基于特征异常检测与伪标签回归的无监督对抗域适应

潘 杰[1],刘 波[1],邹筱瑜[2*]
(1. 中国矿业大学信息与控制工程学院,江苏徐州 221116;2. 中国矿业大学机电工程学院,江苏徐州 221116)

摘 要: 无监督域适应任务中源域和目标域通常不满足独立同分布假设. 为生成目标域可用标签,经典域适应方法选择分类器预测概率最大的类别作为目标样本伪标签,使伪标签中可能包含噪声信息,造成域适应"负迁移". 此外,传统对抗域适应方法往往考虑对齐领域间全局分布,较少关注样本类别信息,如何在域适应任务中提取判别性类别级特征至关重要. 为此,本文提出一种基于特征异常检测与伪标签回归的无监督对抗域适应方法. 通过分类器预测同类别目标样本组成目标域类别子域,引入高斯均匀混合模型检测与类均值特征距离异常的子域样本,计算样本后验概率并以此度量子域中样本伪标签的正确性,作为损失因子限制伪标签在训练中对模型的影响. 同时,采用伪标签回归函数减小分类器预测标签与高置信度伪标签差异,对无标签目标域进行类别约束,提高特征类别可辨别性. 实验表明,所提方法在数据集 Office-31、Image-CLEF 和 Office-Home 上平均识别精度分别为 90.2%、89.6% 和 69.5%,较相关主流算法均有提升.

关键词: 对抗域适应;特征检测;高斯均匀混合模型;伪标签回归;无监督学习;图像分类
基金项目: 国家自然科学基金(No.62176258,No.62273349,No.61806207);中央高校基本科研业务费专项资金项目(No.2021YCPY0111)
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112(2025)01-0128-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240074

Feature Anomaly Detection and Pseudo-Label Regression for Adversarial Domain Adaptation

PAN Jie[1], LIU Bo[1], ZOU Xiao-yu[2*]
(1. *School of Information and Control Engineering*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*;
2. *School of Mechanical and Electrical Engineering*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116, *China*)

Abstract: In unsupervised domain adaptation tasks, the source and target domains usually do not satisfy the independent and identical distribution assumption. In order to generate the usable labels for the target domain, classical domain adaptation methods select the category with the highest prediction probability of the classifier as the pseudo-label of the target sample. Thus, the pseudo-label inevitably contains certain noise information, which may cause negative transfer to the domain adaptation model. In addition, traditional adversarial domain adaptation methods usually consider the global distribution between domains and ignore the category information of samples. How to extract discriminative category-level features in domain adaptation tasks is also an important problem. Therefore, an unsupervised adversarial domain adaptation method is proposed using feature anomaly detection and pseudo-label regression. The target samples of the same class predicted by the classifier are formed into the category subdomain within the target domain. The Gaussian uniform mixture model is used to detect the subdomain samples with abnormal distance from the class mean. The posterior probability of the samples is calculated and the correctness of the sample pseudo-labels in the subdomain is measured, which is used as a loss factor to limit the influence of pseudo-labels on the model in training. Meanwhile, the pseudo-label regression function is used to reduce the difference between the predicted label and the high-confidence pseudo-label of the classifier. The category constraint of the unlabeled target domain is adopted to improve the distinguishability of feature categories. Experimental results

收稿日期:2024-01-15;修回日期:2024-09-06;责任编辑:李勇锋
*通讯作者:邹筱瑜

show that the average recognition accuracy of the proposed method on datasets Office-31, Image-CLEF, and Office-Home are 90.2%, 89.6%, and 69.5%, respectively, which are all higher than the related popular algorithms.

Key words: adversarial domain adaptation; feature detection; Gaussian uniform mixture model; pseudo-label regression; unsupervised learning; image classification

Foundation Item(s): National Natural Science Foundation of China (No.62176258, No.62273349, No.61806207); Fundamental Research Funds for the Central Universities (No.2021YCPY0111)

1 引言

随着计算机硬件的发展,依靠大规模训练数据的深度学习模型在视觉领域取得了显著成功[1~3]. 然而,人工标记数据代价昂贵,受图像来源、拍摄环境与设备差异等因素影响,相同视觉任务在不同领域存在显著差异,导致源域模型在目标域上的性能不佳[4]. 为此,无监督域适应方法[5]被提出,旨在利用源域与目标域相似知识缩小领域差异,提升源学习器在无监督目标域上的泛化性能. 其核心思想在于学习领域不变性特征[6]或估计源域样本重要性以对齐数据分布[7]. 近来,深度无监督域适应[8]受益于深度神经网络多隐层结构与特征表示能力,在图像分类、语义分割、超分辨率重构等视觉任务上取得了丰富成果,主要包括分布距离度量、对抗域适应与伪标签学习等方法.

一般而言,分布距离度量利用数据分布函数衡量领域间距离,最小化该距离实现源域与目标域对齐,其中最大均值差异[9~11]是典型方法. 此外,诸多研究中引入了不同距离度量模式[12]. 关联对齐[13]通过领域间协方差矩阵对齐数据分布. 中心矩方法[14]利用领域间不同阶中心向量进行特定域特征匹配. Wasserstein 方法[15]则以最小传输代价将源域数据转化为目标域分布,度量不重叠领域的分布差异.

不同于距离方法需要设计一个复杂的分布距离度量函数,对抗域适应[16~19]采用“对抗思想”在领域判别器与特征提取器间形成博弈,将领域分布的距离关系转化为潜在特征空间的领域混淆过程,生成具有判别性和领域不变性的特征表示. 然而,当前对抗域适应方法中主要考虑对齐领域的边缘分布,较少关注不同类别样本间的条件分布,使领域判别器缺乏类别语义信息,造成样本区分度下降. 为此,多对抗域适应[20,21]与条件对抗域适应[22,23]被提出,利用样本特征与预测类别构造高阶特征判别模型,对齐领域条件概率分布. 另外,最大分类器差异法[24]采用分类器同时实现领域判别与类别预测,能够学习到具有类别信息的更具判别性的特征表示. 最小最大熵方法[25]通过交替最大化分类器无标签目标样本条件熵与最小化特征提取器条件熵来提取判别性特征. 需要指出的是,多对抗与条件对抗的方法未考虑目标域预测类别正确性,易引起特征类别结构混淆. 而最大分类器差异与最小最大熵方法在领域适应过程中需要先验知识,包括对模型进行预训练或需求目标域标签数据. 因此,如何自适应提取具有类别判别性与领域不变性的特征表示,仍是目前亟需解决的问题.

另一方面,伪标签学习等方法将分类器输出概率最高类别作为样本伪标签,引入训练提升模型域适应性能. 为确保可靠性,不对称多学习器法[26]采用两个分类器进行结果印证,但面对差异较大领域仍会产生“负迁移”. 语义表征法[27]避免了使用伪标签数据直接训练模型,而利用其对齐源域与目标域同类别样本中心以学习语义信息. 协同对抗网络[28,29]基于分类器得分设计目标样本权重函数,以降低目标域伪标签监督损失在训练中对域适应模型的影响. 能够看出,上述方法均依赖于分类器预测精度,在训练初始阶段或领域差异较大时,分类器获得的伪标签依然包含噪声信息,影响最终域适应效果.

综上,为解决无监督领域适应中出现的类别级特征表示难、目标域伪标签噪声大等问题,本文提出一种基于特征异常检测与伪标签回归的对抗域适应方法,主要分为伪标签后验概率预测与对抗域适应模型学习两个阶段. 伪标签预测阶段,利用特征提取器和分类器获取目标域特征与标签信息,更新高斯均匀混合模型参数,获得目标样本后验概率进行特征异常检测. 对抗域适应学习阶段,基于后验概率和高置信度伪标签数据学习对抗域适应网络. 相比直接剔除置信度低的伪标签,本文方法利用后验概率阈值进行剔除,并引入高斯均匀混合模型对目标域子域特征距离异常的样本进行检测,得到目标样本伪标签后验概率,衡量伪标签真实性. 两阶段循环迭代缩小领域差异,提升目标域分类性能. 本文主要贡献如下:

(1)针对目标域伪标签噪声问题,设计类别相关的高斯均匀混合模型进行特征异常检测,计算样本后验概率衡量伪标签正确性,利用高置信度伪标签数据扩充目标域训练集.

(2)针对类别级特征表示问题,设计伪标签回归损失函数,减少目标样本预测结果与高置信度伪标签差异,利用熵最小化方法学习对抗域适应网络,得到类别区分性高的特征表示,提升无监督预测性能.

(3)采用 Office-31、Office-Home 和 Image-CLEF 等典

型数据集,将本文方法与当前无监督域适应主流方法进行实验对比.结果表明,所提方法能够生成可靠的目标样本伪标签并获得高判别性特征,提高无监督对抗域适应性能.

