arXiv:2302.03020v2 [cs.LG] 5 Jun 2023

RLSbench：宽松标签偏移下的域适应

Saurabh Garg [1] **Nick Erickson** [2] **James Sharpnack** [2] **Alex J. Smola** [2] **Sivaraman Balakrishnan** [1] **Zachary C. Lipton** [1 2]

摘要

尽管针对标签偏移的领域自适应已有严谨的方法，但这些方法对类别条件分布偏移的敏感性仍未得到充分探究。与此同时，流行的深度领域自适应启发式方法在面对标签比例偏移时往往表现不佳。虽然有若干论文尝试修改这些启发式方法以应对标签比例偏移，但评估标准、数据集和基线方法之间缺乏一致性，因此难以判断当前的最佳实践。本文介绍了RLS-BENCH，这是一个大规模的宽松标签偏移基准测试，包含 >500 个涵盖视觉、表格数据和语言模态、标签比例各异的分布偏移对。与主要关注类别条件分布偏移的 $p(x|y)$现有基准测试不同，我们的基准测试也关注标签边缘分布偏移。首先，我们评估了 13 种流行的领域自适应方法，发现它们在标签比例偏移下的失效情况比此前所知更为普遍。接着，我们提出了一种有效的两步元算法，可与大多数领域自适应启发式方法兼容：(i) 在每个训练轮次对数据进行伪平衡；(ii) 使用目标标签分布估计调整最终分类器。在标签比例大幅偏移时，该元算法能够改进现有的领域自适应启发式方法，准确率通常提高 2–10 个百分点；而在标签比例没有偏移时，影响极小（<0.5%）。我们希望这些发现以及 RLSBENCH 的开放获取能够鼓励研究人员在宽松标签偏移情境下严格评估所提出的方法。代码已公开发布于
h
t
t
ps://github.com/acmi-lab/RLSbench。

[1]卡内基梅隆大学 [2]亚马逊云科技。通信请联系：Saurabh Garg <sgarg2@andrew.cmu.edu>。

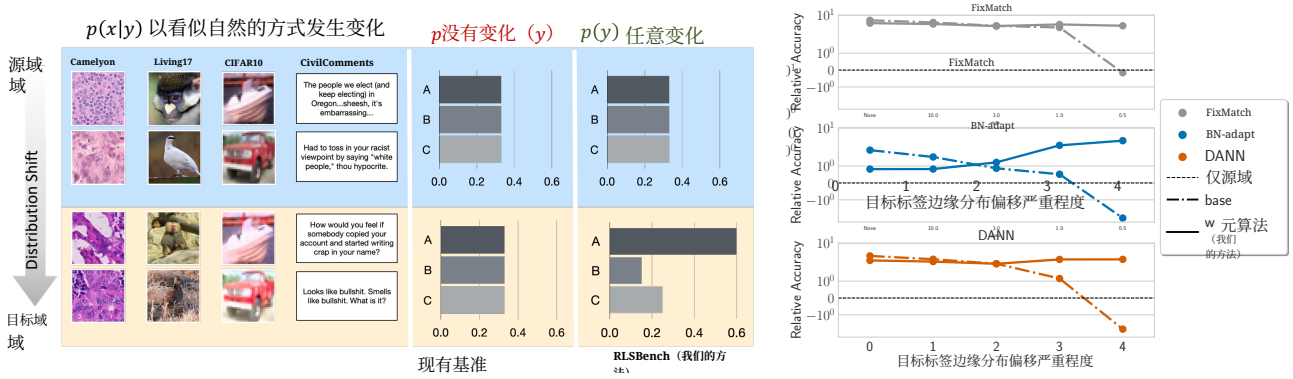
国际机器学习会议 40^{th} 国际机器学习会议，美国夏威夷檀香山。PMLR 202，2023年。版权所有 2023年，归作者所有。

1. 引言

现实世界中机器学习模型的部署通常具有分布偏移的特征，即生产环境中遇到的数据与训练数据在统计特性上存在差异（Quinonero-Candela 等人，2008年；Torralba 和 Efros，2011年；Koh 等人，2021年）。由于持续为数据添加标签的成本可能高得难以承受，研究人员将重点放在无监督域适应（DA）设置上，在该设置中，训练时仅能获得来自源域分布的有标签数据，以及来自目标域分布的未标记数据。在缺乏进一步假设的情况下，域适应问题被认为是不充分确定的（Ben-David 等人，2010b），因此不存在一种普遍适用的方法。

研究人员针对这些挑战提出了多种应对方法。一种方法是探究使域适应问题适定的结构性假设。常见的例子包括协变量偏移和标签偏移；只要源分布和目标分布具有重叠支撑集，就存在相应的识别策略和原则性方法（Shimodaira，2000年；Sch¨olkopf 等人，2012年；Gretton 等人，2009年）。例如，近期关于标签偏移的研究提出了有效方法，这些方法适用于深度学习场景，既能一致地估计目标标签边缘分布，也提供了更新所得分类器的原则性方法（Lipton 等人，2018年；Alexandari 等人，2021年；Azizzadenesheli 等人，2019年；Garg 等，2020年）。然而，这类假设在实际应用中通常会在一定程度上遭到违背。即使是疾病患病率变化这类典型情形，标签偏移假设也可能不成立。例如，在COVID-19疫情期间，疾病阳性率的变化与治疗方案、感染人群的年龄分布以及病毒遗传构成的变化同时发生。

另一条互补的研究路线聚焦于构建基准数据集，希望找到能够利用无标签目标数据的启发式方法，从而改善实际中出现的各类问题。此类基准测试包括 Office-Home（Venkateswara 等人，2017年）、Domainnet（Peng 等人，2019年）、WILDS（Sagawa 等人，2021年）。然而，大多数此类基准数据集中的标签分布几乎没有发生变化 $p(y)$（或者完全没有变化）。因此，基于基准测试的研究



(a) (b)

图 *1.* 宽松标签偏移下的领域自适应。**(a)RLSbench设置概述：** 与标签边缘分布 $p(y)$ 不会发生偏移的现有基准测试不同，在RLSbench中，$p(y)$ 可以任意偏移。类别条件分布 $p(x|y)$ 则会按照常见基准测试所呈现的方式，以看似自然的形式发生偏移。RLSbench涵盖了 14 个多域数据集，涉及视觉、自然语言处理和表格模态。**(b) 主要结果：** 随着目标标签比例偏移程度加剧，现有流行域适应方法的性能会下降，往往低于仅源域分类器。将域适应方法与我们的元算法结合后，其性能显著优于仅源域分类器。

产生了各种启发式方法（Ganin 等，2016；Sohn 等人，2020年；Wang 等人，2021年；Li 等人，2016），尽管这些方法能够提升基准测试性能，但当 $p(y)$ 发生变化时往往会失效。尽管此前已有研究表明域对抗方法存在这一脆弱性（Wu 等，2019；Zhao 等人，2019），我们证明这一问题的普遍程度超出了以往的认识。近期有多篇论文试图应对由自然变化加剧的标签分布变化问题 $p(x|y)$（Tan 等人，2020年；Tachet des Combes 等，2020；Prabhu 等人，2021年）。然而，由于不同论文在如何模拟 $p(y)$ 变化以及评估指标的选择方面存在差异，跨论文比较实验结果可能较为困难。此外，许多方法通过在模型选择和超参数调优期间查看目标域验证性能，违反了无监督设定的约束（Wilson & Cook, 2020；Saito 等人，2021年）。简而言之，对于标签分布发生变化的设定，目前缺乏对域适应方法进行全面且公平比较的研究。

本文构建了 RLSBENCH，这是首个针对宽松标签偏移情形的标准化测试平台，其中 $p(y)$ 可以任意偏移，而类别条件分布 $p(x|y)$ 也可以以看似自然的方式偏移（遵循广受欢迎的域适应基准）。现有域适应基准通常关注 $p(x|y)$ 的偏移，而我们的基准还关注标签边缘分布 $p(y)$的偏移。我们在涵盖视觉、自然语言处理（NLP）和表格模态的 14 个多域数据集上，评估了一系列广受欢迎的域适应方法，包括基于域不变表示学习、自训练和测试时自适应的方法。每个数据集中的不同领域呈现出不同的$p(x|y)$偏移。由于这些数据集中的标签边缘分布偏移较小甚至不存在，我们通过不同程度的分层抽样来模拟目标标签边缘分布偏移。总的来说，我们获得了 560 个不同的源域与目标域分布偏移对，并在我们的测试平台上训练 $> 30k$ 模型。

基于在 RLSBENCH 上开展的实验，我们得出以下几项发现。首先，我们观察到，在目标标签分布未发生偏移时，流行的域适应方法通常优于仅源域分类器；但在目标标签边缘分布严重偏移时，这些方法的性能往往会下降，甚至低于仅源域分类器。接下来，我们提出了一种包含两项简单校正的元算法：(i) 对数据进行重采样，使源域达到平衡，并使目标域达到伪平衡；(ii) 根据对目标标签边缘分布的估计，对最终分类器进行重新加权。我们观察到，在这些宽松标签偏移设置下，将现有域适应方法（例如 CDANN、FixMatch 和 BN-adapt）与我们的元算法结合后，其性能显著优于仅源域分类器。另一方面，专为宽松标签偏移提出的现有方法（例如 IW-CDANN 和 SEN-TRY ）往往无法优于仅源域分类器；与结合我们元算法的现有域适应方法相比，其性能也明显较差。

总体而言，RLSbench 为标签分布偏移提供了一套全面且标准化的基准测试，使现有基准测试更进一步，能够体现我们在实际部署模型时应当预期会遇到的多样性。我们的发现强调了一种简单却此前被忽视的基线方法的有效性。我们希望 RLSBENCH 及我们的元算法（可与任何域适应方法结合使用）能够为宽松标签偏移场景下严谨且可复现的未来研究提供框架。

2. 预备知识与先前工作

我们首先介绍符号并正式定义问题。设 $\mathcal{X}$ 为输入空间，$\mathcal{Y} = \{1, 2, \ldots, k\}$ 为

输出空间。设 P_s，P_t：$\mathcal{X} \times \mathcal{Y} \to [0,\ 1]$ 为源分布和目标分布，p_s 和 p_t 分别表示相应的概率密度（或质量）函数。与标准监督学习设置不同，在无监督领域自适应中，我们拥有有标签源数据 $\{(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)\}$ 和无标签目标数据 $\{x_{n+1},\ x_{n+2},\ \ldots,\ x_{n+m}\}$。用 f：$\mathcal{X} \to \Delta^{k-1}$ 表示一个预测函数，该函数在输入 x 上预测 $\hat{y} = \arg\max_y f_y(x)$。对于向量 v，我们用 v_y 访问索引为 y的元素。

在传统标签偏移设置中，通常假设 $p(x|y)$不会发生变化，但 $p(y)$ 可能会发生变化。在标签偏移下，会出现两个挑战：(i) 估计目标标签边缘分布 $p_t(y)$；以及 (ii) 训练分类器 f，以最大化其在目标域上的性能。本文关注宽松标签偏移设置。具体而言，我们假设标签分布可以从源域到目标域任意变化，但$p(x|y)$在源域与目标域之间仅以某种相对受限的方式变化（例如，在真实世界中自然产生的变化，如从 ImageNet（Russakovsky 等人，2015年）到 ImageNetV2（Recht 等人，2019年））。从数学上讲，我们对 $p(x|y)$施加基于散度的约束。也就是说，对于某个较小的 $\epsilon > 0$ 和分布距离 $\mathcal{D}$，有$\max_y \mathcal{D}(p_s(x|y),\ p_t(x|y)) \leqslant \epsilon$ 并允许标签边缘分布 $p(y)$发生任意偏移。我们在附录G中讨论了几种精确的实例化方式。然而，在实践中，对于足够小的ϵ，使用有限样本很难对这些分布距离进行实证验证。此外，我们尚未对这些偏移在流行的领域自适应基准中产生的具体含义进行严格刻画；鉴于本文工作的重点是使用真实世界数据集进行实证评估，我们将形式化研究留待未来工作。

域适应的目标是将预测器从具有有标签数据的源分布迁移到目标分布，而我们只能观测到来自目标分布的未标注样本。以往研究宽松标签偏移的工作主要关注分类器性能，而我们还单独评估了估计目标标签边缘分布的方法。这有两方面益处。首先，可以更深入地了解改善目标类别比例估计如何提升目标性能。其次，理解类别比例如何变化本身也具有研究价值。

2.1. 先前工作

无监督域自适应　域适应的两种常见且定义良好的设置包括：(i) 协变量偏移（Zhang 等，2013年；Zadrozny，2004；Cortes 等，2010年；Cortes 与 Mohri，2014年；Gretton 等人，2009年），其中 $p(x)$ 可以从源域变到目标域，而 $p(y|x)$ 保持不变；以及 (ii) 标签偏移（Saerens 等人，2002年；Lipton 等，2018年；Azizzadenesheli 等人，2019年；Alexandari 等人，2021年；Garg 等，2020年；Zhang 等，2021年；Roberts 等人，2022年），其中标签边缘分布 $p(y)$ 可以变化，但 $p(x|y)$ 保持共享于源域和目标域之间。当目标分布的支撑集是源域支撑集的子集时，针对这些设置的原则性方法具有强有力的理论保证。Ben-David 等人（2010b;a）；Mansour 等人（2009年）；Zhao 等人（2019年）；Wu 等人（2019年）；Johansson 等人（2019年）对协变量支撑集包含这一假设被违反时的情形进行了理论分析。另一条研究路线中，Elkan 与 Noto（2008年）；Bekker 与 Davis（2020年）；Garg 等（2021; 2022a）将标签偏移设置扩展到目标域中可能出现先前未见类别的问题，并且 $p(x|y)$ 在已见类别中保持不变。最近，大量研究开始探索基准驱动的启发式方法（Long 等人，2015年；2017年；Sun 与 Saenko，2016年；Sun 等人，2017年；Zhang 等人，2019年；2018年；Ganin 等，2016；Sohn 等人，2020年）。然而，对域适应方法的严格评估通常局限于这些精心整理的基准数据集，其中源域到目标域的标签边缘分布几乎没有或完全没有偏移。

宽松标签偏移 在自然变化下探究标签边缘分布从源域到目标域的偏移问题 $p(x|y)$，有少数论文指出，基于域对抗神经网络训练的领域自适应方法在理论和实证方面均存在不足（Yan 等人，2017年；Wu 等人，2019年；Zhao 等人，2019年；Johansson 等人，2019年）。随后，若干论文尝试通过域对抗训练来解决这些问题（Tachet 等人，2020年；Prabhu 等人，2021年；Liu 等人，2021年；Tan 等人，2020年；Manders 等人，2019年）。然而，这些方法往往缺少与其他主流领域自适应方法的比较，而且评估所用的数据集和模型选择标准各不相同。为此，我们在标准化评估框架下，对具有代表性的流行领域自适应方法进行了大规模、严格的比较。

领域泛化 在领域泛化中，模型可以访问来自多个不同领域的数据，目标是在测试时泛化到一个此前未见过的领域（Blanchard 等人，2011年；Muandet 等人，2013年）。关于不同领域泛化算法的综述，读者可参考Gulrajani & Lopez-Paz（2020年）。这里的一个关键区别在于，与领域泛化设置不同，在域适应问题中，我们可以访问来自测试域的未标注样本。

与以往分布偏移基准研究的区别 以往评估分布偏移下鲁棒性的研究主要聚焦于迁移学习和领域泛化设置 Wenzel 等人(2022年)；Gulrajani 与 Lopez-Paz (2020年)；Djolonga 等人(2021年)；Wiles 等人(2021年)；Koh 等人(2021年)。Taori 等人(2020年)；Hendrycks 等人(2021年)研究了鲁棒性干预（例如，数据增强技术、对抗训练）对目标域（分布外）性能的影响。值得注意的是，Sagawa 等人(2021年)专注于评估 WILDS 2.0 上的域适应方法。我们的工作与这些研究相辅相成，因为我们

首次全面研究了分布偏移下的域适应方法，包括 $p(y)$ 和自然变化 $p(x|y)$。

3. RLSBENCH：宽松标签偏移基准测试

在本节中，我们介绍 RLSBENCH，这是一个由数据集和域适应算法组成的套件，也是本研究的核心。受针对（更严格的）标签偏移设定的校正方法（Saerens 等人，2002年；Lipton 等人，2018年）以及不平衡数据集上的学习（Wei 等人，2021年；Cao 等人，2019a）的启发，我们还提出了一种元算法，通过简单的校正即可与几乎任何域适应方法兼容。

3.1. 数据集

RLSBENCH 基于14 个用于分类的多域数据集，涵盖物体分类、卫星图像、医学和毒性检测等应用领域的任务。在这些数据集中，我们总共获得了 56 个不同的源域与目标域组合。关于数据集的更多详情，请参见附录 D。

