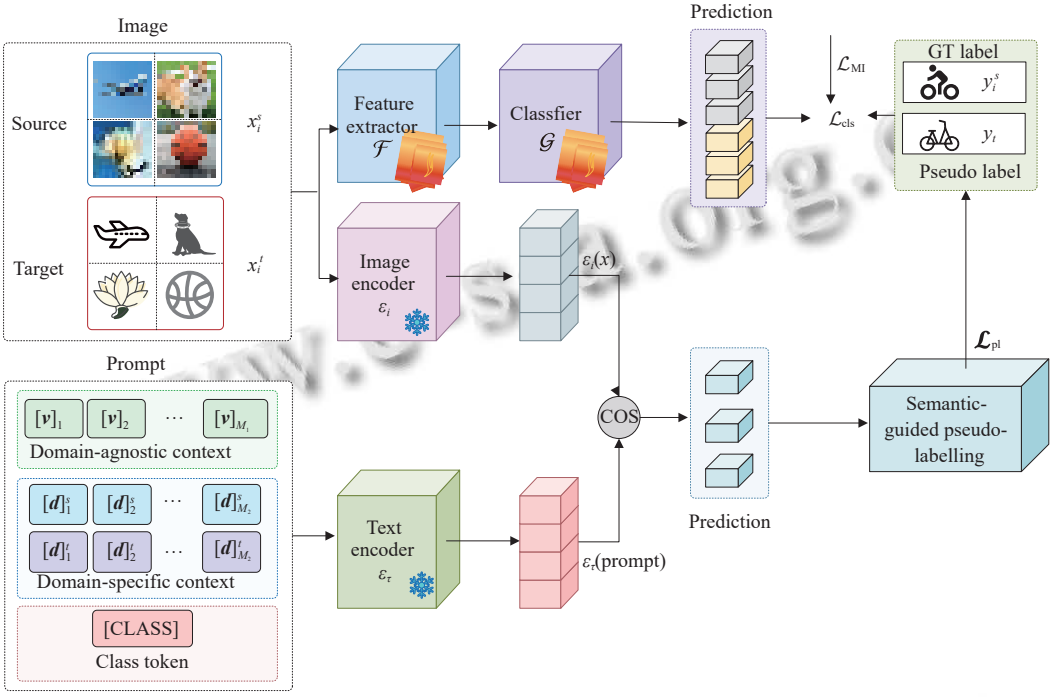

图 2 DAMPL 模型架构图

2.2 特定域文本提示范式

现有的提示学习方法如图 3(a) 所示, 即上下文在所有领域和所有类别之间共享. 它有一个统一的格式:

$$\boldsymbol{t}_k=[\boldsymbol{v}]_1[\boldsymbol{v}]_2\cdots[\boldsymbol{v}]_{M_1}[\mathrm{CLASS}]_k \tag{1}$$

其中, $[\boldsymbol{v}]_{m_1}$, $m_1\in\{1,2,\cdots,M_1\}$, 是与嵌入词具有相同维数的向量, M_1 是在提示符中应用的上下文令牌的数量.

然而考虑到源域和目标域的类标签在语义上可能是不同的, 领域不可知上下文忽略了领域之间的差异等问题, 我们采用融合特定领域的提示学习范式来处理分布移位[21], 如图 3(b) 所示.