2 相关工作

2.1 无监督域适应

给定标签源域$D_s=\{x_i^s,y_i^s\}(i=1,2,\cdots,n_s)$和无标签目标域$D_t=\{x_j^t\}_{j=1}^{n_t}$,其中$n_s$与$n_t$表示源域与目标域样本数,$x_i^s$和$x_j^t$为源域与目标域样本,$y_i^s$为源域标签,同时源域$D_s$和目标域$D_t$服从不同边缘概率分布$P(x^s)\neq P(x^t)$.无监督域适应旨在利用源域$D_s$标签数据学习映射$y=f(x)$,减小不同领域样本分布差异,使目标域$D_t$泛化误差最小化.研究表明,深度神经网络相较于传统方法能够学习到泛化性更强的特征表示[30].因此,无监督域适应距离度量通过设计特征分布的距离函数衡量领域间深度特征差异,最小化特征差异提高目标域泛化性能,其损失函数如下[14]:

$$\min_{\theta_C,\theta_G}\frac{1}{n_s}\sum_{i=1}^{n_s}J\left(C\left(G(x_i^s)\right),y_i^s\right)+\alpha\cdot\tilde{d}\left(P(x^s),P(x^t)\right)\quad(1)$$

式中,θ_G和θ_C为特征提取器G和分类器C参数,$J(\cdot,\cdot)$表示交叉熵损失,$\alpha>0$为损失权重,$\tilde{d}(\cdot,\cdot)$表示不同领域分布距离,n_s表示源域与目标域样本数,x_i^s和x_j^t为源域与目标域样本,y_i^s为源域标签.

2.2 对抗域适应

最小化源域分类误差与领域间差异可以提高泛化性能[31],而对抗域适应方法通过在领域判别器上施加二元交叉熵损失以最小化领域差异.给定标签源域$D_s=\{x_i^s,y_i^s\}(i=1,2,\cdots,n_s)$,利用交叉熵损失实现源域分类误差最小化如下[17]:

$$\min_{\theta_C,\theta_G}L_{\text{sce}}=\frac{1}{n_s}\sum_{i=1}^{n_s}J\left(C\left(G(x_i^s)\right),y_i^s\right)\quad(2)$$

式中,θ_G和θ_C为特征提取器G和分类器C参数,n_s表示源域与目标域样本数,x_i^s和y_i^s为源域样本和源域标签.对抗域适应在特征提取器上最大化域对抗损失,提取领域不变性特征混淆领域判别器,在领域判别器上最小化域对抗损失,区分源域与目标域特征.域对抗损失基于如下二元交叉熵损失函数实现[32]:

$$\min_{\theta_G,\theta_D}L_{\text{adv}}=-\frac{1}{n_s}\sum_{i=1}^{n_s}\log\left(D\left(G(x_i^s)\right)\right)-\frac{1}{n_t}\left(1-\log\left(D(G(x_j^t))\right)\right)\quad(3)$$

式中,θ_G和θ_D为特征提取器G与领域判别器D参数,n_s与n_t表示源域与目标域样本数,x_i^s和x_j^t为源域与目标域样本.此外,为避免分别最大化最小化训练特征提取器和领域判别器,引入梯度反转层[[17],取反前后分别最小最大化域对抗损失,生成具有领域不变性的特征减小领域差异.

3 基于特征异常检测与伪标签回归的对抗域适应模型

为生成具有判别性类别级特征并减小伪标签噪声影响,本文提出一种基于特征异常检测与伪标签回归的对抗域适应模型,如图1.模型由特征提取器G、分类器C、领域判别器D与k个高斯均匀混合模型p_k组成,其中θ_G、θ_C、θ_D和θ_k为相应模块参数,$\hat{d}_j^k$表示第k类目标样本特征距离.模型分为两个阶段,伪标签预测与对抗域适应学习.

阶段1中,初始化θ_G与θ_C并将其固定,由特征提取器G获取源域和目标域样本特征,采用余弦相似度得到目标样本各类别特征距离$\hat{d}_j^1,\hat{d}_j^2,\cdots,\hat{d}_j^k$,将其分别输入$k$个高斯均匀混合模型,结合目标样本伪标签$\tilde{y}_j^t$得到样本后验概率,检测目标域各类别特征异常样本.阶段2中采用随机梯度下降法更新θ_G、θ_C与θ_D,在保留域对抗损失L_{adv}和源域交叉熵损失L_{sce}前提下,引入伪标签回归损失L_{pre}、条件熵损失L_{ce}以及高置信度伪标签交叉熵损失L_{tce},提高模型目标域泛化能力和伪标签复用能力.模型分阶段在无监督异常值检测和对抗域适应学习间循环迭代.一方面训练阶段1高斯均匀混合模型获得目标域伪标签$\tilde{y}_j^t$与样本后验概率$\gamma_j(\theta_k)$,为模型学习阶段提供伪标签先验知识.另一方面基于训练完成的对抗域适应网络,将阶段2中特征提取器G和分类器C的参数共享给阶段1,为下一轮更新高斯均匀混合模型提供训练数据.

3.1 特征异常检测

研究表明,域适应任务中引入目标域样本标签有助于提高域适应性能[33,34].然而,无监督域适应目标域缺乏标签信息,分类器预测的伪标签包含一定噪声,这对无监督域适应精度带来了挑战.如何降低噪声提高伪标签准确性,即检测出特征异常样本显得至关重要.通常来说,与目标域子域类均值特征距离大的样本服从均匀分布,距离小的样本服从高斯分布[35],且高斯均匀混合模型无需训练即可直接应用,在处理伪标签时与直接剔除置信度低的伪标签相比,节省了时间成本.因此,本文采用高斯均匀混合模型对目标域样本特征距离进行无监督检测,利用其输出的后验概率度量子域中伪标签真实性,设定阈值分离特征异常目标样本,限制噪声对模型训练的影响.

3.1.1 特征相似性

由分类器对目标样本预测的同类别伪标签可组成目标域子域,给定源域$D_s=\{D_s^k\}(k=1,2,\cdots,K)$,其中子域$D_s^k$表示第$k$类源样本集合,$K$为类别总数.由于目标

样本x_j^t不含标签,经过特征提取器G提取特征后,利用分类器C预测其伪标签$\tilde{y}_j^t=\text{argmax}_k[C(G(x_j^t))]_k$. 根据类别预测结果,目标域表示为$D_t=\{D_t^k\}(k=1,2,\cdots,K)$,其中子域$D_t^k$表示第$k$类伪标签对应的目标样本集合. 在训练初期或面对领域差异较大的任务,分类器在目标域上准确率较低,导致D_t^k中相当样本被赋予错误标签. 为衡量伪标签真实性,目标域样本x_j^t特征与同类子域的类均值特征相似性表示为

$$d_j^k=\text{CS}\left(G(x_j^t),\mu_k^t\right)+\text{CS}\left(G(x_j^t),\mu_k^s\right) \tag{4}$$

$$\mu_k^t=\frac{1}{n_k^t}\sum_{x_j^t\in D_t^k}G(x_j^t),\mu_k^s=\frac{1}{n_k^s}\sum_{x_j^s\in D_s^k}G(x_i^s) \tag{5}$$

式中,CS(·,·)为余弦距离;$G(x_j^t)$表示G提取的样本特征,且$G(x_j^t)\in\mathbf{R}^{256}$;$\mu_k^t$和$\mu_k^s$分别表示源域子域$D_s^k$和目标域子域$D_t^k$的类均值特征,其中$\mu_k^t\in\mathbf{R}^{256},\mu_k^s\in\mathbf{R}^{256}$;$n_k^t$与$n_k^s$表示第$k$类目标样本数与源样本数.