(i) CIFAR-10，包括原始 CIFAR-10（Krizhevsky 和 Hinton，2009年）、CIFAR-10-C（Hendrycks 和 Dietterich，2019年）和 CIFAR-10v2（Recht 等人，2018年）；(ii) CIFAR-100，包括原始数据集和 CIFAR-100-C；(iii) 全部四个 BREEDs 数据集（Santurkar等，2021年），即 Entity13、实体30、Nonliving26、Living17。BREEDs 利用 ImageNet 中的类别层级（Russakovsky 等人，2015年），将原始类别重新用作子群体，并在超类上定义分类任务。我们考察子群体偏移，以及由 ImageNet 数据收集过程差异引起的自然偏移（即 ImageNetv2（Recht 等人，2019年）），并研究两者共同作用的情况。(iv) OfficeHome（Venkateswara 等人，2017年），包含四个领域：艺术、剪贴画、产品和真实；(v) DomainNet（Peng 等人，2019年），我们考察其中四个领域：剪贴画、绘画、真实、素描；(vi) Visda（Peng 等人，2018年；2017年），包含三个领域：训练集、验证集和测试集；(vii) 来自 WILDS基准的 FMoW（Koh 等人，2021年；Christie 等人，2018年），包含三个领域：训练集、分布外验证集和分布外测试集，其中卫星图像采自不同地理区域和不同时期；(viii) 来自 WILDS基准的 Camelyon（Bandi 等人，2018年），包含三个领域：训练集、分布外验证集和分布外测试集，用于肿瘤识别，领域对应不同医院；(ix) Civilcomments（Borkan 等人，2019年），包含三个领域：训练集、分布外验证集和分布外测试集，用于毒性检测，领域对应不同人口统计子群体；(x) 退休成年人（Ding 等人，2021年），我们考察 ACSIncome 预测任务，其中不同领域代表不同州和时间段；以及 (xi) Mimic 再入院（Johnson 等人，2020年；PhysioBank，2000年），任务是预测再入院风险，其中不同领域代表不同时间段的数据，ious 不同领域代表来自不同时间段的数据期。

模拟目标边缘分布的偏移 上述数据集的标签边缘分布偏移很小，甚至没有偏移。因此，我们通过改变目标标签边缘分布，同时保持源目标分布固定（即保持为原始源标签分布），来模拟这种偏移。请注意，与以往一些研究不同，我们不会改变源标签边缘分布，因为在实践中，我们或许可以选择谨慎地整理训练分布，但可能几乎无法控制目标标签边缘分布。

对于每个目标数据集，我们都有真实标签，因此可以改变目标标签分布。具体而言，我们从狄利克雷分布中采样目标标签边缘分布，其参数为 $\alpha \in \{0.5, 1, 3。0, 10\}$ 倍的原始目标边缘分布。具体来说，$p_t(y) \sim \text{Dir}(\beta)$ 其中 $\beta_y = \alpha \cdot p_{t,\ 0}(y)$ 且 $p_{t,\ 0}(y)$ 为原始目标标签边缘分布。狄利克雷参数 α 控制目标标签边缘分布偏移的严重程度。直观而言，随着 α 减小，偏移会加剧。为完整起见，我们还纳入了具有原始目标标签边缘分布的目标数据集。为便于说明，我们将 NONE（无外部偏移）也纳入狄利克雷参数集合，即 $\alpha \to \infty$时的极限分布。在模拟目标标签边缘分布偏移后（每个 α 使用两个随机种子），我们得到 560 对不同的源数据集和目标数据集。

3.2. 领域自适应方法

我们实现了以下算法（各方法的详细说明见附录L）：

仅源域 作为基线，我们纳入了使用源域上的经验风险最小化（Vapnik，1999年）和交叉熵损失训练的模型。我们还纳入了使用和未使用数据增强训练的仅源域模型。此外，我们还纳入了使用增强后的源数据训练的对抗鲁棒模型（**源域（对抗）**）。具体而言，我们使用针对 ℓ_2 扰动进行对抗训练的模型。

域对齐方法 这些方法采用领域对抗训练方案，旨在学习不同领域间不变的表征（Ganin 等人，2016年；Zhang 等人，2019年；Tan 等人，2020年）。在我们的实验中，我们纳入了以下五种方法：领域对抗神经网络（DANN（Ganin 等人，2016年））、条件领域对抗神经网络（CDANN（Long 等人，2018年））、最大分类器差异（MCD（Saito 等人，2018年））、重要性重加权 DANN 和 CDANN（即 IW-DANN 和 IW-CDANN，Tachet des Combes 等人（2020年））。

自训练方法 这些方法使用模型自身的预测为无标签样本“生成伪标签”，然后将其作为有标签样本进行训练。对于视觉数据集，这些方法通常还会使用一致性正则化，以鼓励模型做出一致的预 con-

算法 1 用于处理标签边缘分布偏移的元算法

输入 源域训练集和验证数据：$(X_S,\ Y_S)$ 和 $(X'_S,\ Y'_S)$，未标记的目标域训练集和验证数据：X_T和 X'_T、分类器 f以及领域自适应算法 $\mathcal{A}$

~ ~1: $X_S,\ Y_S \leftarrow$ 类别平衡采样$(X_S,\ Y_S)$ ▷平衡源数据

2: **针对** $t = 1$ 至 T **执行**

3: ~$Y_T \leftarrow$ 最大化$_y\ f_y(X_T)$

4: $\widetilde{X}_T \leftarrow$ 类别平衡采样$(X_T,\ \widehat{Y}_T)$ ▷伪平衡目标数据

5: 在平衡的源域上运行一轮 $\mathcal{A}$，以更新 f 数据 $\{\widetilde{X}_S, \widetilde{Y}_S\}$ 和目标数据 $\{\ \widetilde{X}_T\}$

6: **结束循环**

7: $\widehat{p}_t(y) \leftarrow$ 估计标签边缘分布(f, X'_S, Y'_S, X'_T)

8: $f'_j \leftarrow \frac{\widehat{p}_t(y=j)\cdot f_j}{\sum_k \widehat{p}_t(y=k)\cdot f_k}$ 对所有 $j \in \mathcal{Y}$ ▷重新加权分类器

输出目标标签边缘分布 $\widehat{p}_t(y)$ 和分类器 f'

对未标记数据样本的增强视图给出一致的预测（Lee 等人，2013年；Xie 等人，2020b；Berthelot 等人，2021年）。我们纳入以下三种算法：**Fix-Match** （Sohn 等人，2020），**NoisyStudent** （Xie 等人，2020a），基于集成委员会一致性的选择性熵优化（**SENTRY** （Prabhu 等人，2021年））。对于自然语言处理和表格数据集，由于没有定义强数据增强，我们考虑**伪标签** 算法（Lee 等人，2013年）。

测试时自适应方法 这些方法以源模型为基础，在无标签目标数据上调整少量参数（例如批归一化参数等），旨在提升目标性能。我们纳入了：**CORAL**(Sun 等人，2016年)或域调整回归（DARE（Rosen- feld 等人，2022年））、批归一化自适应（**BN-adapt** (Li 等人，2016年；Schneider 等人，2020年））、测试熵最小化（**TENT** (Wang 等人，2021年）

3.3. M eta 算法以处理目标标签边缘分布 偏移

这里我们讨论在框架中实现的两种简单通用校正。首先要注意，随着目标标签边缘分布偏移程度的增加，由于训练是在类别比例不同的源数据集和目标数据集上进行的，域适应方法的性能可能会下降。事实上，文献已从理论和实证两方面证明，领域对抗训练方法（深度域自适应方法的一类）可能失效（Wu 等人，2019年；Zhao 等人，2019年）。在我们的实验中，我们表明，由标签分布偏移导致的失效并非仅限于领域对抗训练方法，而是在所有常用域适应方法中都很常见（第4节）。

重采样 为处理标准监督学习中的标签不平衡问题，通过重采样数据来平衡类别边缘分布是一种已知的有效策略（Chawla 等人，2002年；Buda 等人，2018年；Cao 等人，2019b）。在宽松标签偏移中，我们希望处理目标数据中的不平衡问题（相对于源标签边缘分布而言），但无法获得真实标签。我们采用一种替代策略，利用目标数据的伪标签执行伪标签类别平衡重采样[1]（Zou 等人，2018年；Wei 等人，2021年）。对于宽松标签偏移问题，Prabhu 等人，2021年采用了这一技术，并将其用于集成委员会一致性目标 SENTRY。然而，他们没有探索基于重采样的现有域适应方法校正。由于这一技术可以与任何领域自适应方法结合使用，我们将该重采样技术应用于现有领域自适应方法，并发现重采样使所有领域自适应方法均受益；在我们的测试平台上（第4节），其表现通常优于 SENTRY。

重新加权 通过重采样，理想情况下，我们可以在由平衡的源数据集和平衡的目标数据集构成的混合数据上训练$\widehat{}$分类器 f。然而，这仍未解决如何使分类器适应$\widehat{}$原始目标标签分布这一问题，f因为该分布无法获得。如果我们能够估计目标标签边缘分布，就可以事后通过简单的重新加权校正来调整$\widehat{}$分类器 f（Lipton 等人，2018年；Alexandari 等人，2021年）。为了估计目标标签边缘分布，我们采用在更严格的标签偏移假设下开发的技术（回顾一下，这种设定中$p(x|y)$ 保持领域不变）。这些方法利用现成分类器来估计目标边缘分布，并在标签偏移条件下、对分类器仅作温和假设时，给出 $\mathcal{O}(1/\sqrt{n})$ 收敛速率（Lipton 等人，2018年；Azizzadenesheli 等人，2019年；Garg 等，2020年）。

尽管宽松标签偏移情形不满足标签偏移估计技术保持一致性所需的条件，我们仍采用这些技术，并在测试平台中通过实证评估这些方法的效果。具体而言，为了估计目标标签边缘分布，我们尝试了：(i) RLLS（Azizzadenesheli 等人，2019年）；(ii) MLLS（Alexandari 等人，2021年）；以及 (iii) 基线估计器，它仅对分类器在 f 无标签目标数据上的预测取平均值。我们在附录F中详细介绍这些方法。由于这些方法使用现成的分类器，因此可以将采用任何领域自适应方法获得的分类器与这些估计方法结合使用。

摘要 总体而言，在算法 1中，我们展示了如何将重采样和重新加权校正融入现有的领域自适应方法。附录9 中的图E 展示了该方法。算法 $\mathcal{A}$ 可以是任意领域自适应方法；在步骤 7 中，我们可以使用上述三种方法中的任意一种来估计目标标签边缘分布。我们通过若干算法实例化算法1

[1]另一种策略是对目标伪标签边缘分布进行重采样，使其与源标签边缘分布相匹配。为简便起见，我们选择平衡源标签边缘分布和目标伪标签边缘分布。

附录 L 中的第 3.2 节。直观来说，在理想情况下，如果我们元算法中的重采样步骤能够完美校正源域与目标域之间的标签不平衡，我们预期领域自适应方法能够将分类器 f 适应于 $p(x|y)$ 偏移。随后，我们元算法中的重新加权步骤可以将分类器 f 适应于目标标签边缘分布 $p_t(y)$。我们强调，在本研究中，我们 并未声称提出了这些校正方法。但据我们所知，我们的工作首次将这两种校正结合起来，并在多种不同的数据集上进行了广泛实验。

3.4 真实评估的其他选择

为了在不同数据集和域适应算法之间进行公平评估和比较，我们在适用的情况下采用一致的设计选择重新实现了所有算法。我们还做出了一些额外的实现选择，详述如下。更多细节请参见附录。马萨诸塞州。

模型选择标准和超参数 鉴于我们缺乏来自目标分布的独立同分布验证数据，域适应问题中的模型选择无法遵循监督训练中使用的标准流程。先前工作通常省略如何选择超参数的细节，从而留下使用测试集选择超参数的可能性，这可能会造成虚假且不可靠的改进感知。此外，不一致的超参数选择策略会使公平评估变得复杂，将改进错误地归因于所研究的算法。

在我们的工作中，我们使用源域留出集性能来选择最佳超参数。首先，对于 ℓ_2 正则化和学习率，我们对随机超参数执行网格搜索，以最大化仅源域模型在源域留出数据上的性能。然后，对于每个数据集，我们在所有领域自适应算法中固定使用这些超参数。对于领域自适应方法特有的超参数，我们在所有方法中使用相同的超参数，并结合相应论文提出的建议。在一次运行中，我们使用源域上的留出集性能来选择提前停止点。在附录中，我们通过使用目标域准确率选择提前停止点，报告*oracle* 性能。

评估标准 为评估目标标签边缘分布估计，我们报告估计标签分布与真实标签分布之间的 ℓ_1 误差。为评估分类器在目标数据上的性能，我们报告（经适配的）分类器在目标数据留出分区上的性能。

架构与预训练细节 我们测试了不同的架构（例如 DenseNet121、Resenet18、ResNet50、DistilBERT、多层感知机和 Transformer）。我们还测试了随机初始化的模型，以及使用 ImageNet 和 DistilBERT 预训练的模型。对于给定的数据集，我们在不同的领域自适应算法中使用相同的架构。

数据增强 数据增强是训练视觉模型的标准组成部分，能够近似不同领域之间的某些差异。除非另有说明，我们使用标准的强数据增强技术训练所有视觉数据集：随机水平翻转、随机裁剪、Cutout 增强（DeVries与Taylor，2017年）和 RandAugment（Cubuk等人，2020年）。为了了解仅靠数据增强所带来的效果，我们还测试了未使用任何数据增强训练的仅源域模型。对于表格数据和自然语言处理数据集，我们不使用任何数据增强。

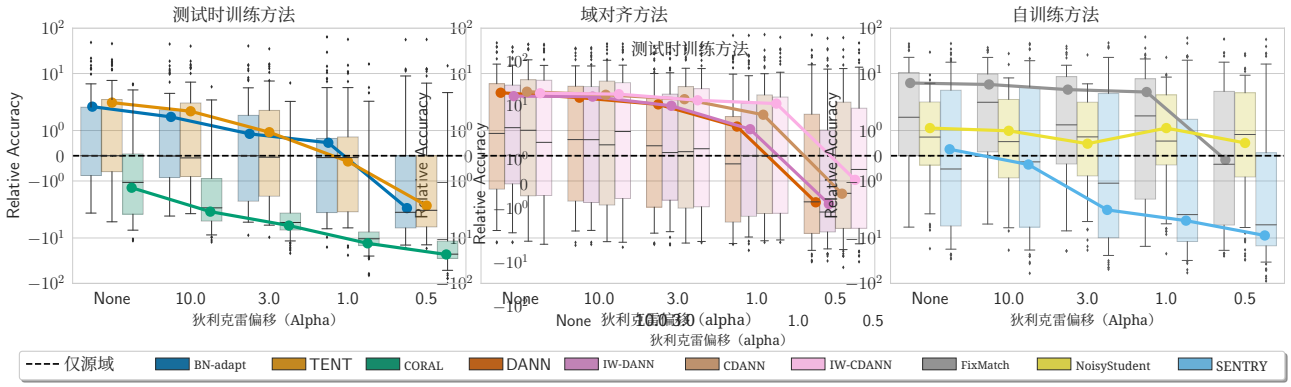
4. 主要结果

我们在图 2 中展示了测试平台中视觉数据集的汇总结果。在附录 B 中，我们展示了自然语言处理和表格数据集的汇总结果。每个数据集的结果见附录 J。请注意，我们未报告仅源域模型的重采样结果，因为该模型仅使用源数据训练；在实验中，我们发现，与仅对源数据进行平衡相比，两者没有差异（因为大多数数据集的源域本身已经平衡）。除非另有说明，我们使用源域验证性能作为早停标准。基于对整个 RLSBENCH 套件的运行结果，我们将发现归纳为以下几点。

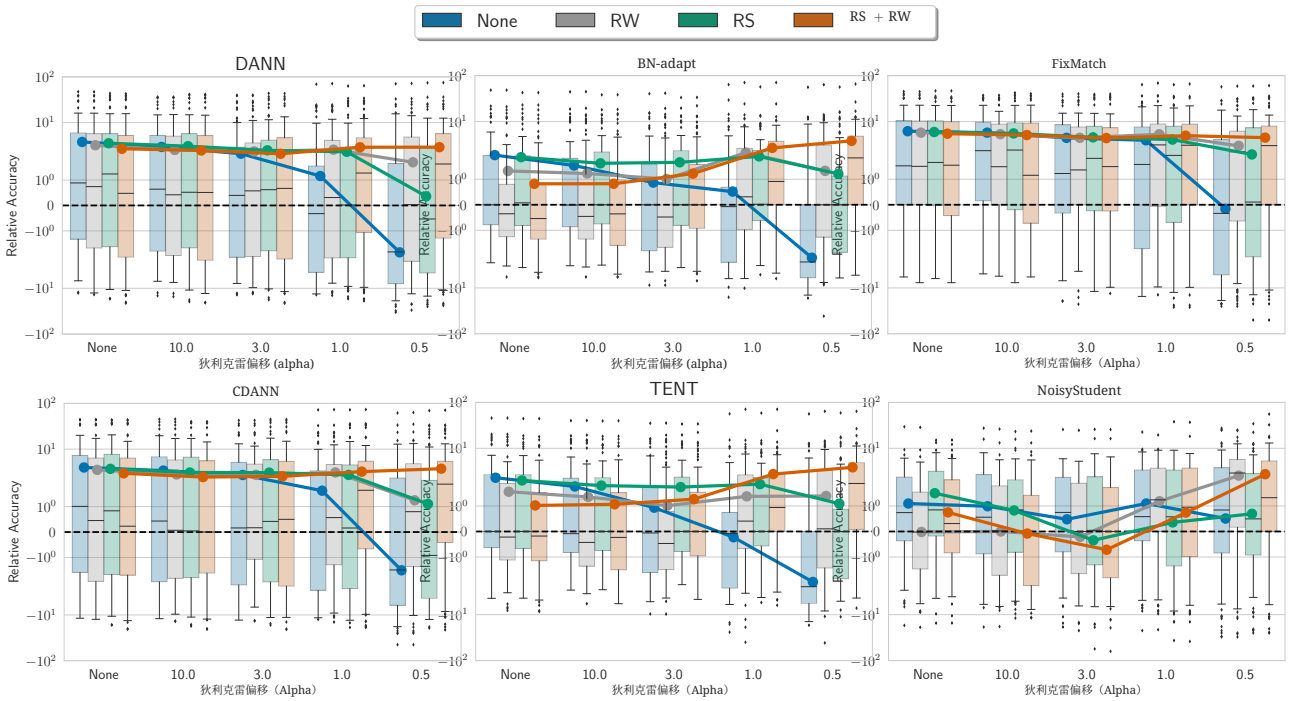
热门深度域自适应方法若不进行任何校正，表现不佳。 当目标标签边缘分布偏移不存在或较小时，领域自适应方法通常优于仅源域分类器；然而，当目标标签边缘分布偏移严重时（即，$\alpha = 0.5$ 图 2a、5a 和6a 所示），除 Noisy Student 外，这些方法的性能会低于仅源域分类器。另一方面，领域自适应方法结合重采样和重加权校正后，即使目标标签边缘分布偏移严重，也能显著优于仅源模型（图 2b、5b 和6b）。