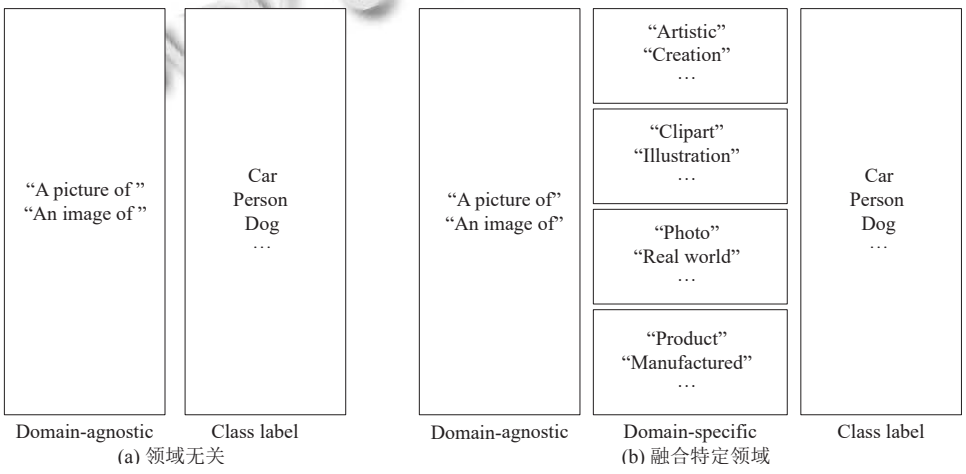

图 3 提示学习范式的比较

融合特定领域的提示学习范式主要包含 3 个部分: 与领域无关的令牌、特定于领域的令牌和类令牌. 使用$[\boldsymbol{d}]^d_{m_2}$, $m_2\in\{1, 2, \cdots, M_2\}$, 用于表示与词嵌入具有相同维度的特定于领域的令牌, 特定领域的令牌在所有类别之间共享, $[\boldsymbol{d}]^s_i \neq [\boldsymbol{d}]^t_j$, $i, j\in\{1, 2, \cdots, M_2\}$, 特定领域的令牌的数量用 M_2 表示, 域指示器 $d\in\{s, t\}$表示源域和目标域. 整个提示符定义为以下格式:

$$\boldsymbol{t}^d_k = [\boldsymbol{v}]_1[\boldsymbol{v}]_2\cdots[\boldsymbol{v}]_{M_1}[\boldsymbol{d}]^d_1[\boldsymbol{d}]^d_2\cdots[\boldsymbol{d}]^d_{M_2}[\text{CLASS}]_k \quad (2)$$

此外, 一个可训练的类感知提示可以学习细粒度的类别表示. 与领域无关的上下文可以遵循由类特定上下文表示的类特定样式. 每个类都可以用不同的令牌初始化:

$$\boldsymbol{t}^d_k = [\boldsymbol{v}]^k_1[\boldsymbol{v}]^k_2\cdots[\boldsymbol{v}]^k_{M_1}[\boldsymbol{d}]^d_1[\boldsymbol{d}]^d_2\cdots[\boldsymbol{d}]^d_{M_2}[\text{CLASS}]_k \quad (3)$$

当类别和域分别匹配时, 图像和提示符才会形成正对, 使用来自相应域的正对更新特定领域的令牌, 使得特定领域的令牌嵌入每个域中共享的信息. 随机初始化源域和目标域的可训练域不可知上下文$[\boldsymbol{v}]^k_i$和特定域上下文$[\boldsymbol{d}]^s_j$、$[\boldsymbol{d}]^t_j$. 类令牌[CLASS]由数据集的类名提供, 分别对源域和目标域应用了不同的提示t^s_k和t^t_k, 共有 $2K$ 个类别. 给定一组训练样本$\{(x^s_i, y^s_i)\}^{N_s}_{i=1}$, 不含标签的样本$\{(x^t_i)\}^{N_t}_{i=1}$, 我们可以根据式 (4)、式 (5) 得到样本属于第 k 类的概率:

$$P\left(\hat{y}^s_i = k \mid x^s_i, t^s_k\right) = \frac{\exp\left(\left\langle \mathcal{E}_{\mathcal{T}}\left(t^s_k\right), \mathcal{E}_{\mathcal{I}}\left(x^s_i\right)\right\rangle / T\right)}{\sum_{d\in\{s,t\}}\sum_{j=1}^{K}\exp\left(\left\langle \mathcal{E}_{\mathcal{T}}\left(t^d_j\right), \mathcal{E}_{\mathcal{I}}\left(x^s_i\right)\right\rangle / T\right)} \quad (4)$$

$$P\left(\hat{y}^t_i = k \mid x^t_i, t^t_k\right) = \frac{\exp\left(\left\langle \mathcal{E}_{\mathcal{T}}\left(t^t_k\right), \mathcal{E}_{\mathcal{I}}\left(x^t_i\right)\right\rangle / T\right)}{\sum_{d\in\{s,t\}}\sum_{j=1}^{K}\exp\left(\left\langle \mathcal{E}_{\mathcal{T}}\left(t^d_j\right), \mathcal{E}_{\mathcal{I}}\left(x^t_i\right)\right\rangle / T\right)} \quad (5)$$

由于已知图像来自源域或目标域, 在给定真值标签y^s_i 的情况下, 利用图像x_i属于 k 类的概率最小化交叉熵损失, 源域的交叉熵损失如下所示:

$$\mathcal{L}_s = -\frac{1}{N_s}\sum_{i=1}^{N_s}\log P\left(\hat{y}^s_i = y^s_i\right) \quad (6)$$