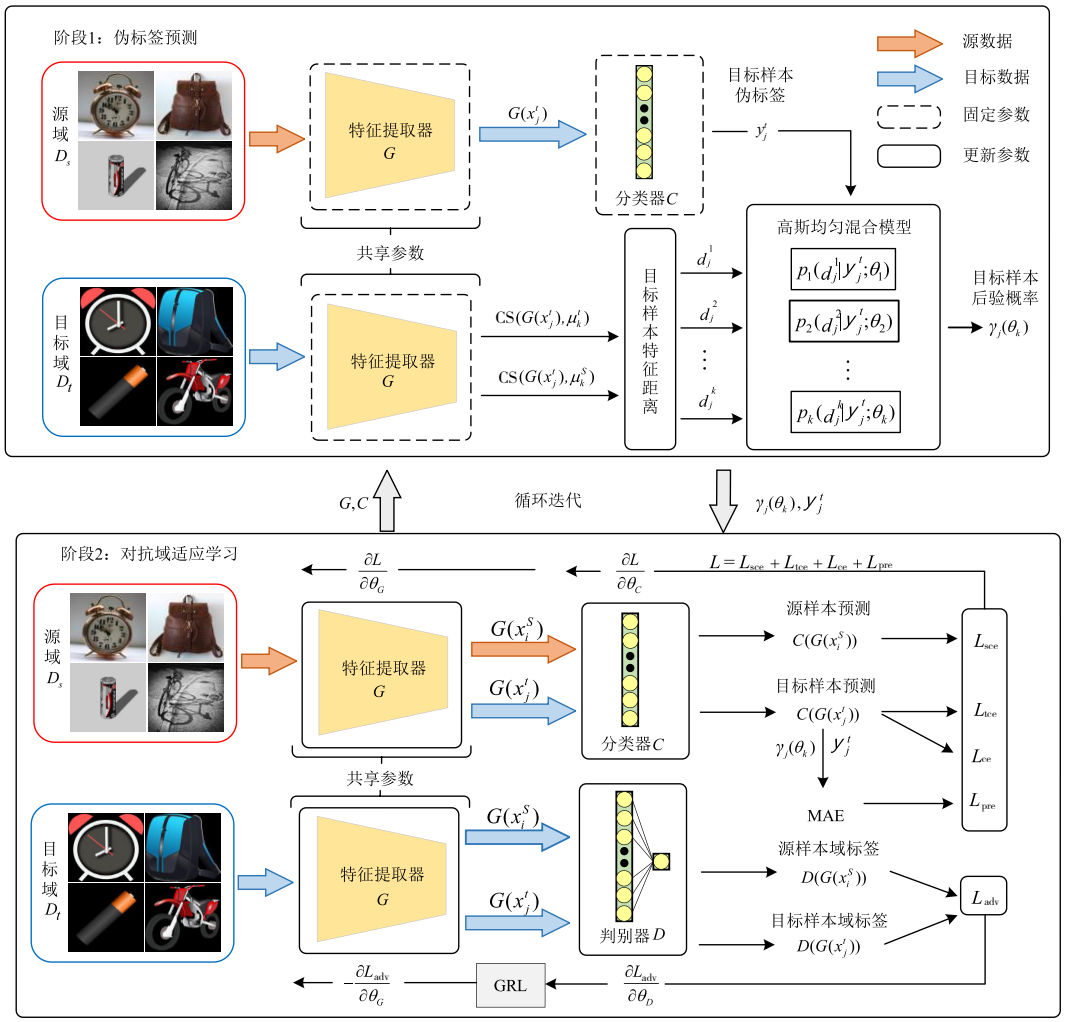


图1　算法框架

3.1.2　特征异常检测

上述余弦距离对于伪标签真实性判别的输出为[0,2],不能直接作为衡量伪标签准确性的指标. 理论上来说,D_t^k中目标样本与类均值特征距离越小,伪标签为真的概率越高,特征距离越大则伪标签为假的概率越高. 对于子域D_t^k中目标样本x_j^t,引入隐变量$z_j\in\{0,1\}$表示目标样本来自不同分布,计算目标样本处于各分布后验概率,判断样本所处分布. 引入高斯均匀混合模型对子域中特征距离异常的目标样本进行检测,衡量伪标签真实性. 并设置阈值分离异常样本伪标签数据. 如图2所示.

高斯均匀混合模型由高斯分布和均匀分布组成,其基于目标样本特征距离$\hat{d}_j^k$进行建模,其中$\hat{d}_j^k=(-1)^{m_j}d_j^k$,

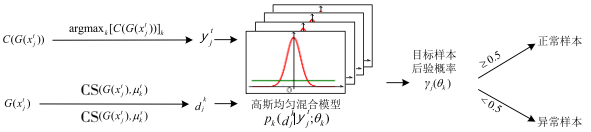


图2　目标样本特征异常检测过程

$j=\{1,2,\cdots,n_t^k\}$，m_j采样自伯努利分布$B(1,0.5)$. 高斯均匀混合模型的概率密度函数如下：

$$p_k\left(\hat{d}_j^k\tilde{y}_j^t;\ \theta_k\right)=\pi_k N\left(\hat{d}_j^k;\ \sigma_k\right)+\left(1-\pi_k\right)\ U\left(\hat{d}_j^k;\ \delta_k\right) \tag{6}$$

式中：$\theta_k=\left\{\pi_k,\sigma_k,\delta_k\right\}(k=1,2,\cdots,K)$表示高斯均匀混合模型参数，$k$为类别数. π_k为样本先验概率，σ_k为高斯分布N标准差，均值$\mu=0$，均匀分布U参数为δ_k，定义区间为$[-\delta_k,\delta_k]$.

目标样本x_j^t特征距离$\hat{d}_j^t$来自高斯分布的后验概率

$$P\left(z_j=1\hat{d}_j^k;\ \theta_k^{(i+1)}\right)=\frac{\pi_k^{(i)}N\left(\hat{d}_j^k;\ \sigma_k^{(i)}\right)}{\pi_k^{(i)}N\left(\hat{d}_j^k;\ \sigma_k^{(i)}\right)+\left(1-\pi_k^{(i)}\right)\ U\left(\hat{d}_j^k;\ \delta_k^{(i)}\right)} \tag{7}$$

式中，$\pi_k^{(i)}$、$\sigma_k^{(i)}$和$\delta_k^{(i)}$为期望最大化算法迭代第i次的样本先验概率、高斯分布方差和均匀分布参数. 为简化表示，令$\gamma_j\left(\theta_k\right)=P(z_j=1\hat{d}_j^k;\ \theta_k^{(i+1)})$.

3.1.3 模型训练过程

高斯均匀混合模型基于期望最大算法更新参数$\theta_k=\left\{\pi_k,\sigma_k,\delta_k\right\}(k=1,2,\cdots,K)$，对含有隐变量数据集的概率模型迭代进行最大似然估计，其在E步计算样本后验概率和M步更新模型参数之间迭代进行，直到似然函数收敛时停止训练. E步输出目标样本x_j^t特征距离$\hat{d}_j^t$来自高斯分布的后验概率$\gamma_j\left(\theta_k\right)$后，M步更新模型参数如下：

$$\pi_k^{(i+1)}=\frac{1}{n_k^t}\sum_{j=1}^{n_k^t}\gamma_j\left(\theta_k^{(i+1)}\right) \tag{8}$$

$$\sigma_k^{(i+1)}=\frac{\sum_{j=1}^{n_k^t}\gamma_j\left(\theta_k^{(i+1)}\right)\left(\hat{d}_j^k\right)^2}{\sum_{j=1}^{n_k^t}\gamma_j\left(\theta_k^{(i+1)}\right)} \tag{9}$$

$$\delta_k^{(i+1)}=\sqrt{3(Q_2-{Q_1}^2)} \tag{10}$$

$$\begin{aligned}Q_1&=\frac{1}{\sum_{j=1}^{n_k^t}\gamma_j\left(\theta_k^{(i+1)}\right)}\sum_{j=1}^{n_k^t}\frac{1-\gamma_j\left(\theta_k^{(i+1)}\right)}{1-\pi_k^{(i+1)}}\hat{d}_j^k,\\ Q_2&=\frac{1}{\sum_{j=1}^{n_k^t}\gamma_j\left(\theta_k^{(i+1)}\right)}\sum_{j=1}^{n_k^t}\frac{1-\gamma_j\left(\theta_k^{(i+1)}\right)}{1-\pi_k^{(i+1)}}\left(\hat{d}_j^k\right)^2\end{aligned} \tag{11}$$

式中，n_k^t表示第k类目标样本数；$\hat{d}_j^k$表示第k类目标样本特征距离；Q_1和Q_2表示1阶和2阶中心矩. 相较于常值参数迭代估计，式(10)中$\delta_k^{(i+1)}$采用离群点中心更新以增强区分内点和离群点的能力. 训练中不断迭代E步和M步直到似然函数收敛，此时E步输出的样本后验概率作为衡量目标样本伪标签$\tilde{y}_j^t$正确性指标.

3.2 伪标签回归

传统对抗域适应方法在训练中领域判别器输入不含样本语义信息，只能有限对齐源域和目标域边缘概率分布. 但源域和目标域条件概率分布并不相同，难以提取区分性高的特征表示. 现有研究表明，对抗域适应虽然提高了特征的可迁移性但也削弱了其可判别性[36]. 而使用伪标签方法可以有效扩充训练数据，减少获取真实标签数据的难度，增强模型泛化能力并节省计算成本. 为此，本文采用伪标签回归与熵最小化方法学习域适应模型，利用置信度高的伪标签扩充可训练数据.

