智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向自闭症辅助诊断的知识蒸馏混合域适应方法

顿家乐, 王骏, 彭汉琛, 李俊诚, 施俊

引用本文:
顿家乐, 王骏, 彭汉琛, 等. 面向自闭症辅助诊断的知识蒸馏混合域适应方法[J]. 智能系统学报, 2025, 20(1): 81-90.
DUN Jiale, WANG Jun, PENG Hanchen, et al. Blended domain adaptation for computer-aided diagnosis of autism through knowledge distillation[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 81-90.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202403030

DOI: 10.11992/tis.202403030

面向自闭症辅助诊断的知识蒸馏混合域适应方法

顿家乐，王骏，彭汉琛，李俊诚，施俊
（上海大学 通信与信息工程学院，上海 200444）

摘　要：使用领域自适应方法构建自闭症辅助诊断模型时，通常会面临目标域中混合了来自多个影像中心的样本的情况（即混合目标域），这使得目标域中包含了多个分布。传统领域自适应方法只能处理目标域包含单一分布的情况，而无法直接处理混合目标域的情况。为此，本文提出了一种基于知识蒸馏的混合目标领域自适应模型。具体地，将图卷积网络（graph convolutional network, GCN）作为教师模型，多层感知机（multilayer perceptron, MLP）作为学生模型。针对混合目标域数据分布的多样性，提出了一种新型的对抗知识蒸馏机制，通过对抗训练特征提取器和域鉴别器来减少源域和目标域之间的分布差异；与此同时，使用知识蒸馏，使教师模型在领域自适应的同时将知识传递给学生模型。在 ABIDE 数据集上验证了算法的有效性，本文方法一方面有效降低了网络的复杂度，另一方面，在混合目标域的分类准确率达到 69.17%，与其他领域自适应方法相比效果更好。

关键词：自闭症谱系障碍；领域自适应；混合目标域；知识蒸馏；图卷积网络；教师网络；学生网络；对抗学习
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1673−4785(2025)01−0081−10

中文引用格式：顿家乐，王骏，彭汉琛，等. 面向自闭症辅助诊断的知识蒸馏混合域适应方法 [J]. 智能系统学报，2025，20(1)：81–90.
英文引用格式：DUN Jiale, WANG Jun, PENG Hanchen, et al. Blended domain adaptation for computer-aided diagnosis of autism through knowledge distillation[J]. CAAI transactions on intelligent systems, 2025, 20(1): 81–90.

Blended domain adaptation for computer-aided diagnosis of autism through knowledge distillation

DUN Jiale, WANG Jun, PENG Hanchen, LI Juncheng, SHI Jun
(School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: In the modeling of computer-aided diagnosis for autism spectrum disorder (ASD) across multiple centers with domain adaptation methods, unlabeled samples from multiple imaging centers are blended together in the target domain. Traditional domain adaptation methods lack the capability to address the clinical scenario of identifying ASD in blended-target domains. To this end, we propose a knowledge distillation blended-target domain adaptation model. Specifically, the graph convolutional network (GCN) is used as the teacher model and the multilayer perceptron (MLP) is used as the student model. To address distribution differences between source and target domains, a novel adversarial knowledge distillation mechanism is proposed to reduce the distribution difference by adversarially training feature extractors and domain discriminators. At the same time, knowledge distillation is used to enable the teacher model to transfer knowledge to the student model while achieving domain adaptation. The ABIDE dataset is employed to validate the effectiveness of the model. Our method not only reduces the complexity of the network but also achieves a classification accuracy of 69.17% in the blended target domains, surpassing other domain adaptation methods.

Keywords: autism spectrum disorder; domain adaptation; blended target domain; knowledge distillation; graph convolutional network; teacher network; student network; adversarial learning

收稿日期：2024−03−18.
基金项目：国家自然科学基金项目（62272289）.
通信作者：王骏. E-mail: wangjun_shu@shu.edu.cn.

自闭症谱系障碍（autism spectrum disorder, ASD）指的是一系列以社交交流缺陷、狭隘兴趣

和重复刻板行为为主要临床表现的复杂神经发育障碍[1]，它给家庭和社会带来了沉重的负担。研究表明，与正常人群相比，ASD 患者的大脑活动存在着较为明显的异常。静息态功能磁共振成像（resting-state functional magnetic resonance imaging, rs-fMRI）通过检测受试者在静息状态时血氧水平依赖（blood oxygenation-level dependent, BOLD）信号来反映大脑的活动情况，为 ASD 的诊断提供了大量神经影像依据。最近，多中心 rs-fMRI 数据已成功用于构建 ASD 分类模型[2-4]。由于各中心在数据的采集设备与方式、招募策略等方面各不相同[5]，所以各中心之间的数据分布存在差异，因此直接将多中心数据组合在一起构建 ASD 分类模型不适合。

为了缓解多个中心之间的数据分布差异，领域自适应（domain adaptation, DA）方法得到了广泛研究[6-7]。其基本思想是将分布不同的源域和目标域的数据映射至一个公共特征空间中，使它们在此空间中的分布尽可能接近。无监督领域自适应（unsupervised domain adaption, UDA）方法通过学习域不变表示，将知识从标记样本（源域）转移到未标记样本（目标域）[8-9]。目前，研究人员提出了多种 UDA 方法，经典的方法有 DAN（deep adaptation network）[10]、LRCDR（low rank and class discriminative representation）[3]、LRDAIC（low-rank domain adaptive method with inter-class）[11]、maLRR（multi-site adaption framework via low-rank representation decomposition）[4] 等。这些方法利用来自单一有标记的源域数据，通过度量学习和对抗学习等方法学习源域与目标域之间的域不变表示。然而，这些研究中的目标域仅包含来自单个领域（即数据中心）的数据。在临床中，目标域中可能混合了来自多个数据中心的数据，这使得已有的 UDA 方法不能直接使用。本文将其作为混合目标领域自适应（blended-target domain adaptation, BTDA）问题进行研究[12]。