2.3 基于语义引导的伪标签校正策略

2.3.1 初始伪标签生成

为了进一步利用未标记的数据, 我们在目标域上利用 CLIP 的零样本预测能力生成对应的初始伪标签. 首先通过视觉编码器 $\mathrm{E_I}$ 生成目标域样本的视觉特征表示$\mathcal{E}_{\mathcal{I}}(x^t)$. 同时, 将特定于领域的提示模板输入到文本编码器 $\mathrm{E_T}$ 中, 以获得文本特征表示$\mathcal{E}_{\mathcal{T}}\left(t^t_k\right)$. 然后使用式 (5) 计算目标域数据的概率$P\left(\hat{y}^t_i = k \mid x^t_i, t^t_k\right)$, 从 K 个预测概率最大的类中选择$\hat{y}^t$作为训练数据x^t的初始伪标签:

$$\hat{y}^t = \arg\max_k P\left(\hat{y}^t = k \mid x^t\right),\ k = \{1, 2, \cdots, K\} \quad (7)$$

损失函数如式 (8) 所示, 我们仅为最大预测概率大于伪标签的固定阈值 τ 的未标记数据生成伪标签, 使用这些无标签图像及其伪标签, 通过对比学习来训练目标域$\boldsymbol{t}^t_k$的提示, 其中$\mathbb{I}\{\cdot\}$为指示函数.

$$\mathcal{L}_t = -\frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{I}\left\{P\left(\hat{y}^t_i = \hat{y}^t_i \mid x^t_i\right) \geqslant \tau\right\}\log P\left(\hat{y}^t_i = y^t_i \mid x^t_i\right) \quad (8)$$

2.3.2 语义引导的校正优化

由于源域和目标域之间的域偏差, 可能会导致生成不可靠的伪标签, 这种不可靠性反过来又会导致次优分类性能, 并对无监督领域自适应模型的自训练泛化能力产生不利影响. 为了提高伪标签的准确性和模型的性能, 我们引入了一种新的伪标签校正策略: 基于语义引导的伪标签. 该策略旨在改进目标域伪标签, 随后利用这些改进的目标域伪标签以监督的方式重新训练模型. 我们通过特征提取器$\mathcal{F}(\cdot)$提取目标域特征$\mathcal{F}(x^t)$来计算目标域中每个类的质心c^t_k:

$$c^t_k = \frac{\sum_{x^t\sim p^t}\left(\mathcal{G}\left(\mathcal{F}\left(x^t\right)\right) + P\left(\hat{y}^t_i = k \mid x^t_i, t^t_k\right)\right)\mathcal{F}\left(x^t\right)}{\sum_{x^t\sim D^t}\left(\mathcal{G}\left(\mathcal{F}\left(x^t\right)\right) + P\left(\hat{y}^t_i = k \mid x^t_i, t^t_k\right)\right)} \quad (9)$$

其中, $\mathcal{G}(\mathcal{F}(x^t))$为来自分类器 G 的目标域样本x^t属于类 k 的初始概率, $P\left(\hat{y}^t_i = k \mid x^t_i, t^t_k\right)$ 表示来自 CLIP 的目标域图像x^t属于类 k 的概率. 利用目标域特定提示t^t_k获得的输出概率来校准质心, 并提高质心对目标数据的可靠性, 然后通过最新的质心分类器为目标域数据x^t分配伪标签:

$$\hat{y}^t = \arg\min_k\left(\mathcal{F}\left(x^t\right), c^t_k\right) \quad (10)$$

最后, 我们利用新分配的伪标签计算目标质心为:

$$c^{t'}_k = \frac{\sum_{x^t\sim D^t}\mathbb{I}\left(\hat{y}^t = k\right)\mathcal{F}\left(x^t\right)}{\sum_{x^t\sim D^t}\mathbb{I}\left(\hat{y}^t = k\right)} \quad (11)$$

$$\hat{y}^t = \arg\min_k \left(\mathcal{F}\left(x^t\right), c_k^{t'}\right)$$

其中, $\hat{y}^t$ 为被修正后的目标域伪标签, 式 (11) 可以迭代更新为多轮, 此处只迭代更新了一轮, 损失函数定义为:

$$\mathcal{L}_{\text{pl}} = \mathbb{E}_{(x^t,\hat{y}^t)\sim D^t} \ell\left(\mathcal{G}\left(\mathcal{F}\left(x^t\right)\right), \hat{y}^t\right) \tag{12}$$

其中, ℓ 是交叉熵损失.