3.2.1 回归损失

目标域标签信息的缺失以及伪标签噪声是影响域适应模型分类性能的重要原因. 特征异常检测为训练模型补充了伪标签信息以及度量伪标签正确性的后验概率指标. 在此基础上，进一步采用目标样本后验概率$\gamma_j(\theta_k)$加权回归作为伪标签损失，减小分类器C预测标签与后验概率高的伪标签间差异. 损失函数为

$$\min_{\theta_G,\theta_C}L_{\text{pre}}=\frac{1}{\gamma_j^t}\sum_{j=1}^{n_t}\gamma_j(\theta_k)\text{MAE}(C(G(x_j^t)),\tilde{y}_j^t) \tag{12}$$

$$\gamma_j\left(\theta_k\right)=\begin{cases}\gamma_j\left(\theta_k\right), & \text{if}\,\gamma_j\left(\theta_k\right)\geqslant 0.5\\ 0, & \text{otherwise}\end{cases} \tag{13}$$

式中，θ_G和θ_C表示G和C模型参数；γ_j^t表示目标样本后验概率和，即$\gamma_j^t=\sum_{j=1}^{n_t}\gamma_j(\theta_k)$；MAE$(\cdot,\cdot)$表示平均绝对误差；$x_j^t$表示目标域样本；$\tilde{y}_j^t$表示概率预测阶段中分类器$C$预测的目标样本伪标签. 训练中舍弃后验概率小于0.5的伪标签数据，限制伪标签噪声对网络带来的影响. 通过样本后验概率加权回归函数，加强高置信度伪标签在域适应网络学习中的作用，抑制低置信度伪标签对网络的影响.

3.2.2 熵最小化

领域间差异以及目标域标签缺失使得目标域样本类别预测概率较小，具有高熵值特性[37,38]. 为此，本文采用熵最小化方法惩罚类别预测概率低的目标样本，增强模型对目标样本类别预测确定性，提高类别可区分性. 熵最小化通过以下损失函数实现：

$$\min_{\theta_G}L_{\text{ce}}=-\frac{1}{n_t}\sum_{j=1}^{n_t}C\left(G(x_j^t)\right)\log C\left(G(x_j^t)\right) \tag{14}$$

式中，θ_G为G模型参数，与文献[37]设定相同；n_t为目标域样本数；x_j^t为目标域样本. 本文使用式(14)更新θ_G，旨在训练目标样本特征远离决策边界，解决较大领域差异导致训练初始阶段目标样本被预测错误难以纠正的问题.

3.2.3 训练数据扩展

源域中可训练样本数量有限,而无监督域适应目标域中不含标签信息. 为此,将上一轮模型更新的较高后验概率伪标签数据作为目标样本真实标签引入训练数据,重新权衡分类损失. 通过交叉熵损失函数监督式更新特征提取器G和分类器C如下:

$$\min_{\theta_G,\theta_C} L_{\text{tce}}=\lambda\frac{1}{n_t'}\sum_{j=1}^{n_t'}J\left(C\left(G(x_j^t)\right),\tilde{y}_j^t\right) \tag{15}$$

式中,$J(\cdot,\cdot)$为交叉熵损失,x_j^t为目标域样本,$\tilde{y}_j^t$为概率预测阶段中分类器C预测的目标样本伪标签,n_t'为满足条件的目标样本数. 为抑制伪标签噪声对此项的影响,设置超参数$\lambda\in(0,1)$为损失权重系数,初始阶段λ趋近0并随训练进程逐渐增大. 此外,参与训练的目标样本伪标签$\tilde{y}_j^t$满足$\gamma_j(\theta_k)\geqslant\omega$,$\omega$为后验概率阈值,旨在控制伪标签数据中含有的噪声比例.

3.3 总损失与算法流程

本文方法在源分类损失L_{sce}和域对抗损失L_{adv}基础上,加入式(12)伪标签回归损失和式(14)的条件熵损失,并结合式(15)目标域交叉熵损失,优化目标如下:

$$\min_{\theta_G,\theta_C,\theta_D} L_{\text{total}}=L_{\text{sce}}+L_{\text{adv}}+L_{\text{pre}}+L_{\text{ce}}+L_{\text{tce}} \tag{16}$$

伪标签预测阶段,固定特征提取器与分类器参数预测目标样本伪标签,通过期望最大化算法更新高斯均匀混合模型获得目标样本后验概率. 对抗域适应学习阶段,通过式(16)训练域适应网络,前向传播计算总损失,反向传播更新深度神经网络参数. 所提基于特征异常检测与伪标签回归的对抗域适应(feature anomaly Detection and pseudo-label Regression for Adversarial Domain Adaptation,DR-ADA)算法流程如算法1所示.

算法1　基于特征异常检测与伪标签回归的无监督对抗域适应

数据: 源域$D_s=\{x_i^s,y_i^s\}(i=1,2,\cdots,n_s)$,目标域$D_t=\{x_j^t\}(j=1,2,\cdots,n_t)$.

伪标签预测阶段: 训练高斯均匀混合模型

1. **输入:** 目标样本伪标签$\tilde{y}_j^t$,目标样本和源样本特征.

2. 根据式(4)获得目标特征与类中心的余弦距离$\hat{d}_j^k$.

$$d_j^k=CS(G(x_j^t),\mu_k^t)+CS(G(x_j^t),\mu_k^s)$$

3.FOR n IN 1:n_1

根据式(7)计算目标样本特征距离的后验概率.

$$P(z_j=1\hat{d}_j^k;\theta_k^{(i+1)})=\frac{\pi_k^{(i)}N(\hat{d}_j^k;\sigma_k^{(i)})}{\pi_k^{(i)}N(\hat{d}_j^k;\sigma_k^{(i)})+(1-\pi_k^{(i)})U(\hat{d}_j^k;\delta_k^{(i)})}$$

根据式(8)~(10)更新高斯均匀混合模型参数θ_k.

$$\pi_k^{(i+1)}=\frac{1}{n_k^t}\sum_{j=1}^{n_k'}\gamma_j(\theta_k^{(i+1)})$$

$$\sigma_k^{(i+1)}=\frac{\sum_{j=1}^{n_k'}\gamma_j(\theta_k^{(i+1)})(\hat{d}_j^k)^2}{\sum_{j=1}^{n_k'}\gamma_j(\theta_k^{(i+1)})}$$

$$\delta_k^{(i+1)}=\sqrt{3(Q_2-Q_1^{\,2})}$$

END

对抗域适应学习阶段: 训练对抗域适应模型

4. **输入:** 源域和目标域数据,目标样本伪标签和后验概率.

5.FOR n IN 1:n_2

根据式(16)前向传播计算全局损失L_{total}.

$$\min_{\theta_G,\theta_C,\theta_D} L_{\text{total}}=L_{\text{sce}}+L_{\text{adv}}+L_{\text{pre}}+L_{\text{ce}}+L_{\text{tce}}$$

随机梯度下降算法更新领域适应网络权重.

END

6. **输出:** 目标域样本预测标签y_j^t.

3.4 复杂度分析

伪标签预测阶段的计算复杂度为如下:

$$O\left(N_1\cdot mM^2B^2G_{i-1}G_i+N_2\cdot\sum_{k=1}^{K}(n_k^t)^2+N_1\cdot m\sum_{j=1}^{N_C}C_{j-1}C_j+\sum_{k=1}^{K}(n_k^s+n_k^t)\right) \tag{17}$$

式中,$O\left(N_1\cdot m\sum_{i=1}^{N_G}M^2B^2G_{i-1}G_i\right)$为特征提取器的计算复杂度,$N_1$为迭代次数,$m$为批大小,$N_G$为特征提取器卷积层数,$M$为卷积输出特征图大小,$B$为卷积核边长,$G_i$表示特征提取器$G$第$i$个卷积层卷积核数量;$O\left(N_2\cdot\sum_{k=1}^{K}(n_k^t)^2\right)$为高斯均匀混合模型计算复杂度,$N_2$为相应迭代次数,$K$为类别数,$n_k^t$为第$k$类目标样本数;$O\left(N_1\cdot m\sum_{j=1}^{N_C}C_{j-1}C_j\right)$为分类器网络的计算复杂度,$N_C$为分类器的全连接神经网络层数,$C_j$为分类器$C$的第$j$层神经元数. $O\left(\sum_{k=1}^{K}(n_k^s+n_k^t)\right)$为余弦距离复杂度,$n_k^s$为第$k$类源样本数.

对抗域适应学习阶段计算复杂度如下:

$$O\left(N_3\cdot m\left(\sum_{i=1}^{N_G}M^2B^2G_{i-1}G_i+\sum_{j=1}^{N_C}C_{j-1}C_j+\sum_{l=1}^{N_D}D_{l-1}D_l\right)\right) \tag{18}$$

式中,$O\left(N_3\cdot m\sum_{i=1}^{N_G}M^2B^2G_{i-1}G_i\right)$为特征提取器的计算复杂度,$N_3$为对抗域适应学习阶段迭代次数. $O\left(N_3\cdot m\sum_{l=1}^{N_D}D_{l-1}D_l\right)$为领域判别器计算复杂度,$N_D$为判别器的全连接神经网络层数,$D_l$为判别器$D$的第$l$层神经元数.