对目标域进行重采样以实现伪平衡，通常有助于提升所有模态下的所有领域自适应方法。 当目标标签边缘分布的偏移不存在或非常小时（即 $\alpha \in \{\text{NONE}\ 10.0\}$见图 2b、5b 和 6b），重采样前后的性能没有（显著）差异。然而，随着目标标签边缘分布的偏移严重程度增加（即 $\alpha \in \{3.0, 1。0, 0.5\}$ 见图 2b、5b 和 6b），我们观察到，重采样通常能提升测试平台中的所有领域自适应方法的性能。

事后重加权分类器的效果取决于偏移严重程度和底层领域自适应算法。 对于域对齐方法（即 DANN 和 CDANN）以及自训练方法，尤其是 FixMatch 和 PseudoLa-bel，我们观察到，当目标标签边缘分布偏移严重时（即 $\alpha \in \{3.0, 1。0, 0.5\}$ 图 2b、5b 和6b），RW 校正通常会带来显著提升（相较于不进行校正）；而当目标标签边缘分布偏移不存在或非常小时（即 $\alpha \in \{\text{NONE}\ 10.0\}$图 2b、5b 和 6b），则没有（显著）影响。对于 BN-adapt、TENT 和 NoisyS-



(a) 随着目标标签边缘分布偏移程度增加，领域自适应方法相对于仅源域训练的性能



(b) 领域自适应方法与我们的元算法（RS 和 RW 校正）结合时，相对于仅源域训练的性能

图 2. 视觉数据集中所有分布偏移对上的不同领域自适应方法相对于仅源模型的性能，按标签边缘分布偏移程度分组。对于每个分布偏移对和领域自适应方法，我们绘制采用该领域自适应方法训练的模型的相对准确率，即用该模型的准确率减去仅源模型的准确率。因此，位于 0 处的黑色虚线表示仅源模型的性能。狄利克雷偏移参数越小，目标类别比例偏移越严重。 **(a)** 偏移程度为 $\alpha = \{\text{None}10.0, 3.0\}$ 对不同领域自适应方法几乎没有影响，而当 $\alpha \in \{1.0, 0.5\}$ 时，所有领域自适应方法的性能都会下降，且往往低于仅源域分类器的性能（Noisy Student 除外）。 **(b)** 在我们的元算法中，RS 和 RW 结合使用，在所有领域自适应方法上都能显著提升总体性能，优于不进行校正。虽然 RS 在不同标签边缘分布偏移程度下均能持续带来帮助（相较于不进行校正），但当偏移程度较小时，RW 会使 BN-adapt、TENT 和 NoisyStudent 的性能略有下降。然而，对于严重偏移（$\alpha \in \{3.0, 1.0, 0.5\}$），RW 能显著提升所有方法的性能。表格数据集和语言数据集上的类似结果见附录 B。单个数据集上所有方法的详细结果见附录 J。绘图方法的更详细说明见附录 A。

tudent，当目标标签边缘分布偏移不存在或较小时，RW 校正可能会略微降低效果（即 $\alpha \in \{\text{None}\ 10.0\}$ 图 2b），但当目标标签边缘分布偏移严重时（即 $\alpha \in \{3.0,\ 1.0,\ 0.5\}$ 图中图 2b）。此外，我们观察到，在真实世界偏移的某些特定情形下，$p(x|y)$（例如 BREEDs 数据集中的子群体偏移、Camelyon 偏移，以及 CIFAR-10 中的复现研究；相较于其他视觉数据集，这些偏移较为温和

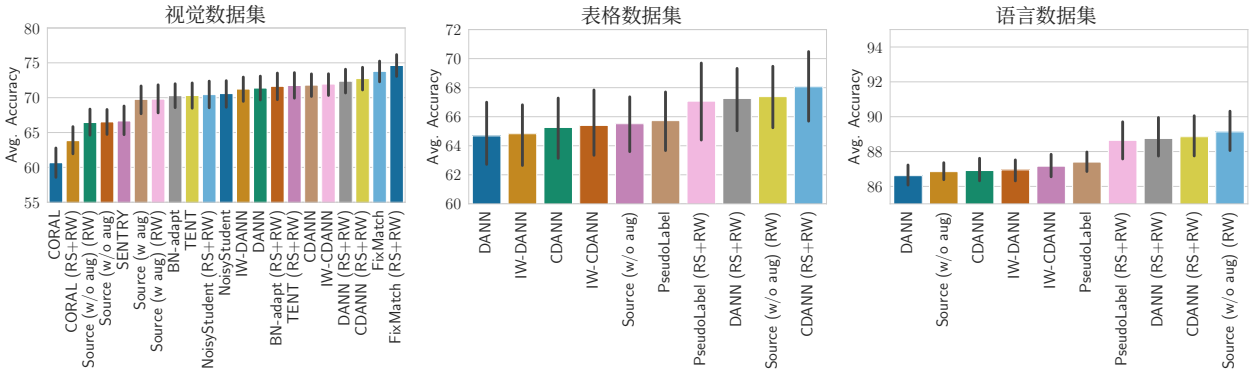


图3。 各模态中不同领域自适应方法在所有分布对上的平均准确率。 附录中给出了各方法在单个数据集上的对应结果。 J。

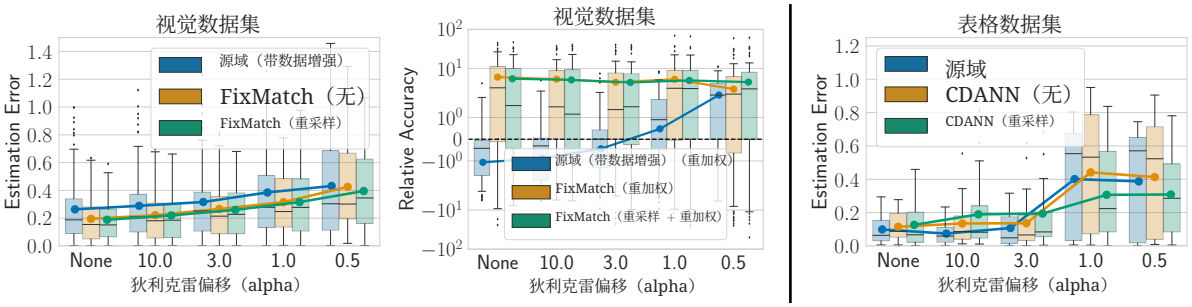


图4。目标标签边缘分布估计（ℓ_1）误差，以及 *RLLS* 与采用不同领域自适应方法获得的分类器的准确率。 **（左）**在视觉数据集的所有偏移程度下，使用领域自适应方法获得的分类器的 RLLS 表现优于使用仅源域分类器的 RLLS。 **（右）** 对于表格数据集，在目标标签边缘分布发生严重偏移时，使用领域自适应方法获得的分类器的 RLLS 表现优于使用仅源域分类器的 RLLS。附录中提供了各领域自适应方法和所有数据集的图。 H。

在我们的测试平台中出现的偏移），RW 校正不会损害 BN-adapt、TENT 和 NoisyStudent 的性能，即使目标标签边缘分布偏移较轻微或不存在（参见附录 J 中的数据集）。

与我们的元算法结合时，领域自适应方法通常优于仅源域分类器，但没有哪种方法始终表现最佳。 首先，我们发现仅源域模型的结果优于此前发表的结果。与以往研究（Gulrajani & Lopez-Paz，2020年）类似，这可以归因于设计选择的改进（例如数据增强、超参数），而我们在所有方法中统一采用了这些选择。虽然没有一种方法能在所有数据集上始终表现最佳，但总体而言，采用重采样和重加权（我们的元算法）的 FixMatch 在视觉数据集上表现最佳。在自然语言处理数据集上，采用重加权（我们的元算法）的仅源域方法总体表现最佳。在表格数据集上，采用重采样和重加权（我们的元算法）的 CDANN 总体表现最佳（图3）。

现有域适应方法与我们的元算法结合后，显著优于其他专为宽松标签偏移提出的域适应方法。 我们观察到，在不同方法采用一致的实验设计时，采用 RS 和 RW 校正的现有域适应方法通常优于先前提出的专门用于应对宽松标签偏移的方法，即 IW-CDANN、IW-DANN 和 SENTRY（图 7）。当目标域标签边际分布发生严重偏移时，IW-DANN、IW-CDANN 和 SENTRY 的表现往往不如仅使用源域的模型。此外，虽然重要性加权（即 IW-CDANN 和 IW-DANN）分别优于 CDANN 和 DANN（图 2a、5a 和 6a），但 RS 和 RW 校正带来的提升远超这些改进（图 7）。

使用我们的元算法的 BN-adapt 和 TENT 是简单而强大的基线方法。 对于带有批归一化参数的模型，采用 RS 和 RW 步骤的 BN-adapt（和 TENT）是一种计算效率高且效果强大的基线方法。我们观察到，当目标标签边缘分布发生偏移时（即 $\alpha \in \{1.0,\ 0.5\}$ 图2(a) 中），BN-adapt（和 TENT）的性能可能大幅下降；而 RS 和 RW 校正则能提升性能，并且当目标标签边缘分布偏移极端时（即 $\alpha = 0.5$ 图 2(b) 中），BN-adapt（和 TENT）的表现往往优于所有其他领域自适应方法。

域适应方法能够更准确地估计目标标签边缘分布，因此相比仅源域分类器，能带来更大的准确率提升（通过重新加权）。

重新加权，而非仅源域分类器。请回想一下，我们实验了利用现成分类器获得估计值的目标标签边缘分布估计方法。我们观察到，对于表格数据集和视觉数据集，利用领域自适应分类器的估计器往往比使用仅源域分类器的效果更好（图4）。对于自然语言处理，我们观察到领域自适应分类器与仅源域分类器的表现相当（仅源域分类器通常略胜一筹）。相应地，正如预期，更准确的估计在应用我们的 RW 校正时会带来更大的准确率提升。具体而言，在视觉数据集和表格数据集上，采用域适应方法的 RW 校正优于仅源域分类器；而在自然语言处理数据集上则相反。（图4）。

早停准则很重要。我们观察到，在所有方法下，视觉和表格数据集分别存在一致的 ≈2% 和 ≈8% 准确率差异（图 14）。在自然语言处理数据集上，尽管未采用 RW 和 RS 校正时，早停准则会导致 ≈2% 的准确率差异，但采用这些校正后，差异变得可以忽略不计（图 14）。这些结果表明，后续研究应描述其评估中所使用的早停准则。

数据增强有所帮助。与此前在其他场景中的研究发现一致（Gulrajani & Lopez-Paz，2020年；Sagawa 等人，2021年），我们观察到，在宽松标签偏移场景下，数据增强可以提升仅源模型在视觉数据集上的性能（参见附录 J中各数据集的结果）。因此，只要适用，后续方法都应使用数据增强。

5. 结论

我们的工作是首个在宽松标签偏移场景下对相关方法开展大规模研究的工作。相较于严格遵循标签偏移假设的研究，RLSBENCH 提供了进行敏感性分析的机会，使研究人员能够衡量其方法在类别条件分布受到各种扰动时的鲁棒性。相较于基准驱动的深度领域自适应文献，我们提供了一套全面且标准化的评估方案，用于评估标签分布发生偏移时的表现，使这些基准测试更接近于展现我们在实际部署模型时所应预期会遇到的多样性。一方面，观察到标签偏移校正带来的稳定改进令人鼓舞。与此同时，鉴于该问题的定义并不充分，实践者必须保持警惕，不应盲目看待任何基准测试上的性能，而应考虑它是否以及在多大程度上代表了其具体应用中可能出现的各种情形。

未来，我们希望将 RLSBENCH 扩展到来自医疗保健和自动驾驶等重要领域实际应用的数据集，在这些领域中，标签边缘分布和类别条件分布可能会因地点和时间而发生变化。我们还希望纳入自监督方法，通过代理任务（如重构（Gidaris 等人，2018年；He 等人，2022年）和对比学习（Caron 等人，2020年；Chen 等，2020年）来学习表征，训练时使用源域与目标域无标签数据的并集。尽管利用目标标签分布估计值对预测进行重新加权能带来显著提升，但我们的结果与预言机性能之间仍存在差距，这应当推动未来工作致力于改进估计器。此外，我们观察到，目标标签边缘分布估计技术的效果取决于偏移的性质 $p(x|y)$。当标签偏移假设不成立时，从数学上刻画标签偏移估计技术的行为，将是一项重要贡献。

可复现性声明

我们将在 GitHub 上发布包含所有结果的代码。`https://github.com/acmi-lab/RLSbench`。我们使用RLSBENCH 库，并基于 PyTorch（Paszke 等人，2017年）实现了该库，同时提供了用于运行所有实验并生成相应结果的基础设施。我们保存了所有模型并记录了所有超参数，以便于复现。在附录中，我们提供了有关数据集和实验的更多细节。在附录D中，我们介绍数据集；在附录M中，我们提供超参数的详细信息。

致谢

感谢 Amrith Setlur、Pratyush Maini 和 Aditi Raghunathan 对 RLSbench 早期稿件提出宝贵意见。我们还要感谢 Xingjian Shi 和 Weisu Yin 在运行大规模实验初期提供的帮助。SG 感谢亚马逊研究生奖学金和摩根大通人工智能博士奖学金的支持。

参考文献

Amr Alexandari、Anshul Kundaje 和 Avanti Shrikumar。《通过偏差校正校准适应标签偏移》。发表于国际机器学习会议（*ICML*），2021年。

Jing An、Lexing Ying 和 YuhuaZhu。《为什么在使用随机梯度校正抽样偏差时，重采样优于重新加权》。*arXiv* 预印本 *arXiv:2009.13447*，2020年。

Kamyar Azizzadenesheli、Anqi Liu、Fanny Yang 和 Ani-mashree Anandkumar。标签偏移下领域自适应的正则化学习。发表于 国际学习表征会议 *(ICLR)*，2019年。

Christina Baek、Yiding Jiang、Aditi Raghunathan 和 Zico Kolter。Agreement-on-the-line：预测分布偏移下神经网络的性能。*arXiv* 预印本 *arXiv:2206.13089*，2022年。

Peter Bandi、Oscar Geessink、Quirine Manson、Mar- cory Van Dijk、Maschenka Balkenhol、Meyke Hermsen、Babak Ehteshami Bejnordi、Byungjae Lee、Kyunghyun Paeng、Aoxiao Zhong 等。从单个转移灶的检测到患者层面的淋巴结状态分类：camelyon17 挑战赛。*IEEE* 医学影像汇刊，2018年。

Jessa Bekker 和 Jesse Davis。正例与无标签数据学习：综述。机器学习，2020年。网址：https://doi.org/10.1007%2Fs10994-020-05877-5。Shai Ben-David、John Blitzer、Koby Crammer、Alex Kulesza、Fernando Pereira 和 Jennifer Wortman Vaughan。跨域学习理论。机器学习，79(1-2)，2010a。Shai Ben-David、Tyler Lu、Teresa Luu 和 D´avid P´al。领域自适应的不可能性定理。载于人工智能与统计国际会议（*AISTATS*），2010b。David Berthelot、Rebecca Roelofs、Kihyuk Sohn、Nicholas Carlini 和 Alex Kurakin。Adamatch：半监督学习与领域自适应的统一方法。*arXiv* 预印本 *arXiv:2106.04732*，2021年。Gilles Blanchard、Gyemin Lee 和 Clayton Scott。将多个相关分类任务的学习泛化到新的无标签样本。神经信息处理系统进展，24，2011年。Daniel Borkan、Lucas Dixon、Jeffrey Sorensen、Nithum Thain 和 Lucy Vasserman。利用真实数据衡量文本分类中非预期偏差的精细指标。载于*2019*年万维网大会附属论文集，2019年。Mateusz Buda、Atsuto Maki 和 Maciej A Mazurowski。卷积神经网络中类别不平衡问题的系统性研究。神经网络，106：249–259，2018年。Jonathon Byrd 和 Zachary C Lipton。深度学习中重要性加权有何影响？载于国际机器学习会议（*ICML*），2019年。Kaidi Cao、Colin Wei、Adrien Gaidon、Nikos Arechiga 和 Tengyu Ma。利用标签分布感知边际损失学习不平衡数据集。载于神经信息处理系统进展，第32卷，2019a。