BTDA 将在单个源域上学习到的模型扩展到混合目标域（目标域中包含了来自不同分布的多个域的数据），充分利用来自单个已标记的源域和多个未标记的目标域的数据，使学习器在混合目标域上表现良好[13-14]。当前已有关于 BTDA 方法的研究。例如，Chen 等[12] 提出了 AMEAN（adversarial meta-adaptation networks），通过结合元学习器并引入对抗性适应损失，实现 BTDA；Xu 等[15] 提出了 MCDA（mutual conditional blended-target domain adaptation），采用基于不确定性的分类域判别器在混合特征空间中对齐分类分布，以缓解混合目标域的分类器偏差，从而实现 BTDA；Roy 等[16] 提出了 CGCT（curriculum graph co-teaching），通过协同学习来聚合跨域的相似样本特征，实现在混合目标域上的领域自适应。尽管这些方法可以把源域上的模型泛化到混合目标域，但由于 ASD 样本之间存在明显的异质性（不同中心受试者之间存在的各种差异），且混合目标域中样本的分布差异较大，使得现有的这些 BTDA 方法难以直接应用于多中心 ASD 辅助诊断建模。

另一方面，图卷积网络（graph convolutional network, GCN）作为一种基于图结构的深度学习模型，能够对患者之间的关系进行建模，近年来已经在 ASD 分类任务中得到了应用[17]。GCN 在训练阶段能够学习到样本之间复杂的关系和上下文信息，从而为构建分类器提供更多的知识。然而，现有的 GCN 方法在处理 ASD 分类任务时也存在明显不足：首先，GCN 的计算复杂度高，特别是在处理 ASD 样本之间的复杂关系时，计算量会显著增加；其次，模型推理依赖全局图结构，在测试阶段模型对未知样本进行分类时，需要将其加入图结构，这会导致图结构发生变化，从而需要重新训练网络。这些不足导致 GCN 对混合目标域中未知样本的分类能力有所下降。

为了解决这些问题，本文将领域自适应与知识蒸馏（knowledge distillation, KD）相结合[18]，提出一种新型的基于知识蒸馏的混合目标领域自适应（knowledge distillation blended-target domain adaptation, KD-BTDA）模型。一方面，通过引入知识蒸馏，将复杂的 GCN（教师模型）中的知识迁移到更简单的模型（学生模型）中[19-20]，使模型在推理时避免了对全局图结构的依赖，无需重新调整网络参数，从而有效降低计算成本和复杂度；另一方面，在知识蒸馏的同时引入对抗学习，使模型能够更好地适应混合目标域分布的多样性，从而增强其泛化能力。

本文创新如下：

1）针对目标域中混合多个影像中心的样本的情况，提出一种新颖的基于对抗学习的知识蒸馏方法，解决 BTDA 中自闭症分类困难的问题；

2）针对混合目标域中因自闭症异质性而导致的网络复杂问题，引入知识蒸馏，将 GCN 作为教师模型，通过学习自闭症样本之间的相互关系指导学生模型训练，将知识迁移到学生模型中，降低模型的复杂度，同时还避免了图卷积网络在测试阶段需要重新训练网络的问题；

3）针对混合目标域中样本分布差异较大的问题，本文在知识蒸馏中结合对抗学习，提出了一种对抗知识蒸馏机制来提高教师模型在混合目标域上的泛化能力。通过对抗训练使教师模型能够适应混合目标域中数据分布的多样性，同时将经过对抗增强后的教师模型中的知识传递给学生模型，提升学生模型的分类能力。

1　本文方法

针对混合目标域场景下的 ASD 分类问题，令 $\boldsymbol{x}_i^s$ 为源域样本 i 经预处理后的特征表示，$\boldsymbol{y}_i^s$ 表示 $\boldsymbol{x}_i^s$ 的标签，$\boldsymbol{x}_j^t$ 为目标域样本 j 经预处理后的特征表示，N_s 表示源域样本数量，N_t 表示混合目标域中的样本总量；给定一个有标记的源域数据集合 $S=\{(\boldsymbol{x}_i^s,y_i^s)\}_{i=1}^{N_s}$ 和混合目标域集合 $T=\{(\boldsymbol{x}_j^t)\}_{j=1}^{N_t}$，其中 T 包含 Q 个未标记的目标域，即 $T_1,T_2,\cdots,T_Q$。与传统领域自适应任务中的基本假设相似，假设源域和各目标域中数据分布不一致[21]。BTDA 的目标是训练一个模型，可以有效地对混合目标域中的所有数据进行分类。

1.1　总体网络架构

本文所提出的 KD-BTDA 模型如图 1 所示，网络结构包括特征提取器 F、教师网络 G、学生网络 H、域鉴别器 D 和知识蒸馏模块。本文在训练过程中，一方面考虑源域到混合目标域的领域自适应损失 $\mathcal{L}_{\mathrm{DA}}$，另一方面考虑教师网络和学生网络的特征输入知识蒸馏模块后产生的知识蒸馏损失 $\mathcal{L}_{\mathrm{KD}}$。