传统对抗域适应[18,32]不考虑目标样本伪标签,计算复杂度与式(18)相同. 本文在伪标签预测阶段增加特征异常检测提取目标样本后验概率和伪标签,相较传统对抗域适应方法计算复杂度有所增加,增加部分

如式(17)所示. 但需要指出的是,式(17)伪标签预测阶段特征提取器与分类器只用于网络前向推理,不需更新网络参数,实际产生的计算开销较低. 与传统对抗域适应方法的计算复杂度相比,需要更新网络参数的部分仅额外增加了 $O\left(N_2\cdot\sum_{k=1}^{K}(n_k^t)^2\right)$,并未显著影响模型的运算速率.

4 实验与分析

为评估本文方法,采用Office-31[39]、Image-CLEF[40]和Office-Home[41]等数据集,设计以下实验来论证DRADA模型有效性. 域适应效果方面,通过特征降维可视化显示域适应前后源域和目标域样本特征分布情况. 伪标签预测方面,在Office-31的4个任务上统计不同后验概率阈值下的伪标签正确率. 通过消融实验进一步验证各损失函数对模型性能的影响,给出训练过程收敛曲线. 参数敏感性实验验证了损失函数中超参数对模型分类准确度的影响.

4.1 数据集

Office-31为基准数据集,包含4 110张图像,由3个领域组成:Amazon(A)、DSLR(D)与Webcam(W). 各领域均为31类,分别有2 817张、498张和795张图像. 其中A来自电商网站,W为网络摄像机低分辨率图像,D为单反相机高分辨率图像,共6个基准域适应任务.

Image-CLEF为基准数据集,包含1 800张图像,Caltech-256(C)、ImageNet ILSVRC 2012(I)和Pascal VOC 2012(P)等3个领域,各领域包含12类600张图像,共6个基准域适应任务.

Office-Home为大型挑战数据集,共有15 588张图像,包含Art(A)、Clipart(C)、Product(P)、Real World(R)等4个领域,各领域均为65类,分别有2 427张、4 365张、4 439张和4 357张图像. A为绘画和素描,C为剪贴画,P为去背景实物图像,R为相机拍摄常规图像,含有12个挑战性域适应任务.

VisDA-2017是挑战数据集,由两个差异较大的域组成,该数据集共280 000张图像,共12种类别. 该数据集作为额外的实验所使用的补充数据集,用以验证本文方法针对复杂任务识别精度的优势.

4.2 实验参数与细节

根据无监督域适应任务条件,训练集只有源样本有标签,目标样本不含标签,而测试集为目标样本. 与其他无监督领域适应方法相同,采用ResNet50[3]作为特征提取器,预处理阶段进行尺寸调整、随机裁剪、随机水平翻转和图像标准化方法[18],所提方法使用GPU进行训练,型号为NVIDIA GeForce RTX 3090,显存为24.0 GB,Pytorch版本1.8.1,python版本3.7,并使用CUDA 11.1进行训练加速. 训练中对特征距离数据 $\hat{d}_j^k$ 进行标准化处理,高斯均匀混合模型初始化参数与文献[33]相同,固定特征提取器和分类器参数,通过期望最大化算法更新混合模型参数.

值得注意的是,在初次高斯均匀混合模型训练中,特征提取器G和分类器C是式(1)和式(2)预训练模型,而模型训练阶段特征提取器G由ImageNet[42]预训练的ResNet50和一层全连接神经网络组成,分类器C和判别器D分别含有1层和2层全连接层. 使用随机梯度下降法更新阶段2的深度神经网络参数,动量为0.9,权重衰减为0.000 1. 按照文献[17]更新学习率和超参数λ,特征提取器G学习率η_0通过$\eta_0=0.001/\left(1+\alpha\rho\right)^{\beta}$进行调整,其中$\beta=0.75$,$\alpha=10$,初始学习率为0.001. 分类器$C$和领域判别器$D$的学习率为特征提取器$G$的10倍. 超参数$\omega$设置为0.8,$\lambda=2/(1+\exp(-\varepsilon\rho))-1$,其中$\varepsilon=10$,$\rho$为优化进度从0到1线性变化,超参数$\lambda$也跟随训练进度从0到1变化.

4.3 基准方法介绍

为验证所提方法域适应性能,在3个数据集上分别选取基准方法进行比较:

·深度领域混淆方法(Deep Domain Confusion, DDC)[11]使用线性核计算最大均值差异,在全连接层嵌入线性核计算两领域样本特征均值,对齐全局分布.

·深度适配网络(Deep Adaptation Netowrk, DAN)[12]在3个全连接层进行适配以提高网络迁移能力,提出多核最大均值差异构建总核解决最优核构造问题.

·深度相关性对齐(CORrelation ALignment for Deep domain adaptation, Deep CORAL)[13]通过最小化源域和目标域协方差矩阵差异来提高网络迁移能力.

·领域对抗神经网络(Domain Adversarial Neural Networks, DANN)[17]将对抗学习的方法引入领域适应问题,提出梯度反转层优化最大化和最小化训练网络.

·对抗判别域适应(Adversarial Discriminative Domain Adaptation, ADDA)[32]使用特征提取器权重不共享的方式提取领域专属信息,提高特征的类别区分性.

·联合适配网络(Joint Adaptation Networks, JAN)[43]在训练过程中引入样本类别信息对齐不同域条件概率分布,从而提高域适应性能.

·条件领域对抗网络(Conditional Adversarial Domain Adaptation, CDAN, CDAN+E)[22]提出特征与分类器预测之间的协方差作为新特征学习判别模型.

·对比自适应网络(Contrastive Adaptation Network, CAN, iCAN)[28]根据网络预测分数对目标样本伪标签进行选择和重新加权,并将目标样本伪标签作为训练数据学习网络.

·球面空间域适应(DANN+R+E)[33]通过计算目标

样本特征间的相似性来选择伪标签,并将置信度高的伪标签引入网络学习.

·深度子域适应网络(Deep Subdomain Adaption Network,DSAN)[44]提出局部最大均值差异,度量源域和目标域相关子域分布,该方法能够捕获各类别细粒度信息.

4.4 实验结果与分析

评估指标:本文使用识别精度作为评价模型性能的标准,识别精度的定义为

$$\text{acurracy}=\frac{\left|\{x|x_j^t\in D_t\wedge f(x_j^t)=y_j\}\right|}{\left|D_t\right|} \tag{19}$$

式中,D_t表示目标域,$f(x_j^t)$和y_j分别为样本x_j^t的预测标签和真值标签,运算符$|\cdot|$表示集合中元素个数.

模型分类性能:数据集Office-31、Image-CLEF和Office-Home的模型识别精度如表1~表3所示,域适应任务表示为:源域→目标域.所有对比方法的结果均来自原文献,在3个数据集任务上报告了各域适应精度和平均精度.

所提算法DR-ADA在各数据集域适应任务上均取得了良好的识别效果.DDC、DAN、D-CORAL和DANN等基准方法未在域适应中引入样本类别信息,仅考虑对齐领域全局分布,易引起类别混淆降低特征区分性,表1~表3结果反映出其性能较弱.相比而言,DR-ADA利用伪标签信息减轻了领域间分布差异,引入后验概率度量同类别伪标签数据真实性,选择真实性高的伪标签样本指导域适应任务,通过伪标签回归损失减小目标域预测结果与高置信伪标签信息差异,从而更准确对齐领域间数据分布,显著提升域适应性能.

JAN、MADA、CDAN和CDAN+E等基准方法考虑了样本类别信息,能更准确对齐源域和目标域条件概率分布,在性能上获得改善.而本文方法相较其中性能最佳的CDAN+E,3个数据集平均识别精度进一步取得了2.5%、1.9%和3.7%的提升.不同于这些方法未对伪标签噪声采取抑制措施,DR-ADA在领域适应之前增加高斯均匀混合模型,无监督条件下检测相同伪标签目标样本特征距离数据,设置后验概率阈值对伪标签进行选择.分类性能提升表明所提方法对于伪标签噪声数据有抑制效果,能够减小目标样本伪标签中噪声对域适应模型带来的"负迁移"影响.