Zhangjie Cao、Kaichao You、Mingsheng Long、Jianmin Wang 和 Qiang Yang。为部分领域自适应学习迁移样本。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2985–2994页，2019b。

Mathilde Caron、Ishan Misra、Julien Mairal、Priya Goyal、Pi- otr Bojanowski 和 Armand Joulin。通过对比簇分配进行视觉特征的无监督学习。神经信息处理系统进展，33：9912–9924，2020年。

Nitesh V Chawla、Kevin W Bowyer、Lawrence O Hall 和 W Philip Kegelmeyer。SMOTE：合成少数类过采样技术。人工智能研究杂志，16：321–357，2002年。Jiefeng Chen、Frederick Liu、Besim Avci、Xi Wu、Yingyu Liang 和 Somesh Jha。利用自训练集成检测错误并估计无标签数据的准确率。神经信息处理系统进展，34：14980–14992，2021年。Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议，第1597–1607页。PMLR，2020年。Gordon Christie、Neil Fendley、James Wilson 和 Ryan Mukherjee。世界功能地图。载于*IEEE*计算机视觉与模式识别会议论文集，2018年。Corinna Cortes 和 Mehryar Mohri。回归的领域自适应与样本偏差校正理论及算法。理论计算机科学，519，2014年。Corinna Cortes、Yishay Mansour 和 Mehryar Mohri。重要性加权的学习界。载于神经信息处理系统进展（*NIPS*），2010年。Ekin D Cubuk、Barret Zoph、Jonathon Shlens 和 Quoc V Le。RandAugment：通过缩小搜索空间实现实用的自动数据增强。载于*IEEE/CVF*计算机视觉与模式识别会议研讨会论文集，第702–703页，2020年。Weijian Deng 和 Liang Zheng。分类器准确率评估是否总是需要标签？载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第15069–15078页，2021年。Terrance DeVries 和 Graham W Taylor。利用 Cutout 改进卷积神经网络的正则化。*arXiv*预印本 *arXiv:1708.04552*，2017年。Frances Ding、Moritz Hardt、John Miller 和 Ludwig Schmidt。告别 Adult：用于公平机器

学习。神经信息处理系统进展，34：6478–6490，2021年。

Josip Djolonga、Jessica Yung、Michael Tschannen、Rob Romijnders、Lucas Beyer、Alexander Kolesnikov、Joan Puigcerver、MatthiasMinderer、AlexanderD'Amour、Dan Moldovan 等。论卷积神经网络的鲁棒性与可迁移性。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16458–16468 页，2021年。

Charles Elkan 和 Keith Noto。仅利用正例与无标签数据学习分类器。载于国际知识发现与数据挖掘会议（*KDD*），第 213–220 页，2008年。

Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pascal Germain、Hugo Larochelle、Franc¸ois Laviolette、Mario Marchand 和 Victor Lempitsky。神经网络的领域对抗训练。机器学习研究杂志，2021年。Jacob Gardner、Geoff Pleiss、Kilian Q Weinberger、David Bindel 和 Andrew G Wilson。GPyTorch：通过 GPU 加速进行黑盒矩阵乘法高斯过程推断。载于神经信息处理系统进展（*NeurIPS*），2018年。Saurabh Garg、Yifan Wu、Sivaraman Balakrishnan 和 Zachary Lipton。标签偏移估计的统一视角。载于神经信息处理系统进展（*NeurIPS*），2020年。Saurabh Garg、Yifan Wu、Alex Smola、Sivaraman Balakrishnan 和 Zachary Lipton。混合比例估计与 PU 学习：一种现代方法。载于神经信息处理系统进展（*NeurIPS*），2021年。Saurabh Garg、Sivaraman Balakrishnan 和 Zachary Lipton。开放集标签偏移下的领域自适应。载于神经信息处理系统进展（*NeurIPS*），2022a。Saurabh Garg、Sivaraman Balakrishnan、Zachary Lipton、Behnam Neyshabur 和 Hanie Sedghi。利用未标注数据预测分布外性能。载于国际学习表征会议（*ICLR*），2022b。Spyros Gidaris、Praveer Singh 和 Nikos Komodakis。通过预测图像旋转进行无监督表征学习。*arXiv* 预印本 *arXiv:1803.07728*，2018年。Arthur Gretton、Alexander J Smola、Jiayuan Huang、Marcel Schmittfull、Karsten M Borgwardt 和 Bernhard Sch¨olkopf。通过核均值匹配处理协变量偏移。机器学习研究杂志（*JMLR*），2009年。

Devin Guillory、Vaishaal Shankar、Sayna Ebrahimi、Trevor Darrell 和 Ludwig Schmidt。对未见分布进行置信预测。*arXiv* 预印本 *arXiv:2107.03315*，2021年。

Ishaan Gulrajani 和 David Lopez-Paz。寻找丢失的领域泛化。*arXiv* 预印本 *arXiv:2007.01434*，2020年。

何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于计算机视觉与模式识别（*CVPR*），2016年。

何恺明、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross Girshick。掩码自编码器是可扩展的视觉学习器。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16000–16009 页，2022年。Dan Hendrycks 和 Thomas Dietterich。对神经网络应对常见损坏和扰动的鲁棒性进行基准测试。*arXiv* 预印本 *arXiv:1903.12261*，2019年。Dan Hendrycks、Steven Basart、Norman Mu、Saurav Ka- davath、Frank Wang、Evan Dorundo、Rahul Desai、Tyler Zhu、Samyak Parajuli、Mike Guo 等。鲁棒性的多重面貌：对分布外泛化的批判性分析。载于*IEEE/CVF*国际计算机视觉会议论文集，第 8340–8349 页，2021年。Gao Huang、Zhuang Liu、Laurens Van Der Maaten 和 Kilian Q Weinberger。密集连接卷积网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 4700–4708 页，2017年。Badr Youbi Idrissi、Martin Arjovsky、Mohammad Pezeshki 和 David Lopez-Paz。简单数据平衡可实现有竞争力的最差群体准确率。载于因果学习与推理会议，第 336–351 页。PMLR，2022年。Junguang Jiang、Yang Shu、Jianmin Wang 和 Mingsheng Long。深度学习中的可迁移性：综述，2022年。Yiding Jiang、Vaishnavh Nagarajan、Christina Baek 和 J Zico Kolter。通过分歧评估随机梯度下降的泛化能力。*arXiv* 预印本 *arXiv:2106.13799*，2021年。Fredrik D. Johansson、David Sontag 和 Rajesh Ranganath。域不变表示中的支撑集与可逆性。载于 Kamalika Chaudhuri 和 Masashi Sugiyama（编），第二十二届国际人工智能与统计学会议论文集，《机器学习研究论文集》。PMLR，2019年。

Alistair Johnson、Lucas Bulgarelli、Tom Pollard、Steven Horng、Leo Anthony Celi 和 Roger Mark。Mimic- iv。*PhysioNet*。在线地址：*https://physionet.org/content/mimiciv/1.0/*（访问于*2021*年*8*月*23*日），2020年。

Pang Wei Koh、Shiori Sagawa、Henrik Marklund、Sang Michael Xie、Marvin Zhang、Akshay Balsubramani、Weihua Hu、Michihiro Yasunaga、Richard Lanas Phillips、Irena Gao、Tony Lee、Etienne David、Ian Stavness、Wei Guo、Berton A. Earnshaw、Imran S. Haque、Sara Beery、Jure Leskovec、Anshul Kundaje、Emma Pierson、Sergey Levine、Chelsea Finn 和 Percy Liang。WILDS：野外分布偏移基准。在国际机器学习会议 *(ICML)*，2021年。

Alex Krizhevsky 和 Geoffrey Hinton。从 Tiny Images 中学习多层特征。技术报告，Citeseer，2009年。

Ananya Kumar、Tengyu Ma 和 Percy Liang。理解渐进式领域自适应中的自训练。在国际机器学习会议，第 5468–5479 页。PMLR，2020年。

Dong-Hyun Lee 等人。伪标签：一种简单高效的深度神经网络半监督学习方法。载于 机器学习领域挑战研讨会，*ICML*，第 3 卷，第 896 页，2013年。

Yanghao Li、Naiyan Wang、Jianping Shi、Jiaying Liu 和 Xiaodi Hou。重新审视批归一化在实际领域自适应中的应用。*arXiv* 预印本 *arXiv:1603.04779*，2016年。

Zachary C Lipton、Yu-Xiang Wang 和 Alex Smola。使用黑盒预测器检测并纠正标签偏移。载于国际机器学习会议（*ICML*），2018年。

Xiaofeng Liu、Zhenhua Guo、Site Li、Fangxu Xing、Jane You、C.-C. Jay Kuo、Georges El Fakhri 和 Jonghye Woo。结合条件偏移与标签偏移的对抗式无监督领域自适应：推断、对齐与迭代。载于 *2021*年 *IEEE/CVF* 计算机视觉国际会议（*ICCV*），第 10347–10356 页，加拿大魁北克省蒙特利尔，2021年10月。IEEE。ISBN 978-1-66542-812-5。doi: 10.1109/ICCV48922.2021.01020。
URL https://ieeexplore.ieee.org/document/9710205/。

Mingsheng Long、Yue Cao、Jianmin Wang 和 Michael Jor- dan。利用深度适应网络学习可迁移特征。载于国际机器学习会议，第 97–105 页。PMLR，2015年。

Mingsheng Long、Han Zhu、Jianmin Wang 和 Michael I Jordan。利用联合适应网络进行深度迁移学习。载于国际机器学习会议。PMLR，2017年。

Mingsheng Long、Zhangjie Cao、Jianmin Wang 和 MichaelI Jordan。条件对抗域适应。神经信息处理系统进展，31，2018年。

Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。*arXiv* 预印本 *arXiv:1706.06083*，2017年。

Jeroen Manders、Twan van Laarhoven 和 Ele na Mar- chiori。《用于领域自适应的类别预测不确定性对抗对齐》，2019年1月。网址 http://arxiv.org/abs/1804.04448。编号：arXiv:1804.04448 arXiv:1804.04448 [cs, stat]。Yishay Mansour、Mehryar Mohri 和 Afshin Rostamizadeh。《领域自适应：学习界与算法》。*arXiv*预印本 *arXiv:0902.3430*，2009年。John P Miller、Rohan Taori、Aditi Raghunathan、Shiori Sagawa、Pang Wei Koh、Vaishaal Shankar、Percy Liang、Yair Carmon 和 Lud wig Schmidt。《准确率表现：分布外泛化与分布内泛化之间的强相关性》。载于国际机器学习会议。PMLR，2021年。Krikamol Muandet、David Balduzzi 和 Bernhard Sch¨olkopf。《通过不变特征表示实现领域泛化》。载于国际机器学习会议，第10–18页。PMLR，2013年。Adam Paszke、Sam Gross、Soumith Chintala、Gregory Chanan、Edward Yang、Zachary DeVito、Zeming Lin、Al- ban Desmaison、Luca Antiga 和 Adam Lerer。《PyTorch中的自动微分》。2017年。Xingchao Peng、Ben Usman、Neela Kaushik、Judy Hoffman、Dequan Wang 和 Kate Saenko。《VisDA：视觉域适应挑战赛》，2017年。Xingchao Peng、Ben Usman、Kuniaki Saito、Neela Kaushik、Judy Hoffman 和 Kate Saenko。《Syn2Real：合成到真实视觉域适应的新基准》。*arXiv*预印本 *arXiv:1806.09755*，2018年。Xingchao Peng、Qinxun Bai、Xide Xia、Zijun Huang、Kate Saenko 和 Bo Wang。《多源域适应的矩匹配》。载于 *IEEE/CVF*国际计算机视觉会议论文集，第1406–1415页，2019年。

PhysioToolkit PhysioBank。Physionet：复杂生理信号新型研究资源的组成部分。《*Circulation*》，101(23):e215–e220，2000年。

Viraj Prabhu、Shivam Khare、Deeksha Kartik 和 Judy Hoffman。SENTRY：通过委员会一致性进行选择性熵优化，以实现无监督域适应。载于*IEEE/CVF* 计算机视觉国际会议论文集，第 8558–8567 页，2021年。

Joaquin Quinonero-Candela、Masashi Sugiyama、Anton Schwaighofer 和 Neil D Lawrence。机器学习中的数据集偏移。MIT出版社，2008年。

Benjamin Recht、Rebecca Roelofs、Ludwig Schmidt 和 Vaishaal Shankar。CIFAR-10 分类器能否泛化到 CIFAR-10？*arXiv* 预印本 *arXiv:1806.00451*，2018年。

Benjamin Recht、Rebecca Roelofs、Ludwig Schmidt 和 Vaishaal Shankar。ImageNet 分类器能否泛化到 I mageNet？载于国际机器学习会议，第 5389–5400 页。PMLR，2019年。

Manley Roberts、Pranav Mani、Saurabh Garg 和 Zachary Lipton。潜在标签偏移下的无监督学习。载于神经信息处理系统进展（*NeurIPS*），2022年。

Elan Rosenfeld、Pradeep Ravikumar 和 Andrej Risteski。域调整回归，或者说：经验风险最小化可能已经学到了足以实现分布外泛化的特征。*arXiv*预印本 *arXiv:2202.06856*，2022年。

Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、San- jeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpa- thy、Aditya Khosla、Michael Bernstein 等。ImageNet大规模视觉识别挑战赛。《国际计算机视觉杂志》，115(3):211–252，2015年。

MarcoSaerens、Patrice Latinne 和 ChristineDecaestecker。将分类器的输出调整为新的先验概率：一种简单方法。《神经计算》，2002年。

Shiori Sagawa、Pang Wei Koh、Tony Lee、Irena Gao、Sang Michael Xie、Kendrick Shen、Ananya Kumar、Wei- hua Hu、Michihiro Yasunaga、Henrik Marklund、Sara Beery、Etienne David、Ian Stavness、Wei Guo、Jure Leskovec、Kate Saenko、Tatsunori Hashimoto、Sergey Levine、Chelsea Finn 和 Percy Liang。《扩展用于无监督适应的WILDS基准》。发表于*NeurIPS*分布偏移研讨会，2021年。

Kuniaki Saito、Kohei Watanabe、Yoshitaka Ushiku 和 Tat- suya Harada。用于无监督领域自适应的最大分类器差异。载于*IEEE*计算机视觉与模式识别会议论文集，2018年。

Kuniaki Saito、Donghyun Kim、Piotr Teterwak、Stan Sclaroff、Trevor Darrell 和 Kate Saenko。以正确的方式进行调优：通过软邻域密度进行领域自适应的无监督验证。载于*IEEE/CVF*国际计算机视觉会议论文集，第9184–9193页，2021年。Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量。*arXiv*，abs/1910.01108，2019年。

Shibani Santurkar、Dimitris Tsipras 和 Aleksander Madry。《BREEDs：子群体偏移基准测试》。发表于国际学习表征会议（*ICLR*），2021年。

Steffen Schneider、Evgenia Rusak、Luisa Eck、Oliver Bring- mann、Wieland Brendel 和 Matthias Bethge。《通过协变量偏移适应提高对常见损坏的鲁棒性》。*arXiv*预印本 *arXiv:2006.16971*，2020年。

Bernhard Sch¨olkopf、Dominik Janzing、Jonas Peters、Eleni Sgouritsa、Kun Zhang 和 Joris Mooij。《因果与反因果学习》。载于国际机器学习会议 *(ICML)*，2012年。

Hidetoshi Shimodaira。通过对数似然函数加权，改善协变量偏移下的预测推断。《统计规划与推断杂志》，2000年。

Kihyuk Sohn、David Berthelot、Nicholas Carlini、Zizhao Zhang、Han Zhang、Colin A Raffel、Ekin Dogus Cubuk、Alexey Kurakin 和 Chun-Liang Li。FixMatch：通过一致性和置信度简化半监督学习。神经信息处理系统进展，33，2020年。

Baochen Sun 和 Kate Saenko。Deep CORAL：用于深度领域自适应的相关性对齐。载于欧洲计算机视觉会议。施普林格，2016年。

Baochen Sun、Jiashi Feng 和 Kate Saenko。《令人沮丧地简单的领域自适应的回归》。载于*AAAI*人工智能会议论文集，2016年。

Baochen Sun、Jiashi Feng 和 Kate Saenko。《无监督域适应的相关性对齐》。载于《计算机视觉应用中的域适应》。施普林格，2017年。

RemiTachet, Han Zhao, Yu-XiangWang 和 Geoff Gordon。通过条件分布匹配和广义标签偏移实现领域自适应。*arXiv:2003.04475 [cs,stat]*, 2020年12月。网址 http://arxiv.org/abs/2003.04475。arXiv: 2003.04475。

Remi Tachet des Combes、Han Zhao、Yu-Xiang Wang 和 Geoffrey J Gordon。通过条件分布匹配和广义标签偏移实现领域自适应。神经信息处理系统进展，33，2020年。

Shuhan Tan、Xingchao Peng 和 Kate Saenko。类别不平衡的域适应：一场实证探索。载于欧洲计算机视觉会议，第 585–602 页。施普林格，2020年。