2.4 互信息最大化损失

最后, 为了减少源域和目标域之间的分布差异, 促使模型学习到域不变的特征表示, 保留关键语义的特征, 我们通过最大化源域预测 $P\left(\hat{y}_i^s = k \mid x_i^s, t_k^s\right)$ 和目标域预测 $P\left(\hat{y}_i^t = k \mid x_i^t, t_k^t\right)$ 之间的无偏互信息来同步双方的知识, 从而增强模型在目标域上的泛化能力, 损失函数如下:

$$\mathcal{L}_{\text{MI}} = -\frac{1}{|X_t|} \sum_{X_i \in X_t} I\left(P\left(\hat{y}_i^s = k \mid x_i^s, t_k^s\right), P\left(\hat{y}_i^t = k \mid x_i^t, t_k^t\right)\right) \tag{13}$$

其中, $I(\cdot,\cdot)$ 为计算互信息[27]. 因此, 总的损失函数为:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_t + \alpha\mathcal{L}_{\text{pl}} + \beta\mathcal{L}_{\text{MI}} \tag{14}$$

其中, α 和 β 是平衡参数.

3 实验与结果分析

3.1 实验数据

本文提出的 DAMPL 方法通过在 3 个数据集上进行广泛的实验来评估其有效性, 分别是: Office-Home[28]、miniDomainNet[29]和 VisDA-2017[30]. Office-Home 数据集包含 4 个域, 分别为 Artistic (A)、Clipart (C) 、Product (P) 和 Real-World (R), 每个域包括来自 65 个类别的图像, 总共产生大约 15 500 张图像. miniDomainNet 数据集是 DomainNet 数据集的子集, 包含了 4 个域, 分别为: 包含 18 703 张图像的 clipart (c)、31 202 张图像的 painting (p)、65 609 张图像的 real world (r)、24 492 张图像的 sketch (s), 每个域包含 126 个类别. VisDA-2017 数据集具有 12 个类别, 包含 152 397 张合成图像和 55 388 张真实图像, 用于具有挑战性的合成到真实的领域自适应, 如图 4 所示.

图 4 VisDA-2017 数据集的示例图像

3.2 实验设置

在所有实验中, 我们使用 ResNet 作为批次大小为 32 的特征提取器, ResNet-50 模型作为除 VisDA-2017 的 ResNet-101 之外的所有数据集的主干. 在 Office-Home 上进行了 200 个 epoch 的训练, 在 VisDA-2017 上进行了 25 个 epoch 的训练, 在 miniDomainNet 上进行了 15 个 epoch 的训练, 其中批大小设置为 32. 采用小批量随机梯度下降优化器, 学习率初始化为 0.003, 并使用余弦退火规则进行衰减. 值得注意的是, 在训练期间, CLIP 中的视觉编码器和文本编码器是冻结的. 至于超参数部分, 本文参考文献[21]中的标准实验设置, 上下文令牌 M_1 和特定领域的令牌 M_2 的长度都设置为 16, 上下文向量使用标准偏差为 0.02 的零均值高斯分布随机初始化. 对于 Office-Home, 伪标签阈值 τ 设置为 0.6, VisDA-2017 设置为 0.5, miniDomainNet 设置为 0.7. 平衡参数 α 和 β 的初始值分别设置为 1.0、1.0.

3.3 评价指标

本文采用准确率 (accuracy) 和任务平均准确率 (average accuracy) 作为指标来评估模型的域适应性能.

(1) 准确率 (accuracy): 该指标衡量模型在单个特定迁移任务上的性能, 计算公式为:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right) \tag{15}$$

其中, N 为目标域测试集的样本总数, $\hat{y}_i$ 为模型预测标签, y_i 为真实标签, $\mathbb{I}\{\cdot\}$ 为指示函数.

(2) 任务平均准确率 (average accuracy): 该指标用于评估模型在所有迁移任务上的综合性能与稳健性,

其值为所有迁移任务准确率的算术平均值:

$$Avg = \frac{1}{T}\sum_{t=1}^{T} Acc_t \tag{16}$$

其中, T 为迁移任务的总数, Acc_t 为模型在第 t 个任务上的准确率.

3.4 实验结果分析

3.4.1 对比实验

在本研究中, 我们评估了 DAMPL 及其他 SOTA 域适应方法在 Office-Home、miniDomainNet 以及 VisDA-2017 数据集上的性能, 通过比较各方法的准确率 (*Acc*) 和平均准确率 (*Avg*) 来验证其有效性. 其中基线模型 CLIP 伪标签通过零样本 CLIP 模型生成, 采用人工设计的提示模板“a photo of [CLASS]”.