表1 Office-31上的识别精度　　单位:%

	方法	A→W	D→W	W→D	A→D	D→A	W→A	Avg.
通用	ResNet50[3]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
基于度量	DDC[11]	75.8	95.0	98.2	77.5	67.4	64.0	79.7
	DAN[12]	83.8	96.8	99.5	78.4	66.7	62.7	81.3
	D-CORAL[13]	77.7	97.6	99.7	81.1	64.6	64.0	80.8
	JAN[43]	85.4	97.4	99.8	84.7	68.6	70.0	84.3
基于对抗	ADDA[32]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
	DANN[17]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
	MADA[20]	90.0	97.4	99.6	87.8	70.3	66.4	85.2
	CDAN[22]	93.1	98.2	**100.0**	89.8	70.1	68.0	86.6
	CDAN+E[22]	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7
	CAN[28]	81.5	98.2	99.7	85.5	65.9	63.4	82.4
	iCAN[28]	92.5	98.8	**100.0**	90.1	72.1	69.9	87.2
	DANN+R+E[33]	—	—	—	—	—	—	89.8
	DR-ADA	**95.3**	**99.7**	**100.0**	**94.5**	**75.4**	**76.5**	**90.2**

基于伪标签的域适应方法,如iCAN和DANN+R+E不仅含有伪标签选择模块,并且在域适应中利用了样本伪标签学习模型,这些方法在多个基准数据集上取得了较好效果,如表1、表2所示.所提方法相比iCAN在Office-31和Image-CLEF数据集上平均分类准确度各提升了3%和2.2%.如表1~表3所示,与DANN+R+E相比所提方法在3个数据集上的平均分类准确度提高了0.4%、0.2%和0.7%.在伪标签选择上不同于这两种方法,DR-ADA引入异常值检测获得后验概率度量伪标签正确性,同时考虑目标样本特征与同类别子域中类均值特征间差异,为高斯均匀混合模型提供更准确建模信息.表4给出了VisDA-2017数据集不同模型对图像各类别识别精度.可以看出,本文方法DR-ADA在knife、person、skateboard和truck等4个类别中取得了所有对比模型中最优结果,并在总的平均识别精度中与DSAN同为最优.

表5为本文方法DR-ADA与DAN、DANN及DANN+R+E等在Office-31上运行时间比较.运行环境为GPU:NVIDIA GeForce RTX 3090,参数设置:batch为32,epochs为100.表中以10 epochs为运行阶段,统计每

表2 Image-CLEF上的识别精度 单位:%

	方法	I→P	P→I	I→C	C→I	C→P	P→C	Avg.
通用	ResNet50[3]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
基于度量	DDC[11]	74.6	85.7	91.1	83.2	68.3	88.8	81.8
	DAN[12]	75.0	86.2	93.3	84.1	69.8	91.3	83.3
	JAN[43]	76.8	88.0	94.7	89.5	74.2	91.7	85.8
	D-CORAL[13]	76.9	88.5	93.6	86.8	74.0	91.6	85.2
基于对抗	DANN[17]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
	MADA[20]	75.0	87.9	96.0	88.8	75.2	92.2	85.8
	CAN[28]	78.2	87.5	94.2	89.5	75.8	89.2	85.7
	iCAN[28]	79.5	89.7	94.7	89.9	78.5	92.0	87.4
	CDAN[22]	76.7	90.6	97.0	90.5	74.5	93.5	87.1
	CDAN+E[22]	77.7	90.7	**97.7**	91.3	74.2	94.3	87.7
	DANN+R+E[33]	—	—	—	—	—	—	89.4
	DR-ADA	**79.3**	**93.5**	97.5	**92.8**	**78.1**	**96.3**	**89.6**

表3 Office-Home上的识别精度 单位:%

	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.
通用	ResNet50[3]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
度量	DAN[12]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
	JAN[41]	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
对抗	DANN[17]	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
	CDAN[22]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
	CDAN+E[22]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	**56.7**	81.6	65.8
	DANN+R+E[33]	—	—	—	—	—	—	—	—	—	—	—	—	68.8
	DR-ADA	**55.0**	**77.4**	**80.7**	**65.8**	**73.0**	**74.3**	**64.0**	**52.3**	**81.7**	**71.8**	54.6	**83.1**	**69.5**

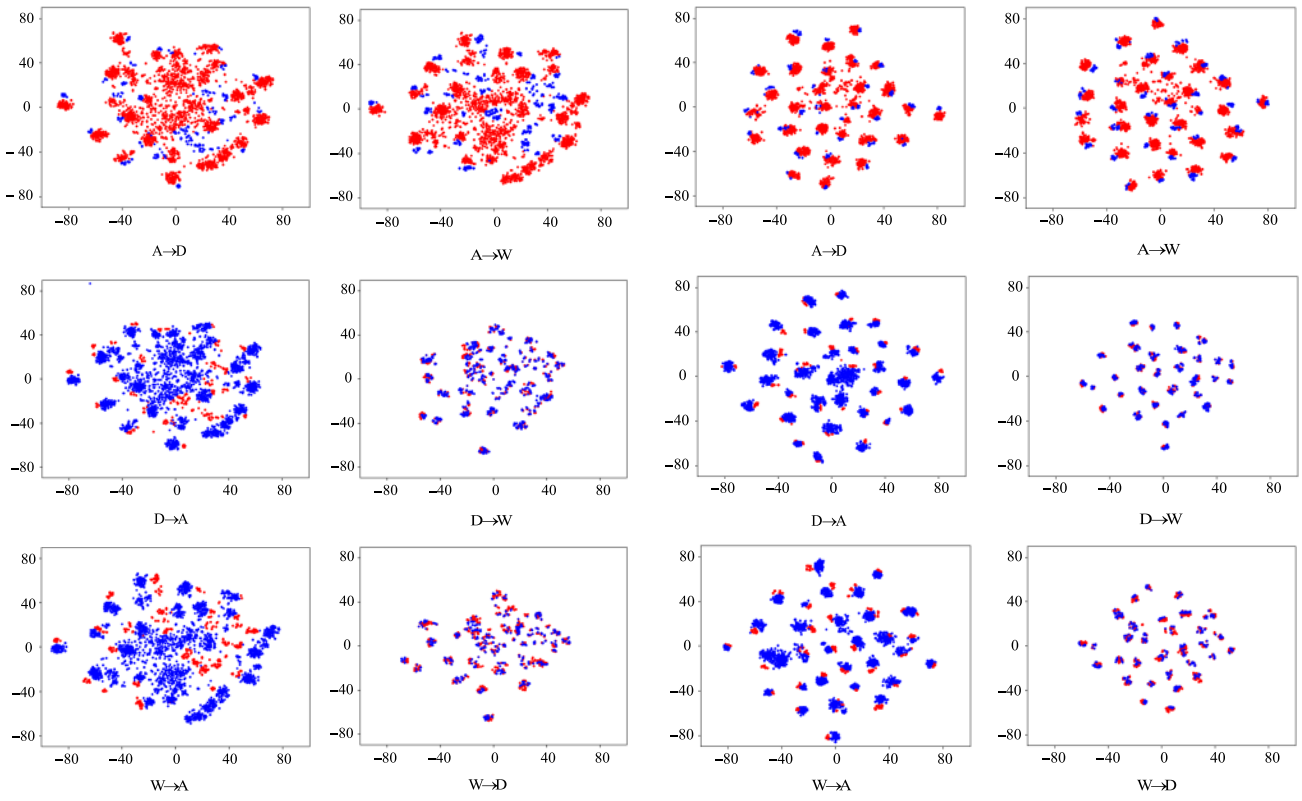
epoch的平均运行时间. 相比DAN及DANN,DR-ADA在运行时间上稍显劣势,这是因为其运算成本额外增加了高斯均匀混合模型计算复杂度这一部分,而相比DANN+R+E,DR-ADA在不同数据集的精度均领先,在运行时间上也取得了显著的优势,充分表明本文模型DR-ADA具有兼顾精度与效率的综合优势.

特征可视化分析:为直观呈现所提方法域适应效果,本文在Office-31数据集6个迁移任务上,利用t-SNE嵌入进行特征降维可视化. 图3(a)为ImageNet预训练ResNet-50网络提取的2048维特征在领域适应前的可视化结果. 图3(b)为本文DR-ADA模型提取的2 048维特征在领域适应后的结果. 图中红点表示源域样本特征,蓝点表示目标域样本特征. 由图3(a)可以看出,领域适应前存在同领域同类别样本特征分布发散、源域与目标域同类别样本特征分布距离大和领域间整体分布差异大的问题. 为此,本文在对抗域适应方法学习模型的基础上,引入特征异常检测与伪标签数据损失函数,使得领域适应后的源域和目标域同类别样本特征分布更为接近,并且同类别样本特征间的类内距离更小,不同类别样本特征类间距离更大,这表明所提方法可以学到具有类别区分性和领域不变性的特征表示,能够准确匹配不同领域的数据分布.