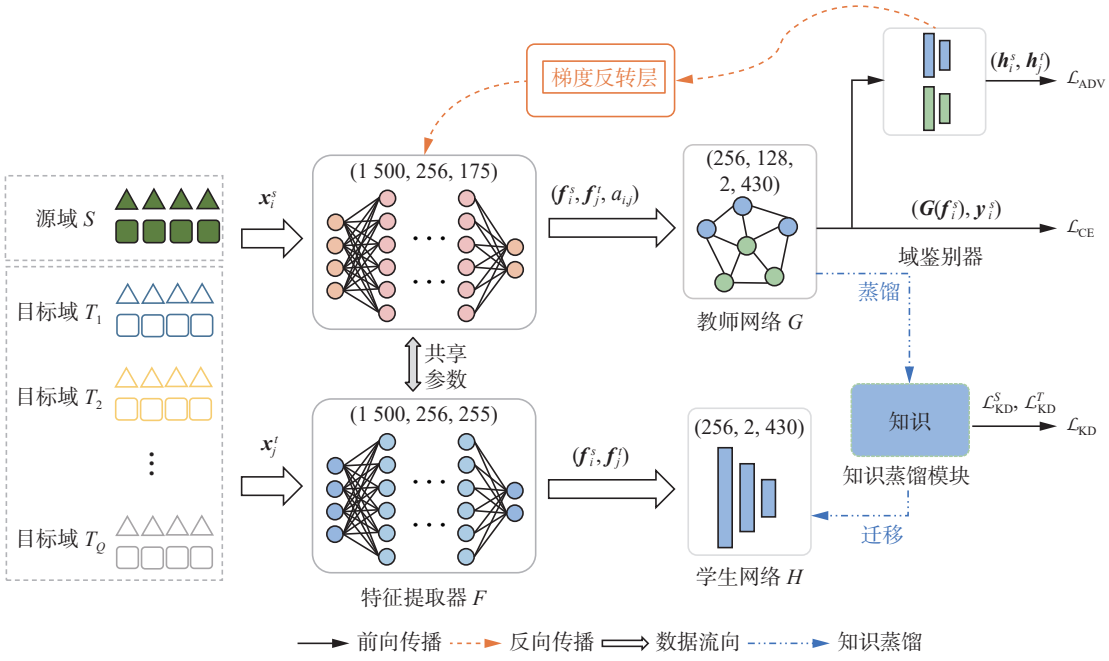

图 1　基于知识蒸馏的混合目标领域自适应网络模型框架

Fig. 1　Framework for knowledge distillation blended-target domain adaptation network

1.2　教师和学生网络框架

由于 GCN 能够有效地捕捉样本之间的复杂关系和上下文信息，本文选择 GCN 作为教师网络主干，对源域和混合目标域的样本特征进行聚合[22]。GCN 通过图卷积操作来聚合源域和混合目标域中样本及其邻接节点的特征，生成更具代表性的节点特征表示，为后续的知识蒸馏和领域自适应提供支持[17]。学生网络采用结构较为简单的多层感知机（multilayer perceptron, MLP）作为分类器。

经预处理后的源域样本特征 $\boldsymbol{x}_i^s$ 和混合目标域样本特征 $\boldsymbol{x}_j^t$ 送入特征提取器 F 中，得到源域特征 $\boldsymbol{f}_i^s=F(\boldsymbol{x}_i^s)$ 和目标域特征 $\boldsymbol{f}_j^t=F(\boldsymbol{x}_j^t)$，然后将源域与混合目标域中的所有样本构成一个无向全连接图 $\Gamma=(\mathcal{V},\mathcal{E},\mathcal{A})$。其中节点 $v_i\in\mathcal{V}$ 表示源域和混合目标域中的样本，节点 v_i 的特征向量为 $\boldsymbol{f}_i$；$e_{i,j}\in\mathcal{E}$ 表示节点 v_i 和 v_j 之间的边；$a_{i,j}\in\mathcal{A}$ 表示图中节点 v_i 和 v_j 之间的关系。本文选择根据节点之间的相似性来构造节点之间边的权重：

$$\mathrm{Sim}(\boldsymbol{f}_i,\boldsymbol{f}_j)=\exp\left(-\frac{\rho^2(\boldsymbol{f}_i,\boldsymbol{f}_j)}{2\sigma^2}\right)\tag{1}$$

式中：$\rho(\cdot)$ 是欧氏距离函数，σ 是核宽度。令 $a_{i,j}=\mathrm{Sim}(\boldsymbol{f}_i,\boldsymbol{f}_j)$，在网络前向传播过程中，更新节点特征的计算公式为

$$v_i^{(l)}=G^{(l)}\left(\boldsymbol{f}_i^{(l-1)},\sum_{j\in B(i)}a_{i,j}\cdot\boldsymbol{f}_j^{(l-1)}\right)\tag{2}$$

式中：l表示层次深度，$G^{(l)}$表示 GCN 的第l层，$B(i)$包含当前节点i的所有邻居节点，$\boldsymbol{f}_i^{(l-1)}$和$\boldsymbol{f}_j^{(l-1)}$分别是第$l-1$层节点v_i和邻接节点v_j的特征表示。式(2)描述了节点v_i的第l层特征更新，综合考虑了当前节点和其邻居的信息。在教师网络中，通过领域自适应部分迭代更新节点特征，捕捉图结构的复杂关系，使模型更好地适应混合目标域的数据分布。

1.3　基于对抗学习的知识蒸馏

本节通过知识蒸馏，将知识从教师网络蒸馏到学生网络；同时在知识蒸馏时引入对抗学习，从而解决源域到混合目标域领域自适应的过程中目标域样本分布多样性的挑战。

1.3.1　知识蒸馏

ASD 样本之间差异大，关系复杂且异质性强，为此，本文采用 GCN 进行特征学习。但是，GCN 计算复杂度高，因此本文通过知识蒸馏将知识从 GCN 迁移到 MLP 网络中。具体地，将 GCN 作为教师模型，将 MLP 作为学生模型。这一方面降低了预测时网络的复杂程度，另一方面，由于 MLP(学生模型)不依赖图结构，因此能够在无需重新训练网络的情况下对未见过的样本进行分类。图 2 给出了基于对抗学习的知识蒸馏的过程。

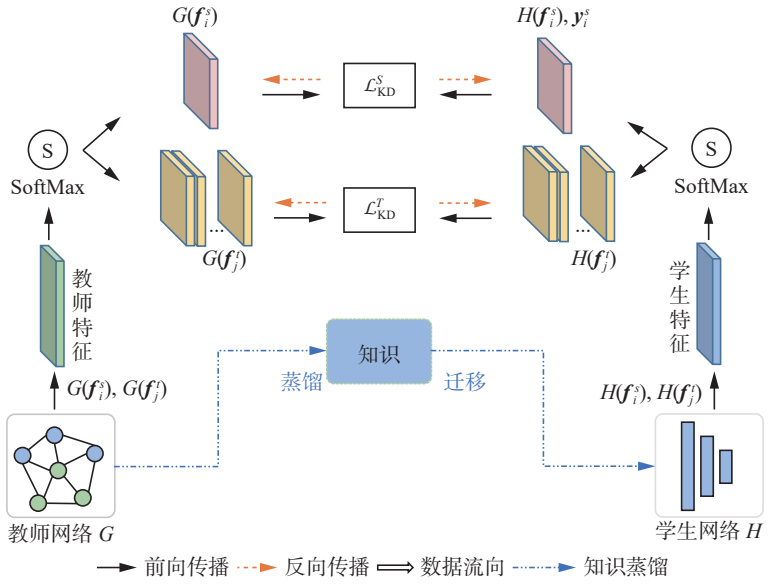

图 2　教师网络对学生网络的知识蒸馏

Fig. 2　Knowledge distillation from teacher network to student network

在源域中，源域特征$\boldsymbol{f}_i^s$同时被送入教师网络和学生网络中，获取的网络输出作为教师特征和学生特征。基于 KL 散度(Kullback-Leibler divergence)，构造源域的蒸馏损失为

$$\mathcal{L}_{\mathrm{KD}}^S = -\frac{1}{N_s}\sum_{\boldsymbol{f}_i^s,\boldsymbol{y}_i^s\in S}\alpha\cdot D_{\mathrm{KL}}(G(\boldsymbol{f}_i^s),H(\boldsymbol{f}_i^s)) - \frac{1}{N_s}\sum_{\boldsymbol{f}_i^s,\boldsymbol{y}_i^s\in S}(1-\alpha)\cdot \mathrm{CE}(H(\boldsymbol{f}_i^s),\boldsymbol{y}_i^s) \tag{3}$$