Rohan Taori、Achal Dave、Vaishaal Shankar、Nicholas Carlini、Benjamin Recht 和 Ludwig Schmidt。衡量图像分类对自然分布偏移的鲁棒性。神经信息处理系统进展，33：18583–18599，2020年。

Antonio Torralba 和 Alexei A Efros。审视数据集偏差的无偏视角。载于*CVPR 2011*，第 1521–1528 页。IEEE，2011年。

Antonio Torralba、Rob Fergus 和 William T. Freeman。8000万张微型图像：用于非参数物体和场景识别的大型数据集。*IEEE* 模式分析与机器智能汇刊，30(11)：1958–1970，2008年。

Vladimir N Vapnik。统计学习理论概述。*IEEE*神经网络汇刊，10(5):988– 999，1999年。

Hemanth Venkateswara、Jose Eusebio、Shayok Chakraborty 和 Sethuraman Panchanathan。用于无监督域适应的深度哈希网络。载于*IEEE*计算机视觉与模式识别会议论文集，第 5018–5027 页，2017年。

Dequan Wang、Evan Shelhamer、Shaoteng Liu、Bruno Ol- shausen 和 Trevor Darrell。TENT：通过熵最小化实现完全测试时自适应。载于 国际学习表征会议，2021年。网址：https://openreview.net/forum?id=uXl3bZLkr3c。

Chen Wei、Kihyuk Sohn、Clayton Mellina、Alan Yuille 和 Fan Yang。Crest：面向不平衡半监督学习的类别再平衡自训练框架。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10857–10866 页，2021年。

Florian Wenzel、Andrea Dittadi、Peter Vincent Gehler、Carl-Johann Simon-Gabriel、Max Horn、Dominik Ziet- low、David Kernert、Chris Russell、Thomas Brox、Bernt Schiele 等。迁移学习中的分布外泛化评估。*arXiv* 预印本 *arXiv:2207.09239*，2022年。

Olivia Wiles、Sven Gowal、Florian Stimberg、Sylvestre Alvise-Rebuffi、Ira Ktena、Taylan Cemgil 等。对分布偏移的细粒度分析。*arXiv* 预印本 *arXiv:2110.11328*，2021年。

Garrett Wilson 和 Diane J Cook。无监督深度域适应综述。*ACM*智能系统与技术汇刊 *(TIST)*，2020年。

Yifan Wu、Ezra Winston、Divyansh Kaushik 和 Zachary Lipton。采用非对称松弛分布对齐的领域自适应。载于国际机器学习会议 *(ICML)*，2019年。

Qizhe Xie、Minh-Thang Luong、Eduard Hovy 和 Quoc V Le。采用 Noisy Student 自训练提升 ImageNet 分类性能。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10687–10698 页，2020a。

Xinpeng Xie、Jiawei Chen、Yuexiang Li、Linlin Shen、Kai Ma 和 Yefeng Zheng。用于保持目标的图像到图像域适应的自监督CycleGAN。在计算机视觉—*ECCV 2020*：第*16*届欧洲计算机视觉会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第*XX*卷第*16*册，第498–513页。施普林格，2020b。

Da Xu、Yuting Ye 和 Chuanwei Ruan。理解重要性加权在深度学习中的作用。*arXiv* 预印本 *arXiv:2103.15209*，2021年。

Hongliang Yan、Yukang Ding、Peihua Li、Qilong Wang、Yong Xu 和 Wangmeng Zuo。《注意类别权重偏差：用于无监督领域自适应的加权最大均值差异》。载于*IEEE* 计算机视觉与模式识别会议论文集，第 2272–2281 页，2017年。

Huaxiu Yao、Caroline Choi、Bochuan Cao、Yoonho Lee、Pang Wei Koh 和 Chelsea Finn。《Wild-Time：一个用于评估真实环境中随时间变化的分布偏移的基准测试》。载于第三十六届神经信息处理系统会议：数据集与基准测试专题，2022年。网址：https://openreview.net/forum?id=F˙9ENmZABB0。

Bianca Zadrozny。《样本选择偏差下分类器的学习与评估》。载于国际机器学习会议（*ICML*），2004年。

Jingzhao Zhang、Aditya Menon、AndreasVeit、Srinadh Bhojanapalli、Sanjiv Kumar 和 Suvrit Sra。《通过分布鲁棒优化应对标签偏移》。载于国际学习表征会议（*ICLR*），2021年。

Kun Zhang、Bernhard Sch¨olkopf、Krikamol Muandet 和 Zhikun Wang。《目标域与条件偏移下的领域自适应》。载于国际机器学习会议（*ICML*），2013年。

Richard Zhang。使卷积网络再次具有平移不变性。载于*ICML*，2019年。

Weichen Zhang、Wanli Ouyang、Wen Li 和 Dong Xu。用于无监督域适应的协作式对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，2018年。

Yuchen Zhang、Tianle Liu、Mingsheng Long 和 Michael Jordan。连接领域自适应理论与算法。发表于国际机器学习会议。PMLR，2019年。

Han Zhao、Remi Tachet Des Combes、Kun Zhang 和 Geoffrey Gordon。关于领域自适应的不变表示学习。发表于国际机器学习会议，第 7523–7532 页。PMLR，2019年。

Yang Zou、Zhiding Yu、BVK Kumar 和 Jinsong Wang。通过类别均衡自训练进行语义分割的无监督领域自适应。载于欧洲计算机视觉会议论文集（*ECCV*），第289–305页，2018年。

附录

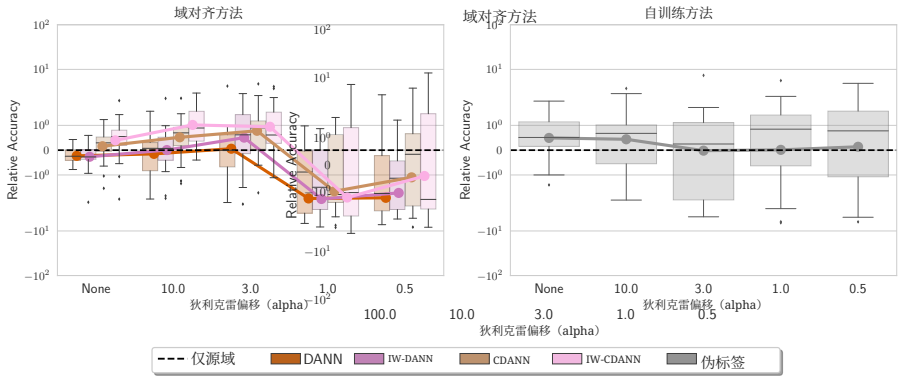
A. 图的说明

对于图 2中的每个图表，我们都会获取具有特定 alpha 值（即 x 轴上的值）的所有分布偏移对。然后，对于每个分布偏移对（具有特定的 alpha 值），我们通过将该分布偏移对的源数据集上训练的仅源模型的性能，从使用感兴趣的域适应算法在该分布偏移对上训练的模型性能中减去，得到相对性能 。因此，对于每个 alpha 值和每种域适应方法，我们会得到112个相对性能值，并绘制这些相对性能值的箱线图及其均值。

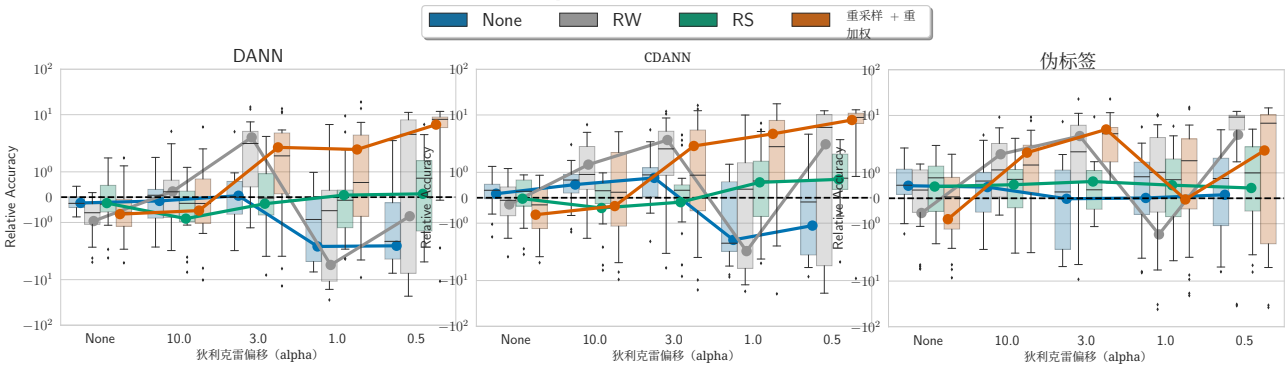
对于（外观相似的）图，我们在全文中使用相同的技术。唯一变化的是执行聚合所涉及的点的组。

B. 主论文中省略的表格数据与自然语言处理结果

B.1. 表格数据集



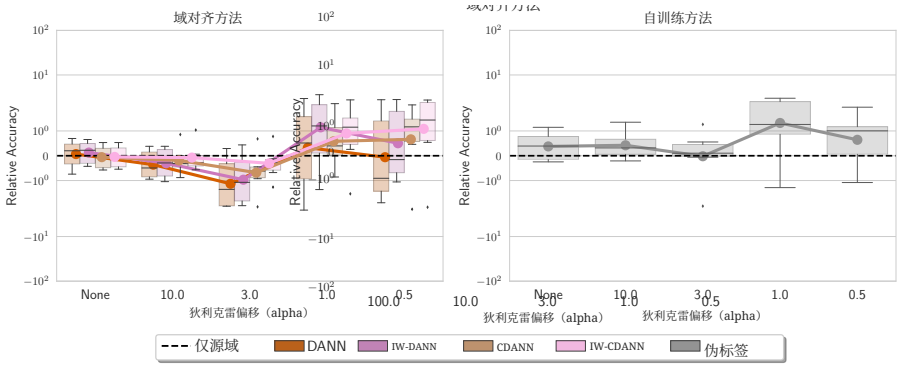
(a) 随着目标标签边缘分布偏移增大，领域自适应方法相对于仅源域训练的性能



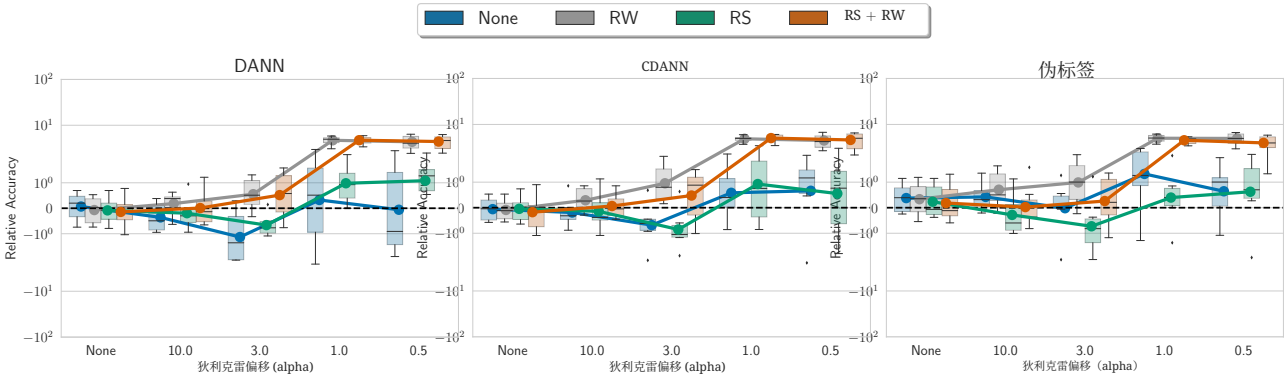
(b) 领域自适应方法与我们的元算法（RS 和 RW 校正）结合时的相对性能

图 5. 表格数据集中，不同领域自适应方法相对于仅源模型在所有分布偏移对上的性能，按标签边缘分布偏移程度分组。对于每个分布偏移对和领域自适应方法，我们绘制采用该领域自适应方法训练的模型的相对准确率，即从其准确率中减去仅源模型的准确率。因此，位于 0 处的黑色虚线表示仅源模型的性能。狄利克雷偏移参数越小，目标类别比例的偏移越严重。 **(a)** 当偏移参数为 $\alpha = \{$ NONE $10.0, 3.0\}$ 时，对不同领域自适应方法几乎没有影响；而当 $\alpha \in \{1.0, 0.5\}$ 时，所有领域自适应方法的性能都会下降，且通常低于仅源域分类器的性能。 **(b)** 在我们的元算法中，RS 和 RW 结合使用，相较于不进行校正，显著提升了所有领域自适应方法的总体性能。相较于不进行校正，RS 在不同标签边缘分布偏移程度下始终有帮助，而当偏移程度较小时，RW 会造成轻微损害。不过，在偏移严重时（$\alpha \in \{3.0, 1.0, 0.5\}$），RW 会显著提升所有方法的性能。各方法在单个数据集上的完整结果见附录 J。

B.2. 自然语言处理数据集



(a) 随着目标标签边缘分布偏移增大，领域自适应方法相对于仅源域训练的性能



(b) 将领域自适应方法与我们的元算法（RS 和 RW 校正）结合时的相对性能

图 *6.* 自然语言处理数据集中，不同领域自适应方法相对于仅源模型在所有分布偏移对上的性能，按标签边缘分布偏移程度分组。对于每个分布偏移对和领域自适应方法，我们绘制采用该领域自适应方法训练的模型的相对准确率，即从其准确率中减去仅源模型的准确率。因此，位于 0 处的黑色虚线表示仅源模型的性能。狄利克雷偏移参数越小，目标类别比例的偏移越严重。 **(a)** 随着目标标签边缘分布偏移程度加剧，DANN 和 IW-DANN 的性能会下降，且通常低于仅源域分类器的性能（ Noisy Student 除外）。PsuedoLabel、CDANN 和 IW-CDANN 对目标边缘偏移程度加剧的敏感性较低。 **(b)** 在我们的元算法中，RS 和 RW 结合使用，相较于不进行校正，显著提升了所有领域自适应方法的总体性能。相较于不进行校正，RS 在不同标签边缘分布偏移程度下始终有帮助，而当偏移程度较小时，RW 会稍微损害 BN-adapt、TENT 和 NoisyStudent 的性能。不过，在偏移严重时（$\alpha \in \{3.0, 1.0, 0.5\}$），RW 会显著提升所有方法的性能。各方法在单个数据集上的详细结果见附录 J。

C. IW-CDANN、IW-DANN 和 SENTRY 与结合我们元算法的现有领域自适应方法的比较

图 7展示了相关对比结果。

注意。 在 OfficeHome 数据集上，我们发现 SENTRY 在我们的实验结果与论文中最初报告的数据之间存在轻微差异（Prabhu 等人，2021年）。我们发现，这是由于原始工作与我们的实验所采用的批量大小不同（我们的实验对所有算法均采用相同的批量大小）。在附录N中，我们报告了采用更新后批量大小的 SENTRY 结果。采用新的批量大小后，SENTRY 的结果与原报告一致，但我们也观察到 FixMatch 的结果显著提升。有关更详细的讨论，请参阅附录N。

D. 数据集详情

本节将进一步介绍我们基准研究中使用的数据集。

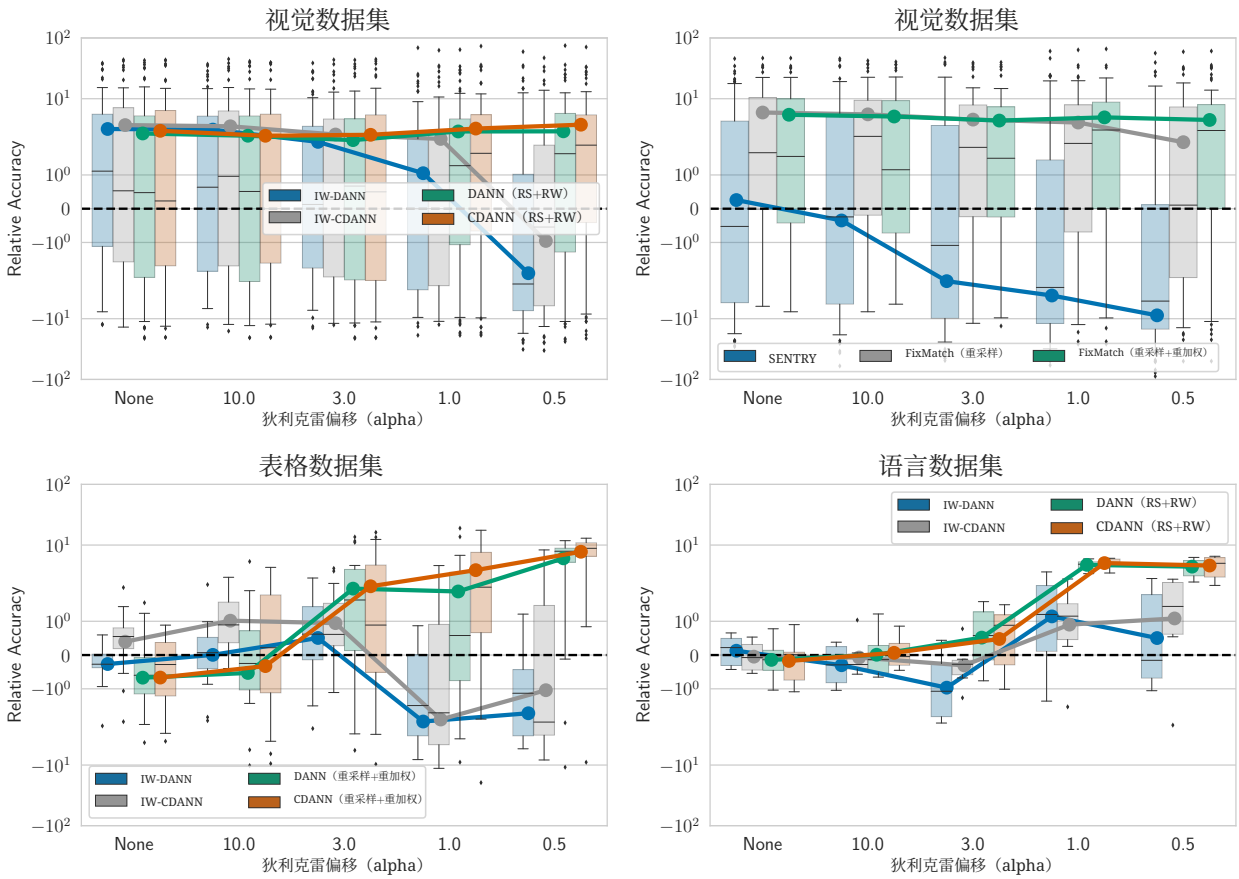


图 7。 将现有域适应方法与我们的重采样和重加权校正相结合，并与专门针对宽松标签偏移问题提出的域适应方法进行比较。 在视觉和表格数据集上，我们观察到，随着目标标签边缘分布偏移加剧，IW-DAN、IW-CDAN 和 SENTRY 容易受到影响。特别是在严重的目标标签边缘分布偏移下，IW-DAN、IW-CDAN 和 SENTRY 的性能通常低于仅源模型。然而，将现有域适应技术与重采样 + 重加权校正结合后，其性能相较于仅源模型有显著提升。对于自然语言处理数据集，我们观察到类似的行为，但程度相对较轻。