我们在 Office-Home 数据集上进行了对比实验, 其实验结果如表 1 所示. 从表 1 中可以看出: DAMPL 方法在多个任务上均取得了最佳性能, 显著优于其他方法. DAMPL 的平均准确率为 83.8%, 相比基线模型 CLIP 的平均准确率提高了 11.8%. 与次优的 DIFO 方法相比, 平均准确率提高了 4.4%, 在具有显著的域偏移的 R→A 和 R→C 任务上的准确率分别提高了 5.8% 和 7.6%, 表明 DAMPL 在从具有挑战性的域迁移到简单域时具有出色的鲁棒性和泛化能力.

为进一步验证模型的有效性, 在 miniDomainNet 数据集上进行了对比实验, 其结果如表 2 所示. 提出的 DAMPL 方法以 79.7% 的平均准确率超越现有的所有方法, 与基线方法 CLIP 相比提高了 8.5%, 与现有的 SOTA 方法相比提升 1.7%, 验证了其伪标签校正策略与互信息最大化的有效性. 尤其是在具有挑战性的迁移任务 (如 r→c 和 s→c), DAMPL 也展现出性能的显著提升. 然而, DAMPL 在部分子任务 c→r 和 p→r 中却表现出次优性能, 可能是在视觉差异极大的子任务中, 文本模态的适配能力受限于预训练 CLIP 的视觉-文本对齐强度而导致的性能下降.

表 1 在 Office-Home 数据集上的准确率 (%)

方法	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	*Avg*
DANN[2]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN[4]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
PLUE[31]	49.1	73.5	78.2	62.9	73.5	74.5	62.2	48.3	78.6	68.6	51.8	81.5	66.9
COWA[32]	56.9	78.4	81.0	69.1	80.0	79.9	67.7	57.2	82.4	72.8	60.5	84.5	72.5
TPDS[33]	59.3	80.3	82.1	70.6	79.4	80.9	69.8	56.8	82.1	74.5	61.2	85.3	73.5
DAPL[21]	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5
PDA[34]	55.4	85.1	85.8	75.2	85.2	85.2	74.2	55.2	85.8	74.7	55.8	86.3	75.3
DAMP[35]	59.7	88.5	86.8	76.6	**88.9**	87.0	76.3	59.6	87.1	77.0	61.0	89.9	78.2
DIFO[36]	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4
CLIP[8]	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
DAMPL	**69.8**	**90.1**	**91.3**	**84.2**	88.7	**90.5**	**83.6**	**68.5**	**90.6**	**85.8**	**70.7**	**91.2**	**83.8**

注: 加粗字体表示最优结果

表 2 在 miniDomainNet 数据集上的准确率 (%)

方法	c→p	c→r	c→s	p→c	p→r	p→s	r→c	r→p	r→s	s→c	s→p	s→r	*Avg*
PLUE[31]	59.8	74.0	56.0	61.6	78.5	57.9	61.6	65.9	53.8	67.5	64.3	76.0	64.7
TPDS[33]	62.9	77.1	59.8	65.6	79.0	61.5	66.4	67.0	58.2	68.6	64.3	75.3	67.1
SHOT[37]	63.5	78.2	59.5	67.9	81.3	61.7	67.7	67.6	57.8	70.2	64.0	78.0	68.1
COWA[32]	64.6	80.6	60.6	66.2	79.8	60.8	69.0	67.2	60.0	69.0	65.8	79.9	68.6
GKD[38]	61.4	77.4	60.3	69.6	81.4	63.2	68.3	68.4	59.5	71.5	65.2	77.6	68.7
DAPL[21]	72.4	87.6	65.9	72.7	87.6	65.6	73.2	72.4	66.2	73.8	72.9	87.8	74.8
ADCLIP[39]	71.7	88.1	66.0	73.2	86.9	65.2	73.6	73.0	68.4	72.3	74.2	**89.3**	75.2
DAMP[35]	76.7	88.5	71.7	74.2	88.7	70.8	74.4	75.7	70.5	74.9	76.1	88.2	77.5
UniMoS[40]	**76.0**	**88.9**	72.1	75.5	**89.2**	71.1	75.1	75.9	70.5	76.4	**76.3**	88.9	78.0
CLIP[8]	67.9	84.8	62.9	69.1	84.8	62.9	69.2	67.9	62.9	69.1	67.9	84.8	71.2
DAMPL	74.6	86.8	**75.1**	**81.3**	87.3	**75.9**	**82.1**	**77.0**	**73.7**	**82.2**	73.8	86.3	**79.7**

注: 加粗字体表示最优结果

为证明模型具有广泛应用性, 使用具有挑战性的 VisDA-2017 数据集, 实验结果如表 3 所示. 相较于零样本 CLIP 基线模型的性能, DAMPL 方法实现了 5.4% 的显著提升, 平均分类准确率高达 89.8%, 显著优于其他所有方法. 进一步仔细观察, 与忽略文本模态信息的纯视觉方法 JAN 和 BNM 相比, DAMPL 在 skateboard 和 train 等纹理复杂的类别中平均提升 25% 以上, 表明其融合 CLIP 多模态先验知识与互信息最大化的策略能够有效捕捉细粒度特征. 对于类别 person, 基线模型零样本 CLIP 的分类准确率仅为 65.7%, 而 DAMPL 的准确率大幅提升至 85.3%, 进一步表明了所提出优化方法的有效性.