式中：G是作为教师网络的 GCN；H是作为学生网络的 MLP；$D_{\mathrm{KL}}(\cdot,\cdot)$是 KL 散度损失函数；$\mathrm{CE}(\cdot,\cdot)$是交叉熵损失函数；$\alpha$是超参数，表示交叉熵损失的重要性。SoftMax 函数中，τ是温度超参数。该损失函数同时考虑了教师和学生网络之间的 KL 散度和源域上的交叉熵损失。

与源域相似，构造目标域上的蒸馏损失为

$$\mathcal{L}_{\mathrm{KD}}^T = -\frac{1}{N_t}\sum_{\boldsymbol{f}_j^t\in T} D_{\mathrm{KL}}\left(G\left(\boldsymbol{f}_j^t\right),H\left(\boldsymbol{f}_j^t\right)\right) \tag{4}$$

最终得到知识蒸馏的损失为

$$\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{KD}}^S + \mathcal{L}_{\mathrm{KD}}^T \tag{5}$$

1.3.2　对抗学习

为了减小源域和混合目标域中的数据分布差异，本文引入条件域对抗网络(conditional domain adversarial network, CDAN)[23]进行领域自适应，使源域和混合目标域数据的分布尽可能相同或相近。具体地，令$G(\boldsymbol{f}_i^s)$、$G\left(\boldsymbol{f}_j^t\right)$表示样本经过教师网络$G$的预测结果。在源域上，定义交叉熵损失为

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N_s}\sum_{\boldsymbol{f}_i^s,\boldsymbol{y}_i^s\in S}\mathrm{CE}(G(\boldsymbol{f}_i^s),\boldsymbol{y}_i^s) \tag{6}$$

定义域对抗的损失函数公式为

$$\mathcal{L}_{\mathrm{ADV}} = -\frac{1}{N_s}\sum_{\boldsymbol{f}_i^s\in S}\log(D(\boldsymbol{h}_i^s)) - \frac{1}{N_t}\sum_{\boldsymbol{f}_j^t\in T}\log\left(1-D\left(\boldsymbol{h}_j^t\right)\right) \tag{7}$$

式中D是域鉴别器。域鉴别器的输入是联合变量$\boldsymbol{h}_i^s=(\boldsymbol{f}_i^s,G(\boldsymbol{f}_i^s))$和$\boldsymbol{h}_j^t=\left(\boldsymbol{f}_j^t,G\left(\boldsymbol{f}_j^t\right)\right)$，将联合变量$\boldsymbol{h}$输

入域鉴别器D来判别输入样本是来自源域还是混合目标域。最小化$\mathcal{L}_{\mathrm{CE}}$有助于分类器$G$学习到源域数据的分布, 最大化$\mathcal{L}_{\mathrm{ADV}}$有助于优化域鉴别器$D$, 使其难以准确判断样本来自源域还是目标域。基于对抗学习的思想, 构造领域自适应损失为

$$\mathcal{L}_{\mathrm{DA}} = \mathcal{L}_{\mathrm{CE}} - \lambda_{\mathrm{ADV}}\mathcal{L}_{\mathrm{ADV}} \tag{8}$$

式中λ_{ADV}是超参数。通过最小化领域自适应损失$\mathcal{L}_{\mathrm{DA}}$, 可以实现在教师网络上从源域到混合目标域的领域自适应。

1.3.3 总的目标函数

综合考虑知识蒸馏与对抗学习, KD-BTDA 模型的总损失函数$\mathcal{L}_{\mathrm{TOTAL}}$由两部分组成, 一部分为领域自适应的损失$\mathcal{L}_{\mathrm{DA}}$, 另一部分为源域和目标域的蒸馏损失$\mathcal{L}_{\mathrm{KD}}$, 最终得到基于对抗学习的知识蒸馏的目标函数为

$$\mathcal{L}_{\mathrm{TOTAL}} = \mathcal{L}_{\mathrm{DA}} + \beta \cdot \mathcal{L}_{\mathrm{KD}} \tag{9}$$

式中超参数β平衡领域自适应损失和蒸馏损失。

训练过程以伪代码形式给出, 如算法 1 所示。

算法 1　基于知识蒸馏的混合目标领域自适应模型训练过程。

输入　源域$S = \{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\}_{i=1}^{N_s}$, 混合目标域$T = T_1, T_2, \cdots, T_Q$, 其中$T_Q = \{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$, 迭代次数$E$, 根据式(1)计算得到的边权重$a_{i,j}$。

输出　训练完成的网络。

预训练阶段

在源域S上使用交叉熵损失预训练教师网络G和学生网络H。

基于对抗学习的知识蒸馏阶段

1) for $e = 0; e \leqslant E$

2) 将源域样本$\boldsymbol{x}_i^s$和混合目标域样本$\boldsymbol{x}_j^t$送入特征提取器F得到特征$\boldsymbol{f}_i^s$、$\boldsymbol{f}_j^t$;

3) 根据式(2)更新节点特征$\boldsymbol{f}_i$;

4) 根据式(3)和(4)计算教师和学生特征之间的 KL 散度以及学生网络的交叉熵, 得到源域和目标域的蒸馏损失$\mathcal{L}_{\mathrm{KD}}^S$、$\mathcal{L}_{\mathrm{KD}}^T$;