• **CIFAR10** 我们使用原始 CIFAR10 数据集（Krizhevsky 和 Hinton，2009年）作为源数据集。对于目标域，我们考虑：（i）由常见损坏导致的合成偏移（CIFAR10-C）（Hendrycks 和 Dietterich，2019年）；以及（ii）自然分布偏移，即由数据收集策略差异导致的 CIFAR10v2（Recht 等人，2018年；Torralba 等人，2008年）。我们从 CIFAR-10-C 数据集中随机抽取了 3 个数据集。总计，我们得到 5 个数据集（即 CIFAR10v1、CIFAR10v2、CIFAR10C-Frost（严重程度 4）、CIFAR10C-Pixelate（严重程度 5）、CIFAR10-C-饱和（严重程度 5））。

• **CIFAR100** 与 CIFAR10 类似，我们使用原始 CIFAR100 集合作为源数据集。对于目标域，我们考虑由常见损坏导致的合成偏移（CIFAR100-C）。我们采样了 4 个 CIFAR100-C 数据集，共得到 5 个领域（即 CIFAR100、CIFAR100C-Fog（严重程度 4）、CIFAR100C-Motion Blur（严重程度 2）、CIFAR100C-Contrast（严重程度 4）和 CIFAR100C-spatter（严重程度 2））。

• **FMoW** 为了考虑现实环境中面临的分布偏移，我们采用 WILDS 基准测试中的 FMoW-WILDs（Koh 等人，2021年；Christie 等人，2018年）。该数据集包含在不同地理区域和不同时间拍摄的卫星图像。我们使用原始训练集作为源域，并将分布外验证集和分布外测试集划分作为目标域，因为它们是在不同时间段收集的。总共得到 3 个不同的领域。

• **Camelyon17** 与 FMoW 类似，我们采用 WILDS 基准测试中的肿瘤识别数据集（Bandi 等人，2018年）。我们使用默认训练集作为源域，并将分布外验证集和分布外测试集划分作为目标域，因为它们是在不同医院收集的。总共得到 3 个不同的领域。

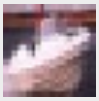
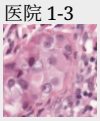
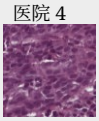
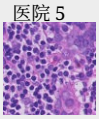
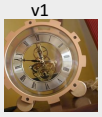
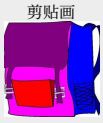
数据集	领域				
CIFAR10	CIFAR10v1 	CIFAR10v2 	CIFAR10C-Frost 	CIFAR10C-Pixelate 	Cifar10C-饱和度 
CIFAR100	Cifar100v1 	Cifar100C-Fog 	Cifar100C-运动模糊 	Cifar100C-对比度 	Cifar100C-Spatter 
Camelyon	医院 1-3 	医院 4 	医院 5 		
Entity13	v1 	v1（不相交子集） 	v2 	v2（分离子集） 	
实体30	v1 	v1（不相交子集） 	v2 	v2（不相交子集） 	
Living17	v1 	v1（不相交子集） 	v2 	v2（不相交子集） 	
Nonliving26	v1 	v1（不相交子集） 	v2 	v2（不相交子集） 	
FMoW	2002—'13 年 	2013—'16 年 	2016—'18 年 		
OfficeHome	产品 	真实世界 	剪贴画 	Art 	
DomainNet	Real 	剪贴画 	素描 	绘画 	
VisDA	渲染	真实 -1	真实 - 2		

图*8.* 每个视觉数据集中所有领域的示例。

数据集	源域	目标域
CIFAR10	CIFAR10v1	CIFAR10v1、CIFAR10v2、CIFAR10C-Frost（严重程度 4），CIFAR10C-Pixelate（严重程度 5）、CIFAR10-C-饱和（严重程度 5）
CIFAR100	CIFAR100	CIFAR100、CIFAR100C-Fog（严重程度 4），CIFAR100C-Motion Blur（严重程度 2）、CIFAR100C-Contrast（严重程度 4），CIFAR100C-spatter（严重程度 2）
Camelyon	Camelyon（医院1–3）	Camelyon（医院1–3）、Camelyon（医院 4）、Camelyon（医院 5）
FMoW	FMoW（2002–2013年）	FMoW（2002–2013年）、FMoW（2013–2016年）、FMoW（2016–2018年）
Entity13	Entity13（ImageNetv1 子群体1）	Entity13（ImageNetv1 子群体1），Entity13（ImageNetv1 子群体1），Entity13（ImageNetv2 子群体1），Entity13（ImageNetv2 子群体2）
实体30	实体30（ImageNetv1 子群体1）	实体30（ImageNetv1 子群体1），实体30（ImageNetv1 子群体2），实体30（ImageNetv2 子群体1），实体30（ImageNetv2 子群体2）
Living17	Living17（ImageNetv1 子群体1）	Living17（ImageNetv1 子群体1），Living17（ImageNetv1 子群体2），Living17（ImageNetv2 子群体1），Living17（ImageNetv2 子群体2）
Nonliving26	Nonliving26（ImageNetv1 子群体1）	Nonliving26（ImageNetv1 子群体1），Nonliving26（ImageNetv1 子群体2），Nonliving26（ImageNetv2 子群体1），Nonliving26（ImageNetv2 子群体2）
OfficeHome	产品	产品、艺术、剪贴画、真实
DomainNet	Real	真实、绘画、素描、剪贴画
VisDA	合成（原称训练集）	合成、Real-1（原称验证集），Real-2（原称测试集）
CivilComments	训练集	训练集、验证集和测试集（均由互不重叠的在线文章划分构成）
MIMIC 再入院	MIMIC 再入院（年份：2008年）	MIMIC 再入院（年份：2008年），MIMIC 再入院（年份：2009年），MIMIC 再入院（年份：2010年），MIMIC 再入院（年份：2011年），MIMIC 再入院（年份：2012年），MIMIC 再入院（年份：2013年）
退休成年人	退休成年人（年份：2014年州：['MD', 'NJ', '马萨诸塞州'])	退休成年人（年份：2015年；州：['马里兰州', '新泽西州', '马萨诸塞州']），退休成年人（年份：2016年；州：['马里兰州', '新泽西州', '马萨诸塞州']），退休成年人（年份：2017年；州：['马里兰州'、'新泽西州'、'马萨诸塞州']），退休成年人（年份：2018年；州：['马里兰州 ', 'NJ', '马萨诸塞州'])

表1. 我们在 RLSBENCH 中考虑的数据集详情。

• **BREEDs** 我们还在设置中纳入 BREEDs 基准（Santurkar 等人，2021年），以评估对亚群偏移的鲁棒性。BREEDs 利用 ImageNet 中的类别层级，将原始类别重新用作子群体，并在超类上定义分类任务。我们考虑由子群体偏移引起的分布偏移，即通过直接令训练分布与测试分布中的子群体互不重叠来实现。BREEDs 基准包含4个数据集**Entity-13**、**Entity-30**、**Living-17**和**Non-living-26**，每个数据集都聚焦于层级结构中不同的子树和层级。我们还考虑由 ImageNet 数据收集过程差异导致的自然偏移（Russakovsky 等人，2015年），例如 ImageNetv2（Recht 等人，2019年）以及两者的组合。总的来说，对于4个 BREEDs 数据集（即 Entity-13、Entity-30、Living-17 和 Non-living-26），我们分别获得四个不同的领域。具体如下：BREEDsv1 子群体1（从 ImageNetv1 抽样）、BREEDsv1 子群体2（从 ImageNetv1 抽样）、BREEDsv2 子群体1（从 ImageNetv2 抽样）和 BREEDsv2 子群体2（从 ImageNetv2 抽样）。对于每个 BREEDs 数据集，我们将 BREEDsv1 子群体 A 用作源域，将其他三个用作目标域。

• **OfficeHome** 我们使用 OfficeHome 数据集中的四个领域（艺术、剪贴画、产品和真实）Venkateswara 等人，2017年）。我们使用产品领域作为源域，其他领域作为目标域。

• **DomainNet** 我们使用 DomainNet 数据集中的四个领域（剪贴画、绘画、真实、素描）Peng 等人，2019年）。我们使用真实领域作为源域，其他领域作为目标域。

• **Visda** 我们使用 Visda 数据集（Peng 等人，2018年）中的三个领域（训练集、验证集和测试集）。其中，“训练集”领域包含对象的合成图像，而“验证集”和“测试集”领域包含真实世界的图像。为避免将领域名称与其作为数据划分的角色混淆，我们将它们分别重命名为“合成”“Real-1”和“Real-2”。我们使用合成领域（原训练集）作为源域，并将其他领域作为目标域。

• **CivilComments** （Borkan 等人，2019年）来自 WILDS 基准测试，其中包含三个领域：训练集、分布外验证集和分布外测试集，用于毒性检测；这些领域对应不同的人口统计学子群体。由于每个领域的数据均来自在线文章的不同分区，该数据集在不同人口统计学群体之间存在子群体偏移。

• **退休成年人** （Ding 等人，2021年）我们考虑其中的 ACSIncome 预测任务，不同领域代表不同的州和时间段；我们随机选择三个州，并考虑这些州之间因时间变化而产生的数据集偏移。有关具体时间段和州的详细信息见表 1。

• **Mimic 再入院**(Johnson 等人，2020年；PhysioBank，2000年），其中任务是预测再入院风险，各个领域代表来自不同时间段的数据。具体时间段详见表 1。

我们提供了脚本，只需一条命令即可在代码中设置这些数据集。为了研究不同方法在更严格的标签偏移设定下的表现，我们还将源域的一个留出分区纳入目标域集合。对于这些源域和目标域均为独立同分布划分的分布偏移对，我们得到更严格的标签偏移问题。我们在表 1中总结了源域和目标域的信息。

训练-测试划分 我们将每个源域和目标数据集划分为 80% 和 20% 的独立同分布子集。使用 80% 的子集进行训练，使用 20% 的子集进行评估（或验证）。我们丢弃 80% 目标子集的标签，仅使用 20% 目标子集的标签进行最终评估。划分目标数据的理由是，在评估时使用一批完全未见过的数据。这样可以避免在模型可能已经过拟合的样例上进行评估。过拟合于未标记数据

用于评估的样例。实际上，如果目的是对所有目标域数据进行预测（即传导），我们可以直接使用（完整的）目标域数据集进行训练和评估。

E. 我们提出的元=算法示意图

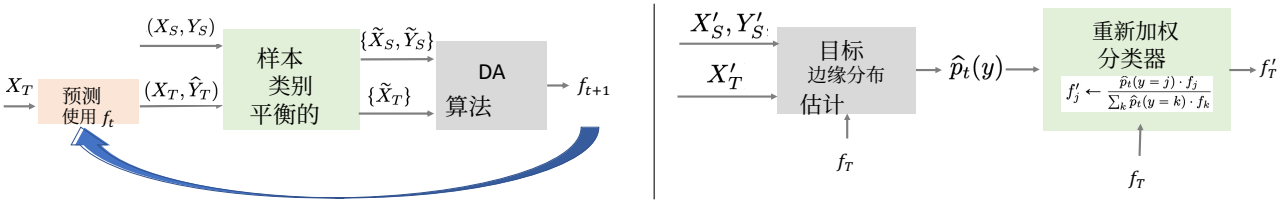


图 *9*。（左）RS 方法在每次迭代中的示意图。（右）使用 RW 方法对分类器进行事后重新加权的示意图。

F. 更严格的标签偏移假设下估计目标边缘分布的方法

本节介绍在更严格的标签偏移假设下估计目标标签边缘分布的方法。回顾一下，在标签偏移假设下，$p_s(y)$ 可以与 $p_t(y)$ 不同，但类别条件分布保持不变，即 $p_t(x|y) = p_s(x|y)$。我们重点讨论近期提出的、利用现成分类器在温和假设下获得一致估计的方法（Lipton 等人，2018年；Azizzadenesheli 等人，2019年；Alexandari 等人，2021年；Garg 等，2020年）。为简单起见，我们假设拥有有标签源数据 $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ 以及无标签目标数据 $\{x_{n+1}, x_{n+2}, \ldots, x_{n+m}\}$。

RLLS 首先，我们讨论 标签偏移下的正则化学习 （RLLS）（Azizzadenesheli 等人，2019年）（黑箱偏移估计 （BBSE，Lipton 等人 （2018年））的一种变体）：利用（可能存在偏差、

未校准或不准确的）预测来估计偏移。RLLS 通过求解以下优化问题来估计重要性权重 $w_t(y) = \frac{p_t(y)}{p_s(y)}$：

$$\widehat{w}_t^{\text{RLLS}} = \arg\min_{w \in \mathcal{W}} \left\| \widehat{C}_f w - \widehat{\mu}_f \right\|_2 + \lambda_{\text{RLLS}} \left\| w - 1 \right\|_2 . \tag{1}$$

其中 $\mathcal{W} = \{w \in \mathbb{R}^d | \sum_y w(y) p_s(y) = 1 \text{ 和 } \forall y \in \mathcal{Y} \quad w(y) > 0\}$。$\widehat{C}_f$ 是分类器 f 在源数据上的经验混淆矩阵，$\widetilde{\mu}_f$ 是分类器 f 对无标签目标数据的预测的经验均值。给定有标签源数据，可按如下方式计算经验混淆矩阵：

$$[\widehat{C}_f]_{i,j} = \frac{1}{n} \sum_{k=1}^{n} f_i(x_k) \cdot \mathbb{I}[y_k = j] .$$

要估计目标标签边缘分布，我们可以将估计得到的重要性权重与源标签边缘分布相乘（源标签边缘分布可直接根据有标签源数据估计）。

在宽松标签偏移问题中，我们使用验证源数据计算混淆矩阵，并使用一部分目标无标签数据计算 μ_f。除非另有说明，本文始终使用 RLLS 估计目标标签边缘分布。我们按照原论文（Azizzadenesheli 等人，2019年）的建议选择 λ_{RLLS}。

MLLS 接下来，我们讨论最大似然标签偏移（MLLS）（Saerens 等人，2002年；Alexandari 等人，2021年）：一种期望最大化（EM）算法，它通过最大化观测到的无标签目标数据的似然来估计目标标签边缘分布，并假设可以访问一个输出源域校准概率的分类器。具体而言，MLLS 使用以下目标函数：

$$\widehat{w}_t^{\text{MLLS}} = \arg\min_{w \in \mathcal{W}} \frac{1}{m} \sum_{i=1} \log(w^T f(x_{i+n})) , \tag{2}$$

其中，f 是在源域上训练的分类器，$\mathcal{W}$ 是上文定义的同一受约束集合。我们再次只需将估计的重要性权重乘以源标签边缘分布，即可估计目标标签边缘分布。

基线估计器 给定分类器 f，我们只需对该分类器在无标签目标数据上的输出取平均，即可估计目标标签边缘分布，也就是：

$$\widehat{p}_t^{\text{baseline}} = \frac{1}{m} \sum_{i=1} f(x_{i+n}) . \tag{3}$$

请注意，之前讨论的所有方法都利用了现成的分类器 f。因此，我们尝试使用通过各种深度领域自适应启发式方法获得的分类器来估计目标标签边缘分布。

得到目标标签边缘分布的估计值后，我们只需使用 $\widehat{p}_t$ 对分类器重新加权，使其 $f'_j = \frac{\widehat{p}_t(y=j) \cdot f_j}{\sum_k \widehat{p}_t(y=k) \cdot f_k}$ 适用于所有 $j \in \mathcal{Y}$。请注意，如果我们在源类别平衡不均匀的情况下训练 f（且未像算法 1 的步骤 1 中那样重新平衡），则可以使用重要性权重 $\widehat{w}_t$ 对分类器重新加权，使其 $f'_j = \frac{\widehat{w}_t(y=j) \cdot f_j}{\sum_k \widehat{w}_t(y=k) \cdot f_k}$ 适用于所有 $j \in \mathcal{Y}$。