表 3 在 VisDA-2017 数据集上的准确率 (%)

方法	Aeroplane	Bicycle	Bus	Car	Horse	Knife	Motorcycle	Person	Plant	Skateboard	Train	Trunk	*Avg*
DANN[2]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
JAN[4]	82.9	18.7	82.3	**86.3**	70.2	56.9	80.5	53.8	92.5	32.2	84.5	54.5	65.7
BNM[41]	89.6	61.5	76.9	55	89.3	69.1	81.3	65.5	90	47.3	89.1	30.1	70.4
SWD[42]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
SDAT[43]	95.8	85.5	76.9	69.0	93.5	**97.4**	88.5	78.2	93.1	91.6	86.3	55.3	84.3
DAPL[21]	83.9	83.1	88.8	77.9	97.4	91.5	64.2	79.7	88.6	89.3	92.5	62	86.9
UniMoS[40]	97.7	88.2	90.1	74.6	96.8	95.8	92.4	84.1	90.8	89.0	91.8	65.3	88.1
DAMP[35]	97.3	91.6	89.1	76.4	97.5	94.0	92.3	84.5	91.2	88.1	91.2	**67.0**	88.4
PDA[34]	96.7	88.8	87.0	82.8	97.1	93.0	91.3	83.0	**95.5**	91.8	91.5	63.0	88.5
CLIP[8]	98.2	**97.8**	90.5	73.5	97.2	84.0	95.3	65.7	79.4	89.9	91.8	63.3	84.4
DAMPL	**99.3**	92.1	**94.1**	77.8	**98.4**	95.9	**95.3**	**85.3**	88	**96.3**	**95.3**	59.8	**89.8**

注: 加粗字体表示最优结果

3.4.2 消融实验

(1) 各模块的消融研究: 为了深入理解不同优化策略对基线模型 CLIP 性能的影响, 本文在 Office-Home 数据集上设计了消融实验, 分别引入了特定领域的提示学习范式 (DP)、基于语义引导的伪标签校正策略 (LG), 以及互信息最大化损失 (IML). 其中基线模型 CLIP 伪标签通过零样本 CLIP 模型生成, 采用人工设计的提示模板“a photo of [CLASS]”. “√”表示使用相应方法, 实验结果如表 4 所示.

我们首先评估了实验 1 基线模型 CLIP 在分类任务中的表现, 其分类准确度为 72.0%. 在实验 2 中, 我们单独引入了特定领域的提示学习范式 (DP) 来增强模型的理解能力, 其分类准确度提升至 74.5%, 比基线模型提高了 2.5%. 这一结果表明特定领域的提示学习范式通过提供与特定领域紧密相关的先验知识, 能够有效地提高模型在分类任务中的准确性. 实验 3 单独引入语义引导的伪标签校正策略 (LG) 后, 模型准确度提升至 73.9%, 比基线模型提高了 1.9%. 这表明通过语义引导的伪标签校正可以有效地减少错误标签的影响, 从而提高模型分类性能. 在实验 4 中, 当 DP 和 LG 组合使用时, 模型准确度显著提升至 82.8%, 比基线模型提高了 10.8%. 这一显著提升表明两种策略之间存在互补性, 共同促进了模型性能的提高. 实验 5 在实验 4 的基础上再加入互信息最大化损失 (IML), 模型准确度进一步提升至 83.8%, 比仅使用 DP 和 LG 的模型又提高了 1.0%. 这表明互信息最大化损失有助于进一步增强模型对特征关联性的学习, 从而提高分类准确度.