5) 根据式(6)和(7)计算教师网络的交叉熵损失$\mathcal{L}_{\mathrm{CE}}$和域对抗损失$\mathcal{L}_{\mathrm{ADV}}$, 得到领域自适应损失$\mathcal{L}_{\mathrm{DA}}$;

6) 根据式(9)计算总损失$\mathcal{L}_{\mathrm{TOTAL}}$;

7) 使用反向传播算法, 反向更新网络参数;

8) end

2 实验结果与分析

2.1 ABIDE 数据集

本文采用 ABIDE(autism brain imaging data exchange)数据库[24] 来验证所提方法的有效性。实验中采用了 ABIDE 中 5 个成像中心的 rs-fMRI 数据, 人口统计学信息如表 1 所示。本文实验选择 NYU 成像中心作为源域, 而将其余 4 个成像中心的数据混合在一起作为混合目标域。

表 1　ABIDE 中 5 个成像中心样本的人口统计学信息

Table 1　Demographic information of the subjects from five sites in the ABIDE dataset

领域	成像中心	样本数	ASD患者/健康对照者数
源域	NYU	175	75/100
目标域	PITT	56	29/27
	UCLA_1	72	41/31
	USM	71	46/25
	YALE	56	28/28
	混合	255	144/111

2.2 数据预处理

本实验采用 C-PAC(configurable pipeline for the analysis of connectomes)流程, 使用 DPARSF 工具箱对 rs-fMRI 数据进行预处理[25]。具体来说, 首先将 rs-fMRI 数据与 HO(Harvard-Oxford)模板对齐, 将所有脑区划分为 111 个感兴趣区域(region of interest, ROI), 计算两个大脑区域的 BOLD 信号之间的皮尔逊相关性作为功能连接(functional connectivity, FC), 并获得 FC 的 111×111 相关矩阵。由于 FC 矩阵是对称矩阵, 本文选择提取其上三角部分, 并将其展平成一个 6 105 维的特征向量, 这个高维特征向量包含了不同脑区之间的复杂关系。

接着, 基于源域数据进一步进行特征选择。具体地, 采用递归特征消除(recursive feature elimination, RFE)方法, 从源域数据中的 6 105 维数据中提取了 1 500 个 FC 特征, 并记录了这些特征的索引。然后, 根据这些索引从目标域中提取相应的特征, 将它们和源域 FC 特征一同作为模型的输入。

2.3 实验细节

在实验中, 本文选用初始学习率$r_1 = 0.000\,5$的指数衰减随机梯度下降(SGD)优化器来进行网络训练。具体地, 设置动量为$M = 0.9$, 权重衰减为$D_{\mathrm{Weight}} = 0.000\,5$。

本文采用 10 折交叉验证来充分评估各方法的性能, 使用训练好的模型分别对 4 个目标域进行测试。在每次交叉验证中, 将每个目标域划分为 10 个子集, 这些子集在样本数量和 ASD 患者与健康对照的数量上保持相近。在每折中, 将源

域的所有样本作为有标签的训练集，而在每个独立的目标域中，选择 9 个无标签的子集作为训练集，剩下 1 个子集作为测试集。需要强调的是，在训练过程中不使用目标域标签，仅在测试中使用目标域标签来评估模型在目标域上的分类性能。对于不同方法使用一致的数据划分以保证交叉验证的公平。

2.4　消融实验

为了研究不同网络作为教师模型与学生模型对性能的影响，本文通过选择不同组合的教师模型和学生模型进行消融实验。本文选择以下 3 种配置进行消融实验：

1) 教师模型 MLP，学生模型 MLP：选择 MLP 作为教师模型，同时选择另一个结构和参数设置相同的 MLP 作为学生网络。

2) 教师模型 MLP，学生模型 GCN：选择 MLP 作为教师模型，GCN 作为学生模型。

3) 教师模型 GCN，学生模型 GCN：选择 GCN 作为教师模型，同时选择另一个结构和参数设置相同的 GCN 作为学生模型。

以上 3 种配置中，MLP 和 GCN 的网络结构和参数设置均与本文方法保持一致。本文在混合目标域上进行了消融实验，表 2 给出了混合目标域上的预测结果（表格中加粗的数据为最佳结果）。

表 2　不同教师与学生模型进行 ASD 分类的结果对比

Table 2　Comparison of results in ASD classification using different teacher and student models　%

教师模型	学生模型	目标域				平均准确率
		PITT	UCLA_1	USM	YALE	
MLP	MLP	70.15±1.61	68.75±0.74	59.19±0.74	70.32±0.97	67.10±0.50
MLP	GCN	64.29±1.88	68.64±3.15	**62.82**±2.32	71.25±1.32	66.75±1.13
GCN	GCN	63.75±2.53	64.03±8.32	58.45±4.80	66.96±5.21	63.30±4.31
GCN	MLP	**71.13**±1.10	**69.96**±0.65	59.80±0.79	**74.00**±0.27	**68.73**±0.39

通过实验可以发现，本文方法明显优于其他方法，在教师与学生模型均采用 MLP 的情况下，实验结果仅次于本文方法。然而，当把 GCN 作为学生模型时，实验结果普遍较差。这一观察结果强调在知识蒸馏中，需要在选择适当的教师和学生模型之间取得平衡，学生模型结构的复杂性可能会影响网络的分类性能[26]。本文选择从复杂网络（GCN）蒸馏到简单网络（MLP），验证了本文方法的有效性。

此外，为验证本文方法中各项损失的有效性，设计了一组消融实验，以验证域对抗损失和蒸馏损失的有效性，结果如表 3 所示。设置 1 仅保留了源域上的交叉熵损失，而设置 2 使用交叉熵损失和蒸馏损失，设置 3 使用交叉熵损失和域对抗损失，设置 4 为本文方法。从表 3 中可以观察到，只保留源域上的交叉熵损失时，实验结果最差，平均准确率比本文方法降低了 2.80 百分点，这说明了蒸馏损失和域对抗损失的重要性。通过进一步比较可以发现，设置 3 实验结果比设置 2 要更好，这表明在解决混合目标域问题时，知识蒸馏的效果更为显著。总体上，本文方法综合了 $\mathcal{L}_{ADV}$ 和 $\mathcal{L}_{KD}$，在混合目标域上取得了最好的效果。