表4 VisDA-2017识别精度 单位:%

类别	ResNet50	DANN	DAN	DSAN	DR-ADA
airplane	72.3	81.9	68.1	**90.9**	88.5
bicycle	6.1	**77.7**	15.4	66.9	70.4
bus	63.4	**82.8**	76.5	75.7	72.7
car	**91.7**	44.3	87.0	62.4	77.6
horse	52.7	81.2	71.1	**88.9**	68.2
knife	7.9	29.5	48.9	77.0	**79.5**
motorcycle	80.1	65.1	82.3	**93.7**	78.6
person	5.6	28.6	51.5	75.1	**80.3**
plant	90.1	51.9	88.7	**92.8**	90.1
skateboard	18.5	54.6	33.2	67.6	**75.0**
train	78.1	82.8	88.9	**89.1**	75.9
truck	25.9	7.8	42.2	39.4	**48.4**
Avg	49.4	57.4	62.8	**75.1**	**75.1**

消融实验与收敛效果:为验证模型的收敛效果以及所提各项损失的有效性,在Office-31数据集A→W、A→D、W→A和D→A等迁移任务上均绘制了5 000次迭代训练中分类错误率变化的收敛曲线并进行了消融实验. 保持模型参数、结构等条件不变,在基础损失上增加条件熵损失、回归损失和伪标签数据上交叉熵损失比较分



(*a*) 领域适应前特征可视化　　(*b*) 领域适应后特征可视化

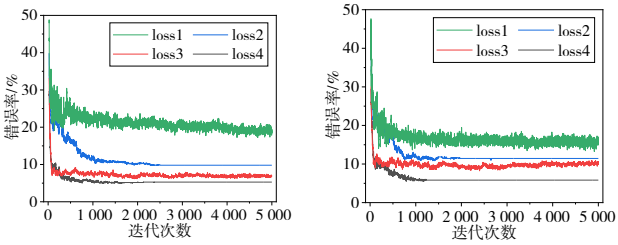
图 3　领域适应前后特征分布对比

表 5　Office-31 每 epoch 平均运行时间　　单位:%

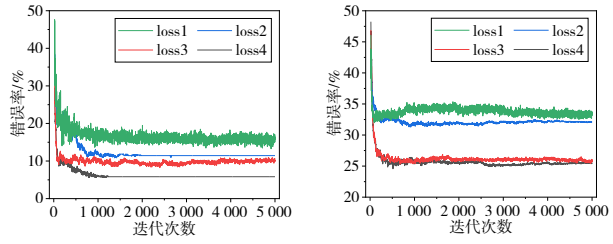
epochs	DANN	DAN	DANN+R+E	DR-ADA
1~10	50.2	23.4	125.2	85.4
11~20	51.5	25.6	121.8	85.3
21~30	53.4	24.3	120.1	87.1
31~40	48.3	22.9	120.5	85.6
41~50	52.6	23.2	123.0	81.3
51~60	47.2	21.5	123.7	80.7
61~70	50.1	24.4	121.3	82.5
71~80	52.6	20.6	120.6	85.3
81~90	51.7	22.4	124.2	81.9
91~100	48.4	21.4	121.0	84.2
Avg.	50.4	23.0	122.0	83.9

类性能,如图 4 所示. 其中 loss1 表示由式(2)和式(3)作为基础损失函数训练网络得到的分类错误率曲线,loss2 表示 loss1+式(14)作为损失函数得到的分类错误率曲线,loss3 表示 loss1+式(12)+式(15)作为损失函数得到的分类错误率曲线,loss4 表示总损失式(16)对应的分类错误率曲线. 可以看出,组合所有损失项的 loss4 收敛稳定且精度最佳. loss3 较基础损失收敛曲线 loss1 在各任务精度显著提升,显示出特征异常检测能够增强模型在伪标签噪声标签下的鲁棒性,基于目标样本伪标签设计的损失能够有效利用目标域伪标签额外匹配样本类别信息,提高目标域样本特征区分性. loss2 与 loss1 相比在各迁移任务上收敛精度均有不同程度提高,表明条件熵损失在训练模型时有助于降低分类错误率. 注意到 A→D 和 A→W 任务相较 D→A 和 W→A 分类错误率下降程度并不相同,这是由于后两个任务领域差异更大使得熵最小化方法对模型的分类性能提升效果降低.

参数敏感性: 为限制伪标签噪声对模型的影响,所提方法在式(15)中引入权重 λ,并对参与此项损失的目标数据设置了后验概率阈值 ω 条件. 在 Image-CLEF 中的 P→C,C→P,P→I 和 I→P 等 4 个迁移任务上对 λ 和 ω 参数取值进行网格搜索,如图 5 所示. 由于 P→I 和 P→C 领域间差异较小,超参数对模型分类性能影响也较小. 但在其他领域差异较大的任务上,如图 5(*a*),阈值 ω 设置在 0.8 左右取得最佳性能,准确率曲线逐步上升. 取值在 0.8 到 0.9 之间,3 个域适应任务上准确率下降. 这种先升后降的趋势是因为较小的阈值使得伪标签数据中包含较大比例噪声,从而降低了模型分类性能. 此外,过大的阈值虽然噪声含量低,但同时也舍弃了大部分伪标签数据,削弱了模型的知识迁移能力. 如图 5(*b*)所示,在



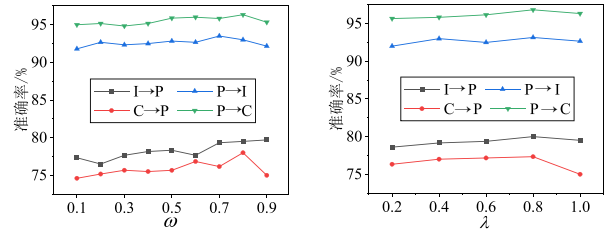
(a) A→W 收敛曲线 (b) A→D 收敛曲线



(c) W→A 收敛曲线 (d) D→A 收敛曲线

图4 不同损失下模型分类错误率收敛曲线

两个领域差异过大的任务 I→P 与 C→P 上，准确率随权重 λ 增加逐渐增大，权重区间在 0.8 到 1.0 时准确率下降，在域差异大的任务上分类器预测准确性较低导致参与训练的伪标签噪声含量高，而过大的取值容易增加伪标签噪声数据对模型的影响，从而引起负迁移.



(a) 阈值 ω 对分类准确率的影响 (b) 权重 λ 对分类准确率的影响

图5 阈值和权重对模型分类准确率的影响

在对分离异常样本的阈值进行选择时，较小或较大的阈值都会影响模型的性能. 较小阈值难以克服噪声干扰，而较大阈值会因丢弃大量伪标签数据而降低迁移能力，且在实际选择阈值时还要考虑不同数据集内部以及不同领域间差异性. 对于 Image-CLEF 数据集，由于 P→I 和 P→C 领域相近，阈值 ω 的选择对模型性能的影响较小，因此选取 C→P 与 I→P 的实验结果作为衡量阈值标准. 如图 5(a)，阈值在 0.8 附近时，分类准确率在两个任务综合表现最佳.

后验概率指标有效性：为了检验特征异常检测任务中目标样本后验概率是否有助于检测伪标签真实性，本文在 Office-31 的 A→W、A→D、D→A 和 W→A 等任务上报告了模型更新 10 轮时，不同后验概率阈值 ω 下的目标样本伪标签预测准确率，表示为后验概率大于阈值 ω 时伪标签为真的目标样本数量与后验概率大于阈值 ω 时的所有目标域样本数量之比. 如表 6 所示，随着 ω 增大伪标签准确率也进一步提升，表明所提特征异常检测方法能够利用高斯均匀混合模型检测特征距离异常的目标样本，通过后验概率指标有效度量伪标签正确性，且提高阈值能够去除错误标记的伪标签数据.

表 6 不同后验概率阈值的伪标签准确率 单位：%

ω	A→W		A→D	D→A	W→A
0.2	95.1±0.4	93.95±0.4	71.29±0.4		78.03±0.1
0.3	95.62±0.3	94.61±0.5	72.81±0.3		79.70±0.2
0.4	95.85±0.2	95.27±0.5	75.11±0.1		82.03±0.2
0.5	96.03±0.4	95.93±0.3	77.30±0.1		84.52±0.5
0.6	96.06±0.6	96.18±0.4	81.61±0.3		87.80±0.7
0.7	96.75±0.6	97.20±0.9	96.83±0.6		97.39±0.6
0.8	100±0.0	100±0.0	100±0.0		100±0.0

5 结论

对抗域适应是无监督领域适应的重要方向，可以有效缓解实际场景中目标任务监督信息不足的问题. 本文针对领域适应中目标域伪标签噪声问题以及如何提取具有判别性的类别级特征表示问题，构建了基于特征异常检测与伪标签回归的对抗域适应框架. 所提方法利用后验概率度量目标域样本伪标签正确性，并作为伪标签损失项乘子降低噪声标签带来的负迁移，提高域适应模型的目标域特征可区分性. 与相关主流基准方法进行对比，表明所提方法在域适应任务中取得了优越的性能. 此外，实验部分也从特征可视化、伪标签预测和损失函数有效性方面进行了系统的比较，证明本文所提特征异常检测方法能有效提取正确性高的目标域伪标签，所提关键损失进一步增强了模型在目标域上的特征表示能力.