G. 宽松标签偏移的理论定义

领域自适应问题通常适定性不良（Ben-David 等人，2010b）。研究者已尝试探究能够使问题适定的附加假设。其中一个例子是标签偏移设定：在该设定下，$p(x|y)$ 保持不变，而 $p(y)$ 可以改变。在标签偏移下，会面临两个挑战：（i）估计目标标签边缘分布$p_t(y)$；以及（ii）训练分类器 f，使其在目标域上的性能最大化。然而，这些假设在实践中通常会在一定程度上被违反。本文旨在放宽这一假设，并聚焦于宽松标签偏移设定。具体而言，我们假设标签分布可以从源域到目标域任意变化，但 $p(x|y)$在源域与目标域之间的变化方式受到相对严格的限制（例如，从 ImageNet（Russakovsky 等人，2015年）到 ImageNetV2（Recht 等人，2019年）这类现实世界中自然发生的偏移）。

从数学角度来说，我们假设 $p(x|y)$受到基于散度的约束，即对于某个较小的 $\epsilon > 0$ 和分布距离$\mathcal{D}$，有$\max_y \mathcal{D}(p_t(x|y), p_t(x|y)) \leqslant \epsilon$ 但允许标签边缘分布发生任意偏移 $p(y)$。先前的研究以不同方式定义了这些约束（Wu 等人，2019年；Tachet des Combes 等人，2020年；Kumar 等人，2020年）。

具体而言，我们可以使用无穷阶瓦瑟斯坦距离来定义约束。首先，我们针对 p, q 上的概率测度 $\mathcal{X}$ 定义瓦瑟斯坦距离：

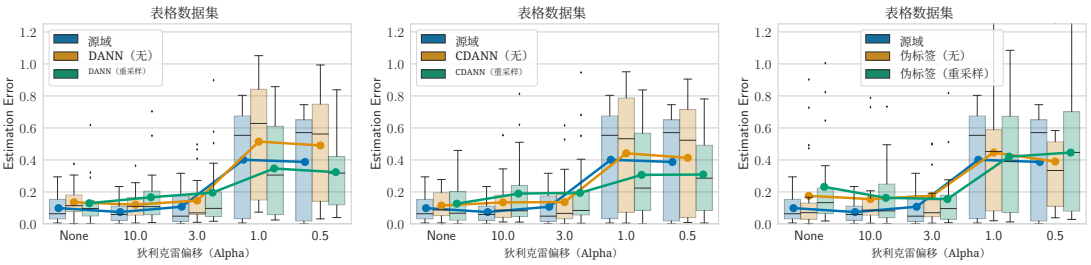
$$W_\infty(p,q) = \inf\{\sup_{x\in\mathbb{R}^d} \|f(x)-x\|_2 : f:\mathbb{R}^d \to \mathbb{R}^d, f_{\#}p = q\},$$

其中 # 表示测度的推前，即对于每个集合 $S \subseteq \mathbb{R}^d$, $p(S) = p(f^{-1}(S))$。直观而言， W_∞ 会将分布 p 中的点移动到 q，移动距离至多为 ϵ，从而使这些分布相匹配。因此，我们的 $D := \max_y W_\infty(p_s(x|y),\ p_t(x|y)) \leqslant \epsilon$。类似地，我们可以用 KL 或全变差距离来定义分布约束。我们可以在表示空间 $\mathcal{Z}$ 中定义约束：通过函数 h： $\mathcal{X} \to \mathcal{Z}$ 将输入 $x \in \mathcal{X}$ 投影得到该表示空间。直观而言，我们希望用某种 h 来定义分布距离，它能涵盖预测目标标签所需的全部信息，同时又能在投影空间中满足较小的分布散度。然而在实践中，使用有限样本很难实证验证这些分布距离是否足够小 ϵ。此外，对于这些偏移在常见领域自适应基准中以何种意义出现，我们缺乏严谨的刻画；而且，由于我们的工作侧重于对真实世界数据集进行实证评估，因此将形式化研究留待未来开展。

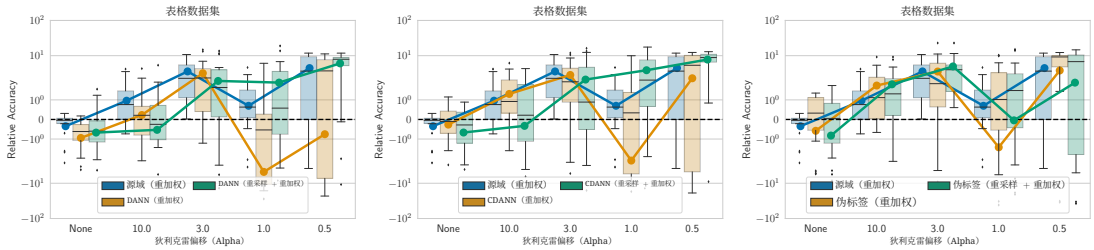
H. 目标边缘分布估计及其对准确率的影响

H.1. 视觉数据集

H.2. 表格数据集



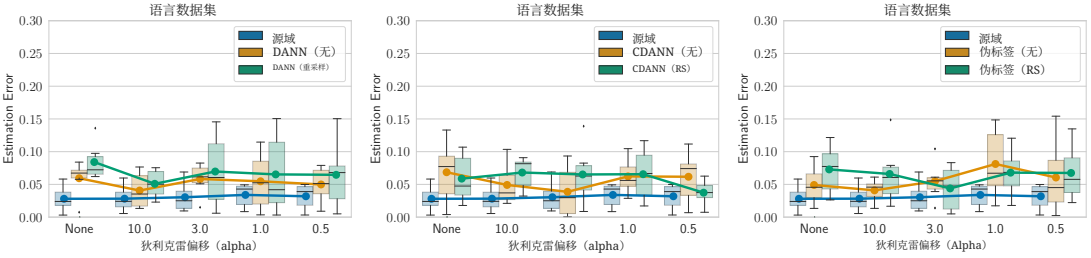
(a) 使用 RLLS 和通过不同领域自适应方法获得的分类器进行目标标签边缘分布估计的误差（ℓ_1）



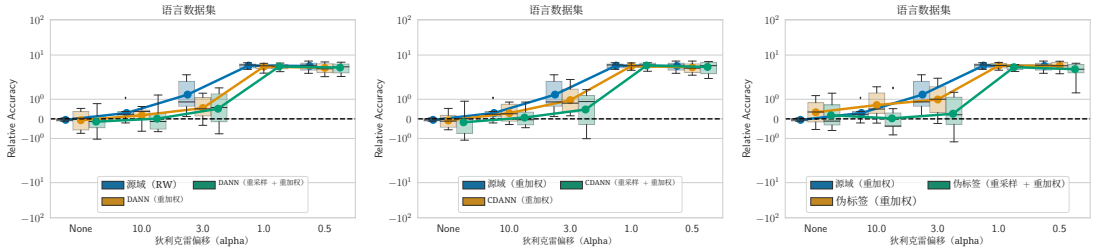
(b) 与 RW 校正结合时领域自适应方法的相对性能

图 *11.* 目标标签边缘分布估计（ℓ_1）误差以及使用 *RLLS* 和通过不同领域自适应方法获得的分类器时的相对性能。 对于表格数据集，在目标标签边缘分布严重偏移时，使用领域自适应方法获得的分类器进行 RLLS，优于使用仅源域分类器进行 RLLS。相应地，在目标标签边缘分布严重偏移时，与将事后 RW 校正应用于仅源域模型相比，将其应用于通过领域自适应方法训练的分类器可获得更好的性能。

H.3. 自然语言处理数据集

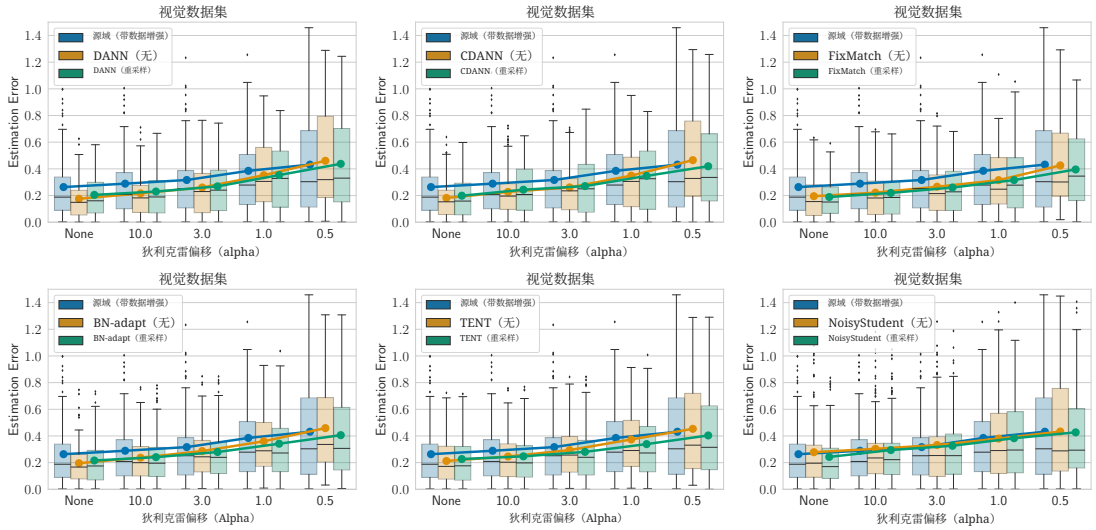


(a) 使用 RLLS 和通过不同域适应方法获得的分类器进行目标标签边缘分布估计（ℓ_1）的误差

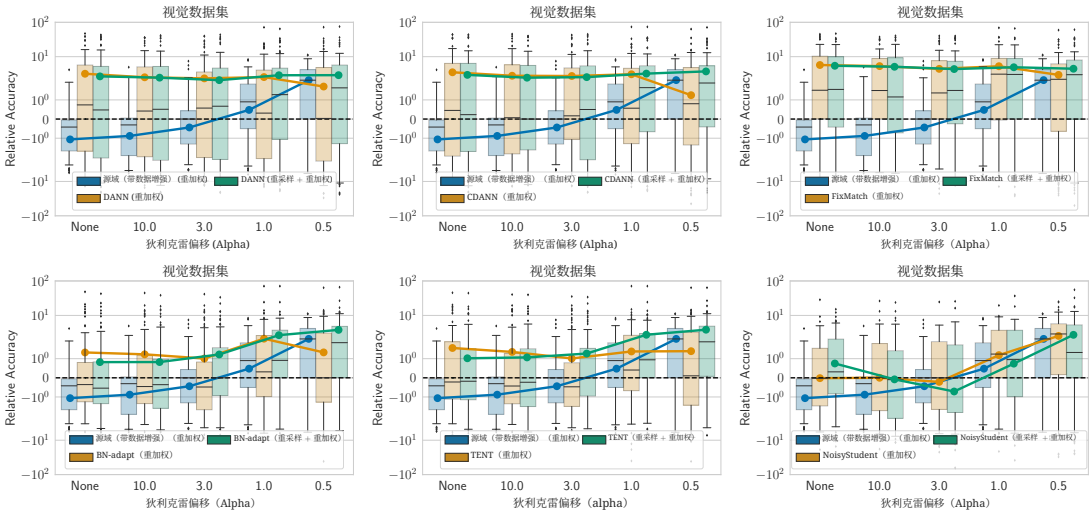


(b) 与 RW 校正配合使用时，域适应方法的相对性能

图 *12.* 使用 *RLLS* 和通过不同域适应方法获得的分类器进行目标标签边缘分布估计（ℓ_1）的误差及相对性能。 对于自然语言处理数据集，使用仅源域分类器的 RLLS 表现优于使用通过域适应方法获得的分类器的 RLLS。相应地，与通过域适应方法训练的分类器相比，对仅源域模型应用事后 RW 校正可提升性能。



(a) 使用 RLLS 进行目标标签边缘分布估计（ℓ_1）的误差，以及采用不同领域自适应方法获得的分类器



(b) 领域自适应方法结合 RW 校正时的相对性能

图 *10.* 目标标签边缘分布估计（ℓ_1）的误差，以及使用 *RLLS* 和采用不同领域自适应方法获得的分类器时的相对性能。在视觉数据集的所有偏移程度下（$\alpha = 0.5$ 除外），采用领域自适应方法获得的分类器与 RLLS 结合使用时，性能优于 RLLS 与仅源域分类器结合使用时的表现。相应地，与将事后 RW 校正应用于仅源模型相比，将其应用于通过领域自适应方法训练的分类器时，性能有显著提升。

H.4. 不同目标标签边缘分布估计方法的比较

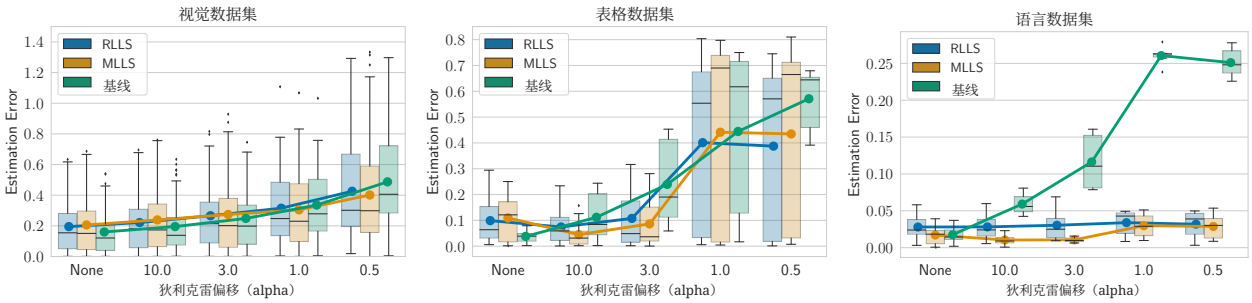


图 *13.* 不同目标标签边缘分布估计方法的比较。 我们使用仅源域分类器，绘制了不同方法的估计误差。对于所有模态，我们观察到，随着目标边缘分布偏移程度加剧，基线方法与 RLLS（或 MLLS）方法的估计误差之间存在权衡。

I. 采用预言机早停准则的结果

本节报告采用预言机早停准则时的结果。在视觉和表格数据集上，使用目标性能与源域留出集性能进行模型选择时，我们观察到性能存在差异。这表明，与准确率在线现象（Miller 等人，2021年；Recht 等人，2019年）相比，实际情况更为微妙。我们希望在未来的工作中更深入地研究这种对比行为。

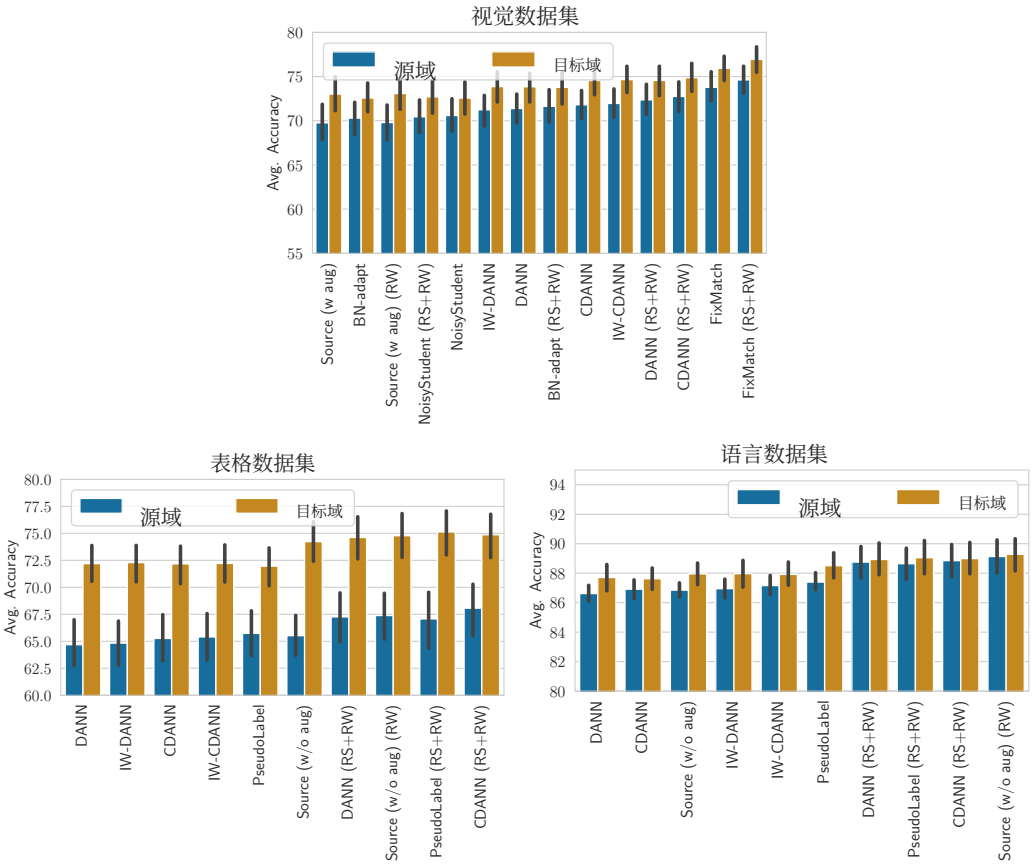


图 *14.*各模态中不同领域自适应方法在所有分布对上的平均准确率。 我们比较了依据源域验证性能和目标域验证性能确定的提前停止点所对应的性能。

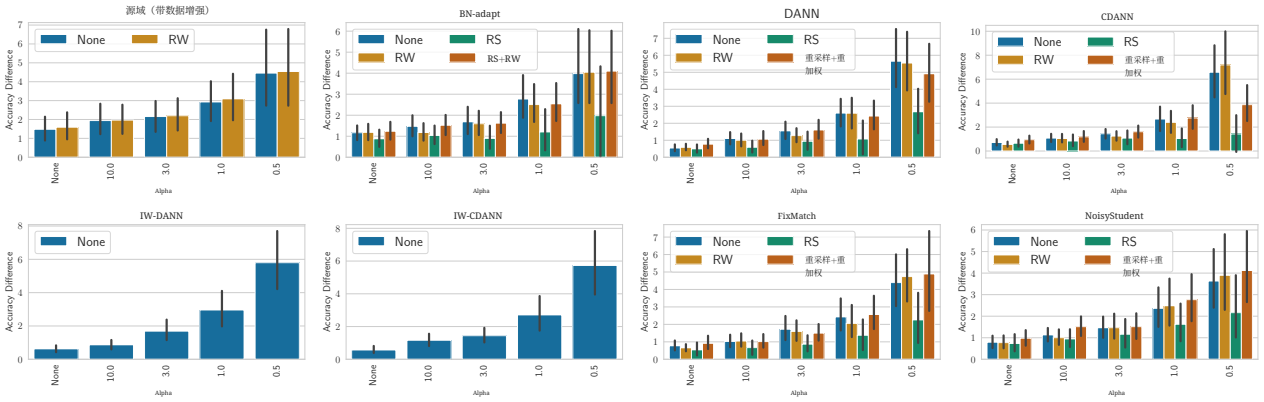


图 *15.* 在视觉数据集的所有分布偏移对上汇总，不同域适应方法以源域性能和目标性能作为早停准则时的准确率差异。我们观察到，随着偏移严重程度增加（即 α 减小），所有方法的准确率差异都会增加。