表 3　不同损失函数消融实验的结果对比

Table 3　Comparison of results in ablation experiments on different loss functions　%

设置	$\mathcal{L}_{CE}$	$\mathcal{L}_{ADV}$	$\mathcal{L}_{KD}$	目标域				平均准确率
				PITT	UCLA_1	USM	YALE	
1	√	—	—	66.96±1.52	67.72±1.53	59.58±0.68	69.47±1.84	65.93±0.87
2	√	√	—	71.27±1.97	68.68±1.01	59.32±0.74	68.97±2.04	67.06±0.61
3	√	—	√	**71.73**±1.70	68.90±0.73	59.68±1.26	70.49±1.32	67.70±0.52
4	√	√	√	71.13±1.10	**69.96**±0.65	**59.80**±0.79	**74.00**±0.27	**68.73**±0.39

2.5　对比实验

为验证本文方法的有效性，把本文方法与近几年 UDA 的代表方法进行对比，包括单目标域的设置和混合目标域的设置。单目标域设置中，采用经典领域自适应方法，分别对每个目标域进行自适应，最后取 4 个目标域上准确率的平均值；而在混合目标域设置中，通过将单一源域直接自适应到混合目标域，训练统一泛化的模型。此外，还与 2 个多目标领域自适应（multi-target domain adaptation, MTDA）方法进行了对比。实验结果如表 4 所示。

表 4　本文方法与其他方法的比较

Table 4　Comparison of the proposed method with other methods

设置	模型	目标域/%				平均准确率/%	p-value
		PITT	UCLA_1	USM	YALE		
单目标域	DANN[27]	69.93±0.96	67.47±2.19	60.37±1.79	67.96±1.39	66.43±1.06	1×10^{-4}
	JAN[28]	70.01±1.88	69.17±0.60	60.22±1.69	66.51±1.63	66.48±0.79	7×10^{-6}
	ADDA[29]	70.13±0.55	67.76±1.75	58.25±1.51	67.06±1.20	65.80±0.94	1×10^{-5}
	CDAN[23]	69.91±1.73	68.12±1.98	59.48±1.74	70.52±2.38	67.01±1.07	3×10^{-4}
混合目标域	DANN[27]	68.07±1.73	67.64±1.75	**60.96**±3.14	65.03±1.65	65.43±1.37	7×10^{-5}
	JAN[28]	68.57±2.17	69.37±0.96	60.32±1.57	67.60±2.25	66.47±0.84	8×10^{-6}
	ADDA[29]	64.44±2.79	64.17±2.59	58.73±4.06	62.07±2.09	62.35±1.82	2×10^{-6}
	CDAN[23]	70.34±2.63	68.70±2.11	59.95±1.28	67.83±3.68	66.70±1.03	3×10^{-4}
	CGCT[16]	66.73±4.27	70.62±1.79	60.39±1.31	71.67±2.38	67.63±0.99	7×10^{-4}
	本文方法	**71.13**±1.10	**69.96**±0.65	59.80±0.79	**74.00**±0.27	**68.73**±0.39	—
多目标域*	D-CGCT[16]	70.03±1.14	68.41±2.48	60.34±0.77	68.97±2.98	66.94±1.11	2×10^{-3}
	MT-MTDA[30]	69.91±2.07	67.22±2.38	59.68±2.30	72.32±2.82	67.28±1.30	7×10^{-3}

从表 4 发现，本文方法在所有对比方法中表现最好，在 4 个目标域上取得了最高的平均准确率。在单目标域的设置中，CDAN 表现最佳，这也验证了本文方法使用 CDAN 的有效性。在混合目标域设置中，我们采用几种典型的对抗学习方法进行混合领域自适应，分别为 DANN（domain-adversarial training of neural networks）[27]、JAN（joint adaptation networks）[28]、ADDA（adversarial discriminative domain adaptation）[29] 和 CDAN。

将本文方法与其他混合目标域方法进行比较，从表 4 中不难发现，其他方法虽然在单目标域领域自适应的任务上能够取得较好的效果，但是将它们应用于混合目标域领域自适应，结果不够理想。其原因是混合目标域的数据分布更为分散，这些方法无法有针对性地对齐混合目标域中的数据分布，这与参考文献 [15] 中的结论是一致的。此外，在混合目标域设置中，还与 CGCT 进行比较。CGCT 方法通过 GCN 与 MLP 协同学习，将由 GCN 预测得到的伪标签加入到源域中进行训练，从而增强了源域信息，这有助于更好地实现混合目标领域自适应。尽管如此，对于 ASD 分类任务而言，不同个体的脑影像数据之间存在很大差异，比如功能连接模式不同、网络组织结构差异、静息态脑区活动差异等，导致样本之间关系复杂，这使得已有的模型难以学习到有效的特征表示。与已有的领域自适应方法不同，KD-BTDA 通过引入基于 GCN 的对抗知识蒸馏机制进行领域自适应：一方面，在进行知识蒸馏时将 GCN 作为教师模型，能够有效处理样本之间的复杂关系，有效缓解 ASD 异质性所引发的脑影像数据信息不一致的问题，同时可以使学生模型也学习到样本间的复杂关系；另一方面，对抗学习能够使教师模型在源域和混合目标域上学习到共同的特征表示，缓解域偏移问题。因此，KD-BTDA 能够在 ASD 分类任务上得到更好的效果。

在与 2 个 MTDA 方法比较中，本文方法也得到了最佳的结果。其中，MT-MTDA 方法[30] 采用多位教师指导单一学生的策略，即对于每个目标域，分别训练一个独立的教师模型来指导该目标域。然而，采用多位教师会增加网络的复杂度，并且在混合目标域中，数据分布差异更大，知识蒸馏难度增加，导致该方法无法有效解决混合目标域的问题。本文方法根据 ASD 数据的特点进行了有针对性的优化，将对抗学习引入知识蒸馏中，同时使用单个教师网络来指导混合目标域，不仅降低了网络的复杂度，还确保在处理 ASD 数据时模型具有更好的适应性。本文方法在应对更为复杂的混合目标域问题时，表现出比多目标域设置的 MT-MTDA 更为出色的效果。表 4 充分显示了本文方法的泛化能力，表明在混合目标域 ASD 分类问题上，该方法具有显著的有效性。