表 4 各模块的消融实验结果

序号	CLIP	DP	LG	IML	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	*Avg*
1	√	—	—	—	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0
2	√	√	—	—	54.1	84.3	84.8	74.4	83.7	85	74.5	54.6	84.8	75.2	54.7	83.8	74.5
3	√	—	√	—	57.2	80.4	82.9	73.9	80.7	81.1	72.8	58.6	83.5	73.3	59.9	81.9	73.9
4	√	√	√	—	69.6	88.9	89	83.1	89.8	89.1	83.0	69.1	89.2	82.8	70.1	89.9	82.8
5	√	√	√	√	69.8	90.1	91.3	84.2	88.7	90.5	83.6	68.5	90.6	85.8	70.7	91.2	83.8

(2) 提示令牌不同长度的消融研究: 本文评估了提示令牌不同长度组合对模型性能的影响, 与领域无关和特定于领域的上下文令牌的长度分别用 M_1 和 M_2 表示, 设置固定总长度为 32, 实验结果如表 5 所示.

从实验结果可以看出, 在数据集 Office-Home、miniDomainNet 上提示令牌长度为 (16, 16) 时性能取得最优, 而在数据集 VisDA-2017 上提示令牌长度为 (24, 8)、(28, 4) 时取得最优. 考虑可能是在偏移程度较大的数据集上, 对通用知识依赖程度较高. 但整体上本文提出的方法性能上对令牌长度不敏感, 影响很小. 这表明 CLIP 的文本编码器能够通过少量关键令牌来学习连续表示.

3.4.3 可视化

为了进一步探索领域特定上下文提示策略的有效性, 本文在图 5 中比较了使用 3 种不同的提示模板对目标域中真实类别预测置信度的影响: (1) 人工设计的提示模板 (与 CLIP 方法一致); (2) 仅包含领域无关上下文的提示模板; (3) 融合领域无关上下文与领域特定上下文的提示模板.

表 5 提示令牌不同长度的分类准确率 (%)

(M_1, M_2)	Office-Home	miniDomainNet	VisDA-2017
(4, 28)	83.5	76.4	89.3
(8, 24)	83.7	79.6	89.6
(16, 16)	83.8	79.7	89.8
(24, 8)	83.6	79.6	89.9
(28, 4)	83.3	78.5	89.9

图 5 在 VisDA-2017 和 Office-Home 数据集上的预测置信度

对于第 1 个例子, 当目标物体公交车被高架桥栏杆所遮挡时, 人工设计的提示无法精准捕捉车身图案与复杂交通场景的关联特征, 从而导致预测置信度显著降低, 相比之下, 包含环境信息细化描述的可学习提示策略则能提升模型表现. 对于最后一个示例, 可学习的领域不可知上下文提示策略甚至比人工设计的提示模板表现得更差, 而包含领域特定上下文的提示通过学习商品域特定知识, 如纯白背景、玳瑁纹理等专属特征, 能够建立更准确的类别判别依据. 总的来说, 比较结果验证了可学习的领域不可知上下文和领域特定上下文相结合的提示模板提高了本文模型的性能.

4 结论与展望

本文提出了一种基于 CLIP 的无监督领域自适应方法 DAMPL, 有效解决了传统 UDA 方法中语义特征丢失的问题. 通过结合 CLIP 模型, 成功地为图像注入了丰富的文本描述信息, 提高了模型对图像语义特征的理解, 从而在分类任务中实现了更高的识别精度. 此外, 特定领域的提示学习机制和基于语义引导的伪标签调整策略, 有效保留了域间特有信息, 提升了类别的可分辨性和模型的泛化能力. 互信息最大化损失的引入进一步增强了源域和目标域特征表示的共享性和域不变性, 降低了过拟合风险. 实验结果表明, DAMPL 方法在无监督域适应任务中展现了优异的性能. 然而, 模型在视觉差异极大的子任务中表现出次优性能, 原因在于文本模态的适配能力受限于预训练 CLIP 的视觉-文本对齐强度而导致的性能下降. 未来工作将进一步探索动态多模态融合与领域感知的伪标签优化.

参考文献

1 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

2 Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 1180–1189.

3 Zellinger W, Grubinger T, Lughofer E, *et al*. Central moment discrepancy (CMD) for domain-invariant representation learning. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

4 Long MS, Zhu H, Wang JM, *et al*. Deep transfer learning with joint adaptation networks. Proceedings of the 34th International Conference on Machine Learning. Sydney:

JMLR.org, 2017. 2208–2217.
5 Xu YC, Cao HZ, Mao KZ, *et al*. Aligning correlation information for domain adaptation in action recognition. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6767–6778. [doi: 10.1109/TNNLS.2022.3212909]
6 Lu ZH, Yang YX, Zhu XT, *et al*. Stochastic classifiers for unsupervised domain adaptation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9108–9117.
7 Kang Q, Yao SY, Zhou MC, *et al*. Effective visual domain adaptation via generative adversarial distribution matching. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(9): 3919–3929. [doi: 10.1109/TNNLS.2020.3016180]
8 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.
9 Pan SJ, Tsang IW, Kwok JT, *et al*. Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks, 2011, 22(2): 199–210. [doi: 10.1109/TNN.2010.2091281]
10 Sun BC, Feng JS, Saenko K. Return of frustratingly easy domain adaptation. Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI). Phoenix: AAAI Press, 2016. 114–120.
11 Long MS, Cao Y, Wang JM, *et al*. Learning transferable features with deep adaptation networks. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 97–105.
12 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.
13 Long MS, Cao ZJ, Wang JM, *et al*. Conditional adversarial domain adaptation. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 1647–1657.
14 Liu H, Long MS, Wang JM, *et al*. Transferable adversarial training: A general approach to adapting deep classifiers. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2020. 4013–4022.
15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
16 He KM, Fan HQ, Wu YX, *et al*. Momentum contrast for unsupervised visual representation learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735.
17 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021. 1–22.
18 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.
19 Zhou K, Yang J, Loy CC, *et al*. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022(130): 2337–2348.
20 Zhou KY, Yang JK, Loy CC, *et al*. Conditional prompt learning for vision-language models. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16795–16804.
21 Ge CJ, Huang R, Xie MX, *et al*. Domain adaptation via prompt learning. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1160–1170. [doi: 10.1109/TNNLS.2023.3327962]
22 Lee DH. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. Proceedings of the 30th International Conference on Machine Learning. Atlanta: PMLR, 2013. 1–6.
23 Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1195–1204.
24 Xie QZ, Dai ZH, Hovy E, *et al*. Unsupervised data augmentation for consistency training. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 525.
25 Zhang YC, Liu TL, Long MS, *et al*. Bridging theory and algorithm for domain adaptation. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 7404–7413.
26 Chen XY, Wang SN, Long MS, *et al*. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. Proceedings of the 36th International

Conference on Machine Learning. Long Beach: PMLR, 2019. 1081–1090.
27 Ji X, Vedaldi A, Henriques J. Invariant information clustering for unsupervised image classification and segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9864–9873.
28 Venkateswara H, Eusebio J, Chakraborty S, *et al*. Deep hashing network for unsupervised domain adaptation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5385–5394.
29 Wang YQ, Liu FR, Chen ZT, *et al*. Contrastive-ACE: Domain generalization through alignment of causal mechanisms. IEEE Transactions on Image Processing, 2023, 32: 235–250. [doi: 10.1109/TIP.2022.3227457]
30 Peng XC, Usman B, Kaushik N, *et al*. VisDA: A synthetic-to-real benchmark for visual domain adaptation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018. 2102–21025.
31 Litrico M, Del Bue A, Morerio P, *et al*. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7640–7650.
32 Lee J, Jung D, Yim J, *et al*. Confidence score for source-free unsupervised domain adaptation. Proceedings of the 39th International Conference on Machine Learning. Baltimore: ICML, 2022. 12365–12377.
33 Tang S, Chang A, Zhang FB, *et al*. Source-free domain adaptation via target prediction distribution searching. International Journal of Computer Vision, 2024, 132(3): 654–672. [doi: 10.1007/s11263-023-01892-w]
34 Bai SH, Zhang M, Zhou WQ, *et al*. Prompt-based distribution alignment for unsupervised domain adaptation. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024. 729–737.
35 Du ZK, Li XY, Li FL, *et al*. Domain-agnostic mutual prompting for unsupervised domain adaptation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 23375–23384.
36 Tang S, Su WX, Ye M, *et al*. Source-free domain adaptation with frozen multimodal foundation model. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 23711–23720.
37 Liang J, Hu DP, Feng JS. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. 560.
38 Tang S, Shi YJ, Ma ZY, *et al*. Model adaptation through hypothesis transfer with gradual knowledge distillation. Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021. 5679–5685.
39 Singha M, Pal H, Jha A, *et al*. AD-CLIP: Adapting domains in prompt space using CLIP. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshop. Paris: IEEE, 2023. 4357–4366.
40 Li XY, Li YK, Du ZK, *et al*. Split to merge: Unifying separated modalities for unsupervised domain adaptation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 23364–23374.
41 Cui SH, Wang SH, Zhuo JB, *et al*. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3940–3949.
42 Lee CY, Batra T, Baig MH, *et al*. Sliced Wasserstein discrepancy for unsupervised domain adaptation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10277–10287.
43 Rangwani H, Aithal SK, Mishra M, *et al*. A closer look at smoothness in domain adversarial training. Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022. 18378–18399.

(校对责编: 李慧鑫)