参考文献

[1] 韩冲, 王俊丽, 吴雨茜, 等. 基于神经进化的深度学习模型研究综述[J]. 电子学报, 2021, 49(2): 372-379.
HAN C, WANG J L, WU Y X, et al. A review of deep learning models based on neuroevolution[J]. Acta Electronica Sinica, 2021, 49(2): 372-379. (in Chinese)

[2] XIANG X, ABDEIN R, LI W, et al. Deep scene flow learning: From 2D images to 3D point clouds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 185-208.

[3] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[4] MITSUZUMI Y, IRIE G, KIMURA A, et al. Phase randomization: A data augmentation for domain adaptation in human action recognition[J]. Pattern Recognition, 2024, 146(1): 1-11.

[5] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[6] PAN S J, TSANG I W, KWOK J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2010, 22(2): 199-210.

[7] AZIZZADENESHELI K. Importance weight estimation and generalization in domain adaptation under label shift[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6578-6584.

[8] 范苍宁, 刘鹏, 肖婷, 等. 深度域适应综述: 一般情况与复杂情况[J]. 自动化学报, 2021, 47(3): 515-548.
FAN C N, LIU P, XIAO T, et al. A review of deep domain adaptation: General situation and complex situation[J]. Acta Automatica Sinica, 2021, 47(3): 515-548. (in Chinese)

[9] 孙晨峰, 吕卫民, 戴洪德, 等. 一种基于TimeGAN和OCSVM的多元退化设备小子样数据增广方法[J]. 电子学报, 2022, 50(11): 2678-2687.
SUN C F, LÜ W M, DAI H D, et al. A small sample data augmentation method for multivariate degradation equipment based on TimeGAN and OCSVM[J]. Acta Electronica Sinica, 2022, 50(11): 2678-2687. (in Chinese)

[10] WANG W, LI H, DING Z, et al. Rethinking maximum mean discrepancy for visual domain adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(1): 264-277.

[11] TZENG E, HOFFMAN J, ZHANG N, et al. Deep domain confusion: Maximizing for domain invariance[EB/OL]. [2024]. https://arxiv.org/abs/1412.3474.

[12] LONG M, CAO Y, WANG J, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the International Conference on Machine Learning. New York: ILMS, 2015: 97-105.

[13] SUN B, SAENKO K. Deep CORAL: Correlation alignment for deep domain adaptation[C]//Proceedings of the European Conference on Computer Vision. Amsterdam: Springer, 2016: 443-450.

[14] ZELLINGER W, GRUBINGER T, LUGHOFER E, et al. Central moment discrepancy for domain-invariant representation learning [C]//Proceedings of the 5th International Conference on Learning Representations. Toulon: Springer, 2017: 1-11.

[15] ANDÉOL L, KAWAKAMI Y, WADA Y, et al. Learning domain invariant representations by joint Wasserstein distance minimization[J]. Neural Networks, 2023, 167(1): 233-243.

[16] 王格格, 郭涛, 余游, 等. 基于生成对抗网络的无监督域适应分类模型[J]. 电子学报, 2020, 48(6): 1190-1197.
WANG G G, GUO T, YU Y, et al. Unsupervised domain adaptation classification model based on generative adversarial network[J]. Acta Electronica Sinica, 2020, 48(6): 1190-1197. (in Chinese)

[17] GANIN Y, LEMPITSKY V S. Unsupervised domain adaptation by backpropagation[C]//International Conference on Machine Learning. Lille: IMLS, 2015: 1180-1189.

[18] ISMAEL S F, KAYABOL K, APTOULA E. Unsupervised domain adaptation for the semantic segmentation of remote sensing images via one-shot image-to-image translation[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20 (1): 1-12.

[19] ZHAO S, YUE X, ZHANG S, et al. A review of single-source deep unsupervised visual domain adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2): 473-493.

[20] PEI Z, CAO Z, LONG M. Multi-adversarial domain adaptation[C]//Proceedings of the Thirty-Second Conference on Artificial Intelligence. Louisiana: AAAI, 2018: 3934-3941.

[21] SAFFARI M, KHODAYAR M. Low-rank sparse generative adversarial unsupervised domain adaptation for multi-target traffic scene semantic segmentation[J]. IEEE Transactions on Industrial Informatics, 2024, 20(2): 2564-2576.

[22] LONG M, CAO Z, WANG J, et al. Conditional adversarial domain adaptation[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. New York: ACM, 2018: 1647-1657.

[23] DAN J, JIN T, CHI H, et al. Trust-aware conditional adversarial domain adaptation with feature norm alignment[J]. Neural Networks, 2023, 168(1): 518-530.

[24] CHO J, KIM D, JUNG Y, et al. MCDAL: Maximum classifier discrepancy for active learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8753-8763.

[25] SAITO K, KIM D, SCLAROFF S, et al. Semi-supervised domain adaptation via minimax entropy[C]//IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 8050-8058.

[26] SAITO K, USHIKU Y, HARADA T. Asymmetric tri-training for unsupervised domain adaptation[C]//Proceedings of the International Conference on Machine Learning. New York: IMLS, 2017: 2988-2997.

[27] XIE B, LI S, LI M, et al. SePiCo: Semantic-guided pixel contrast for domain adaptive semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 9004-9021.

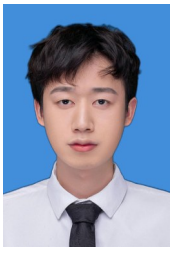
[28] ZHANG W, OUYANG W, LI W, et al. Collaborative and adversarial network for unsupervised domain adaptation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 3801-3809.

[29] ZHANG W, XU D, OUYANG W, et al. Self-paced collaborative and adversarial network for unsupervised domain

adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(6): 2047-2061.
[30] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.
[31] HARARY S, SCHWARTZ E, ARBELLE A, et al. Unsupervised domain generalization by learning a bridge across domains[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5270-5280.
[32] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2962-2971.
[33] GU X, SUN J, XU Z. Spherical space domain adaptation with robust pseudo-label loss[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 9098-9107.
[34] CHEN S, JIA X, HE J, et al. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 11013-11022.
[35] LATHUILIÈRE S, MESEJO P, ALAMEDA-PINEDA X, et al. Deepgum: Learning deep robust regression with a Gaussian-uniform mixture model[C]//European Conference on Computer Vision. Cham: Springer, 2018: 202-217.
[36] HUANG J, XIAO N, ZHAN L. Balancing transferability and discriminability for unsupervised domain adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5807-5814.
[37] ZHANG J, DING Z, LI W, et al. Importance weighted adversarial nets for partial domain adaptation[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8156-8164.
[38] MA A, LI J, LU K, et al. Adversarial entropy optimization for unsupervised domain adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6263-6274.
[39] SAENKO K, KULIS B, FRITZ M. Adapting visual category models to new domains[C]//European Conference on Computer Vision. Berlin: Springer, 2010: 213-226.
[40] WANG S, ZHANG L, ZUO W, et al. Class-specific reconstruction transfer learning for visual recognition across domains[J]. IEEE Transactions on Image Processing, 2019, 29(1): 2424-2438.
[41] VENKATESWARA H, EUSEBIO J, CHAKRABORTY S, et al. Deep hashing network for unsupervised domain adaptation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5018-5027.
[42] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[43] LONG M, ZHU H, WANG J, et al. Deep transfer learning with joint adaptation networks[C]//Proceedings of the International Conference on Machine Learning. New York: IMLS, 2017: 2208-2217.
[44] ZHU Y, ZHUANG F, WANG J, et al. Deep subdomain adaptation network for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4): 1713-1722.

作者简介

潘 杰 男，1986年12月出生于江苏省徐州市. 现为中国矿业大学信息与控制工程学院副教授、硕士生导师. 主要研究方向为机器学习与网络生理学.
E-mail: panjie1616@cumt.edu.cn



刘 波 男，1997年10月出生于湖南省邵阳市. 2023年毕业于中国矿业大学信息与控制工程学院. 现为中国电信股份有限公司长沙分公司工程师. 从事算法设计与云网设施维护工作.
E-mail: liubo7556@163.com

邹筱瑜 女，1990年5月出生于四川省自贡市. 现为中国矿业大学机电工程学院副教授、硕士生导师. 主要研究方向为机器学习与装备运维.
E-mail: zouxiaoyu@cumt.edu.cn