此外，本文进一步使用 paired t-test 对实验结果进行统计分析，并在表 4 中给出了 p 值（p-value）。为了减少随机误差，对每种方法进行了 10 次重复实验，得到了平均准确率。然后，将本文方法与其他方法分别进行了 paired t-test。结果显示，本文方法与所有对比方法之间的准确率差异是显著的（$p<0.05$）。大部分方法的 p 值都小于 0.001，进

一步表明本文方法在性能上的显著优势。这些结果均表明，本文提出的 KD-BTDA 模型在 ASD 分类性能上显著优于大多数对比方法。

值得注意的是，本文方法的目标是在混合目标域上获得一个统一的模型，而非专门针对某个目标域进行训练。虽然本文方法在 USM 这个目标域上的准确率略低于其他方法，但在混合目标域上的平均结果优于其他方法。

2.6　算法时间复杂度分析

KD-BTDA 计算的主要过程包括特征提取器、教师网络、学生网络和域鉴别器 4 个部分。这 4 个部分都为神经网络，它们的时间复杂度都基于网络的层数以及相应每层中结点的数量。下面分别计算这几个部分的时间复杂度：

1）特征提取器

特征提取器是一层全连接层，假设输入维度为P_1，输出维度为P_2。特征提取器F的时间复杂度为

$$O(F)=O(P_1P_2)$$

2）教师网络

教师网络G有两层图卷积层。假设输入节点的特征维度为$N\cdot P_2$，邻接矩阵维度为$N\cdot N$，隐藏层维度为Z，输出维度为类别数目C。

假设第一层图卷积输入特征矩阵维度$N\cdot P_2$，输出特征矩阵维度$N\cdot Z$，时间复杂度为

$$O(G_1)=O(N^2P_2+NP_2Z)$$

假设第二层图卷积的输入特征矩阵维度为$N\cdot Z$，输出特征矩阵维度为$N\cdot C$，时间复杂度为

$$O(G_2)=O(N^2Z+NZC)$$

因此，教师网络的总时间复杂度为

$$O(G)=O(G_1+G_2)=O(N^2P_2+NP_2Z+N^2Z+NZC)$$

3）学生网络

学生网络H是一层全连接层，假设输入维度为P_2，输出维度为C。学生网络的时间复杂度为

$$O(H)=O(P_2C)$$

4）域鉴别器

域鉴别器D包括随机层R和对抗网络A。随机层R有 2 个输入，假设维度分别为R_1和R_2，输出维度为W。随机层R的时间复杂度为

$$O(R)=O(R_1W+R_2W)$$

对抗网络A有 3 层全连接层，假设第一层输入维度为W，输出维度为Z_1，第二层输入维度为Z_1，输出维度为Z_2，第三层输入维度为Z_2，输出维度为U。对抗网络A的时间复杂度为

$$O(A)=O(WZ_1+Z_1Z_2+Z_2U)$$

域鉴别器D的总时间复杂度为

$$O(D)=O(R)+O(A)=$$
$$O(R_1W+R_2W+WZ_1+Z_1Z_2+Z_2U)$$

因此，整个算法的时间复杂度为

$$O(\mathrm{TOTAL})=O(F)+O(G)+O(H)+O(D)$$

2.7　算法收敛分析

本节进一步研究了 KD-BTDA 算法的收敛性。图 3 给出了训练过程中每个 Epoch 的训练损失以及在测试集上的准确率。我们给出了 10 折交叉验证中某一折的损失和准确率。由图 3 可以观察到，模型在训练到 100 个 Epoch时，目标函数趋于稳定，说明训练过程已接近收敛，此时在混合目标域的测试集上平均准确率也接近 68%。

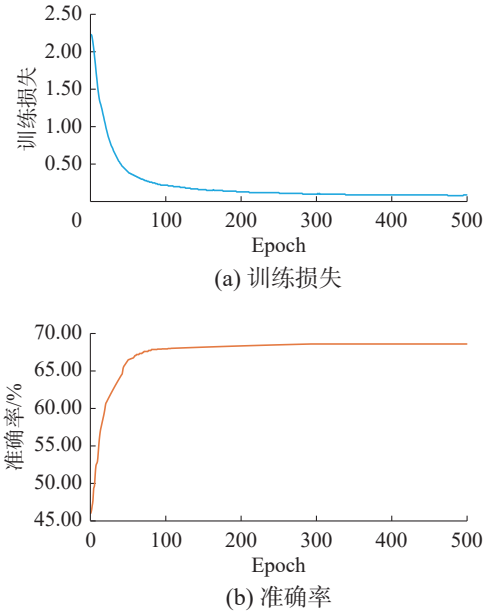

图 3　KD-BTDA 的训练损失和准确率

Fig. 3　Training loss and accuracy of KD-BTDA

3　结束语

本文提出了一种基于知识蒸馏的混合目标领域自适应模型。该模型采用 GCN 作为教师网络，将知识蒸馏到学生网络。教师网络使用 GCN，有助于网络学习到样本之间的相互关系，结合领域自适应方法为特征对齐和分类提供更多的先验知识；引入知识蒸馏解决了 GCN 复杂程度高、无法对单个样本进行测试的问题，同时也避免了测试阶段因引入新样本改变图的结构而需要重新训练网络的问题。在实验部分，本文通过消融实验证明了选择 GCN 作为教师模型的优势，同时验证了将知识蒸馏损失引入混合目标领域自适应的有效性；通过与经典单目标域和混合目标域的领域自适应方法进行比较，验证了本文方法在 BTDA 问题上的有效性。

目前，本文方法需要直接访问源域数据。但是，在临床上，由于隐私保护的问题，可能无法直接接触病人的影像数据，从而在领域自适应时无法将它们作为源域数据直接使用。在此情况下，实现混合目标域的领域自适应更具挑战性。此外，在临床上可能会遇到需要对未知目标域中的样本进行分类的情况，这涉及一个新的问题，即领域泛化。在未来研究中，将对这些问题进行深入研究。

参考文献：

[1] LORD C, ELSABBAGH M, BAIRD G, et al. Autism spectrum disorder[J]. Lancet, 2018, 392(10146): 508−520.

[2] HEINSFELD A S, FRANCO A R, CRADDOCK R C, et al. Identification of autism spectrum disorder using deep learning and the ABIDE dataset[J]. NeuroImage: clinical, 2018, 17: 16−23.

[3] LIU Xingdan, WU Jiacheng, LI Wenqi, et al. Domain adaptation via low rank and class discriminative representation for autism spectrum disorder identification: a multi-site fMRI study[J]. IEEE transactions on neural systems and rehabilitation engineering, 2023: 3233656.

[4] WANG Mingliang, ZHANG Daoqiang, HUANG Jiashuang, et al. Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation[J]. IEEE transactions on medical imaging, 2020, 39(3): 644−655.

[5] ABRAHAM A, MILHAM M P, DI MARTINO A, et al. Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example[J]. NeuroImage, 2017, 147: 736−745.

[6] 蔡鸿顺, 张琼敏, 龙颖. 面向阿尔茨海默症辅助诊断的多尺度域适应网络[J]. 智能系统学报, 2023, 18(5): 1090−1098.
CAI Hongshun, ZHANG Qiongmin, LONG Ying. Multiscale domain adaptation network for the auxiliary diagnosis of Alzheimer's disease[J]. CAAI transactions on intelligent systems, 2023, 18(5): 1090−1098.

[7] 梁艳, 温兴, 潘家辉. 融合全局与局部特征的跨数据集表情识别方法[J]. 智能系统学报, 2023, 18(6): 1205−1212.
LIANG Yan, WEN Xing, PAN Jiahui. Cross-dataset facial expression recognition method fusing global and local features[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1205−1212.

[8] MANSOUR Y, MOHRI M, ROSTAMIZADEH A. Domain adaptation: learning bounds and algorithms[EB/OL]. (2023−11−30)[2024−03−13]. https://arxiv.org/abs/0902.3430v3.

[9] BEN-DAVID S, BLITZER J, CRAMMER K, et al. A theory of learning from different domains[J]. Machine learning, 2010, 79(1): 151−175.

[10] LONG Mingsheng, CAO Yue, WANG Jianmin, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: ACM, 2015: 97−105.

[11] DING Jie, WANG Li, YU Lei, et al. Low-rank domain adaptive method with inter-class difference constraint for multi-site autism spectrum disorder identification[C]// 2022 7th International Conference on Computational Intelligence and Applications. Nanjing: IEEE, 2022: 237−242.

[12] CHEN Ziliang, ZHUANG Jingyu, LIANG Xiaodan, et al. Blending-target domain adaptation by adversarial meta-adaptation networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2243−2252.

[13] LIU Ziwei, MIAO Zhongqi, PAN Xingang, et al. Open compound domain adaptation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12406−12415.

[14] ZHENG J, WU Wenzhao, FU Haohuan, et al. Unsupervised mixed multi-target domain adaptation for remote sensing images classification[C]//2020 IEEE International Geoscience and Remote Sensing Symposium. Waikoloa: IEEE, 2020: 1381−1384.

[15] XU Pengcheng, WANG Boyu, LING C. Class overwhelms: mutual conditional blended-target domain adaptation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2023, 37(3): 3036−3044.

[16] ROY S, KRIVOSHEEV E, ZHONG Zhun, et al. Curriculum graph co-teaching for multi-target domain adaptation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 5347−5356.

[17] PARISOT S, KTENA S I, FERRANTE E, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease [J]. Medical image analysis, 2018, 48: 117−130.

[18] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015−05−09)[2024−03−13]. https://arxiv.org/abs/1503.02531v1.

[19] 黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624−653.
HUANG Zhenhua, YANG Shunzhi, LIN Wei, et al. Knowledge distillation: a survey[J]. Chinese journal of computers, 2022, 45(3): 624−653.

[20] CHEN Defang, MEI Jianping, WANG Can, et al. Online knowledge distillation with diverse peers[C]//Proceedings of the AAAI conference on artificial intelligence. New York: AAAI, 2020, 34(4): 3430−3437.

[21] WILSON G, COOK D J. A survey of unsupervised deep domain adaptation[J]. ACM transactions on intelligent systems and technology, 2020, 11(5): 1−46.

[22] LUO Yadan, WANG Zijian, HUANG Zi, et al. Progressive graph learning for open-set domain adaptation[C]// Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 6468−6478.

[23] LONG Mingsheng, CAO Zhangjie, WANG Jianmin, et al. Conditional adversarial domain adaptation[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: ACM, 2018: 1647−1657.

[24] DI MARTINO A, YAN C G, LI Q, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism[J]. Molecular psychiatry, 2014, 19(6): 659−667.

[25] CAMERON C, SHARAD S, BRIAN C, et al. Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (C-PAC)[J]. Frontiers in neuroinformatics, 2013, 7: 42.

[26] RYU M, LEE G, LEE K. Knowledge distillation for BERT unsupervised domain adaptation[J]. Knowledge and information systems, 2022, 64(11): 3113−3128.

[27] MURPHY K, SCHÖLKOPF B, GANIN Y, et al. Domain-adversarial training of neural networks[J]. Journal of machine learning research, 2016, 17(1): 2096−2030.

[28] LONG Mingsheng, ZHU Han, WANG Jianmin, et al. Deep transfer learning with joint adaptation networks[C]// Proceedings of the 34th International Conference on Machine Learning. Sydney: ACM, 2017: 2208−2217.

[29] ZELLINGER W, MOSER B A, GRUBINGER T, et al. Robust unsupervised domain adaptation for neural networks via moment alignment[J]. Information sciences, 2019, 483: 174−191.

[30] NGUYEN-MEIDINE L T, BELAL A, KIRAN M, et al. Unsupervised multi-target domain adaptation through knowledge distillation[C]//2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021: 1338−1346.

作者简介：

顿家乐，硕士研究生，主要研究方向为深度学习、计算机视觉、迁移学习。E-mail: dunjiale1997@163.com。

王骏，副教授，博士。中国计算机学会高级会员，IEEE 高级会员，中国人工智能学会粒计算与知识发现专业委员会委员、机器学习专业委员会委员，MICS online 委员。主要研究方向为机器学习、医学影像智能计算。发表学术论文 70 余篇。E-mail: wangjun_shu@shu.edu.cn。

彭汉琛，硕士研究生，主要研究方向为深度学习、迁移学习、图像处理。E-mail: phcking0219@163.com。