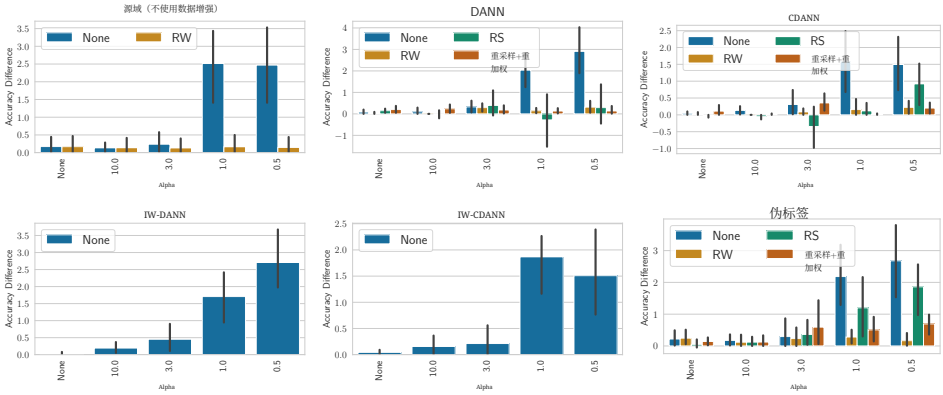


图 *16.* 在语言数据集的所有分布偏移对上汇总，不同域适应方法以源域性能和目标性能作为早停准则时的准确率差异。 我们观察到，随着偏移严重程度增加（即 α 减小），所有未进行任何校正的方法的准确率差异都会增加。采用 RS 和 RW 校正后，我们观察到，随着偏移严重程度增加，准确率差异基本保持不变。

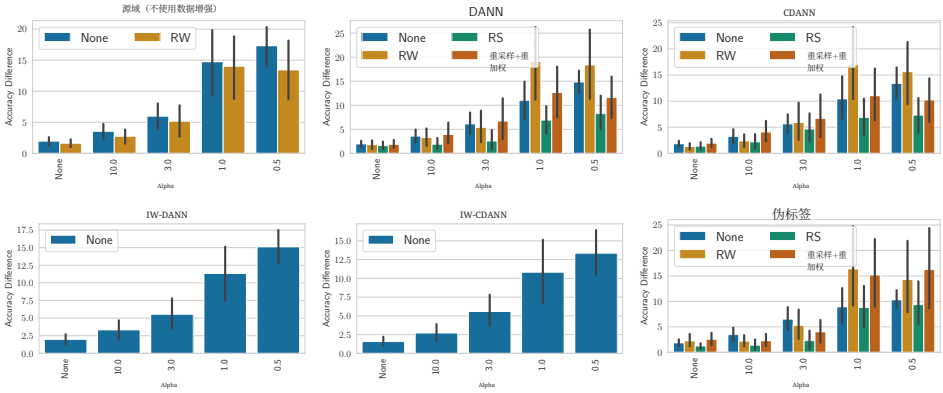


图 *17.* 在表格数据集的所有分布偏移对上汇总，不同域适应方法以源域性能和目标性能作为早停准则时的准确率差异。我们观察到，随着偏移严重程度增加（即 α 减小），所有方法的准确率差异都会增加。

J. 单个数据集上的结果

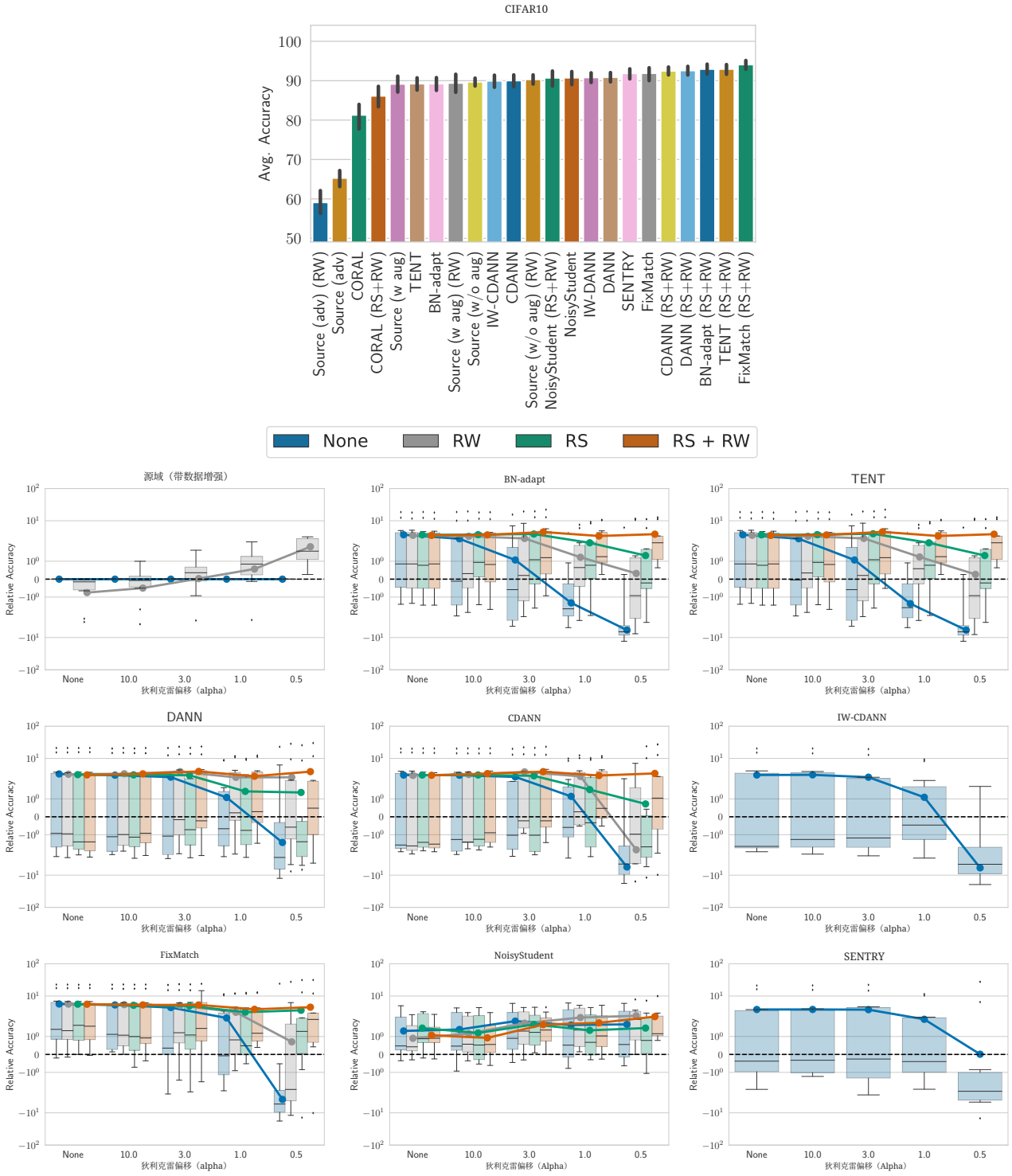


图 *18*。CIFAR10。CIFAR10 中不同域适应算法在各种偏移对上的相对性能和准确率图。

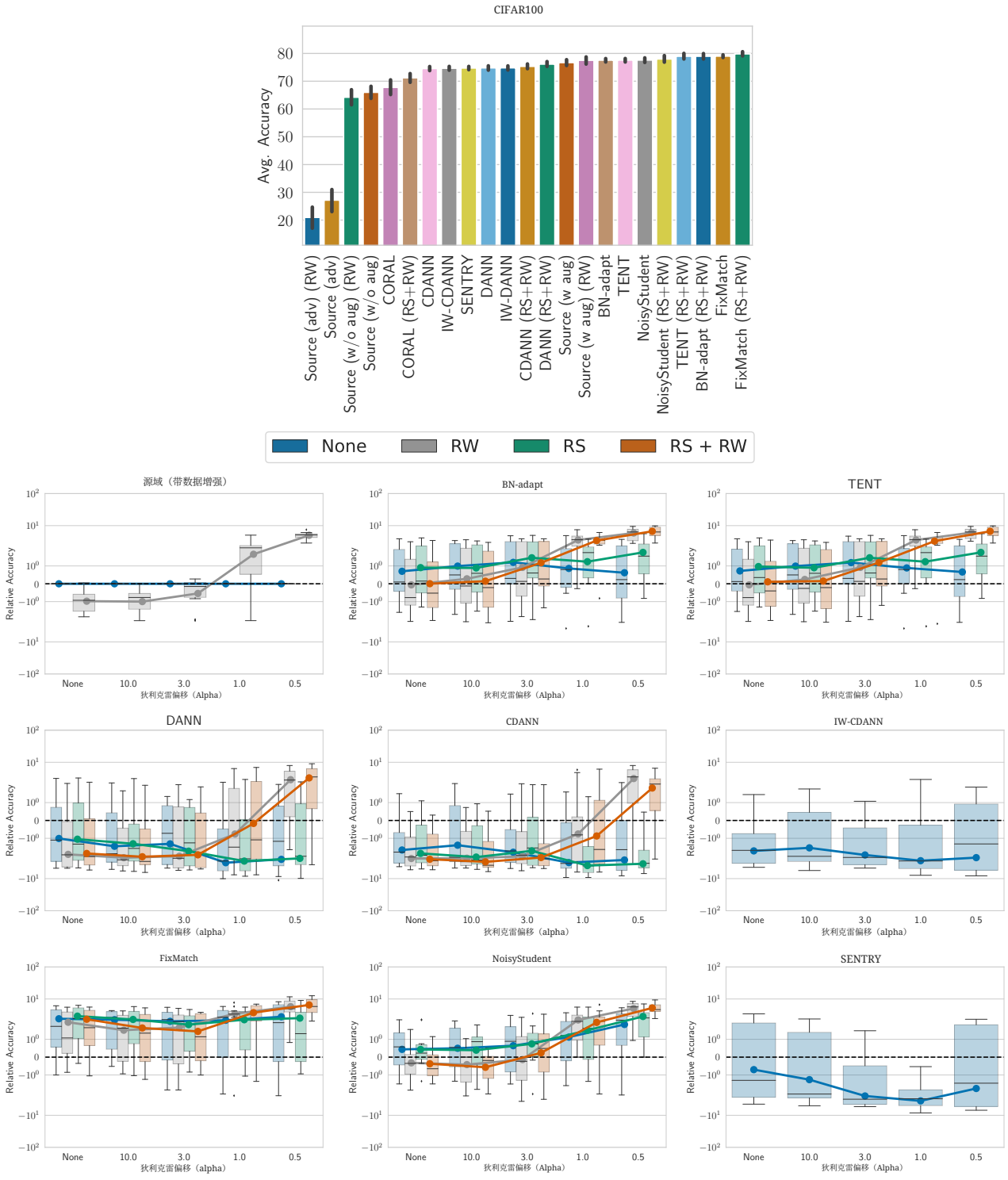


图 *19*。CIFAR100。相 对于 CIFAR100 中不同偏移对的各种领域自适应算法的相对性能和准确率图 .

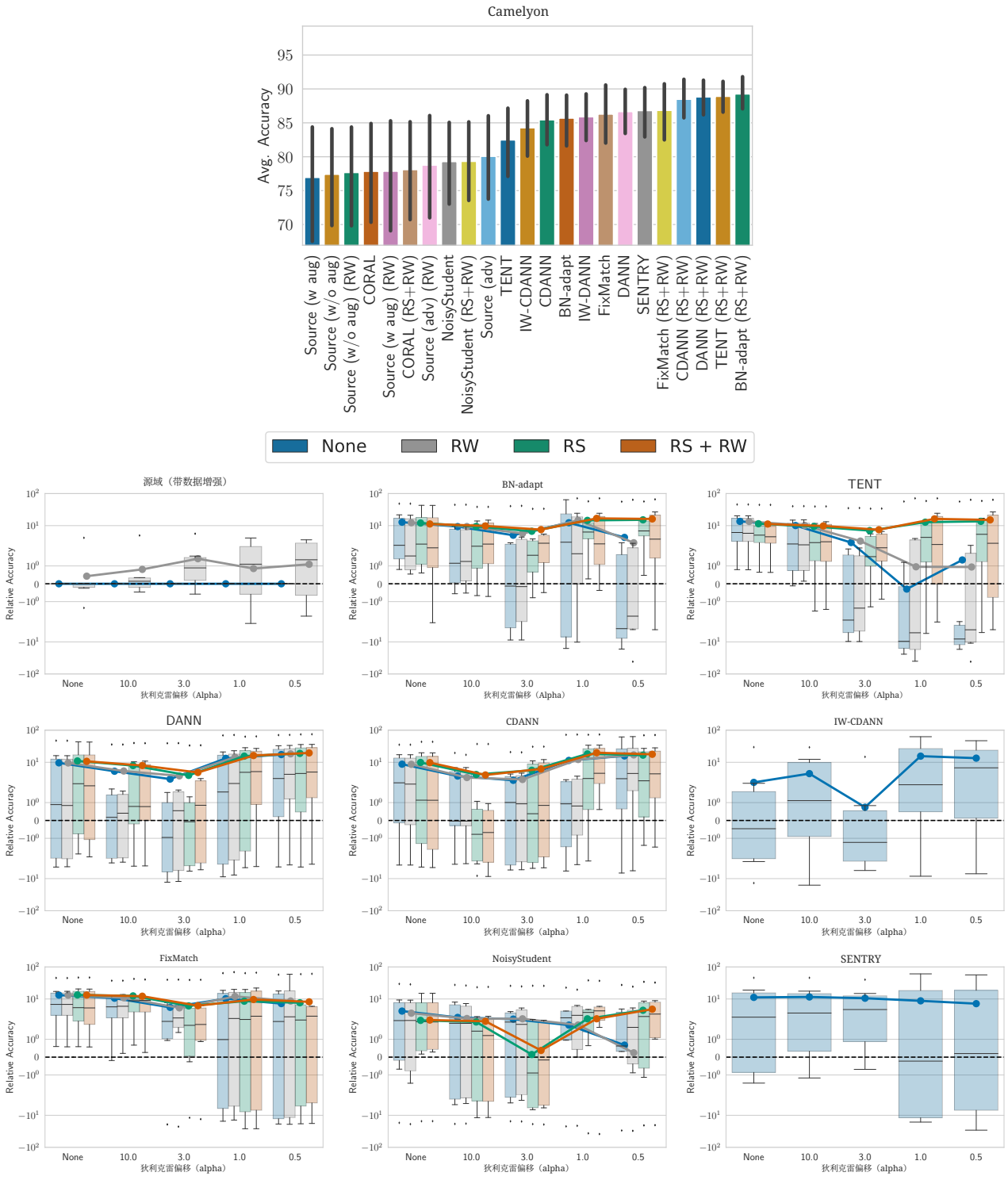


图 *20*。Camelyon。相 对于 Camelyon 中不同偏移对的各种 DA 算法的相对性能和准确率图。

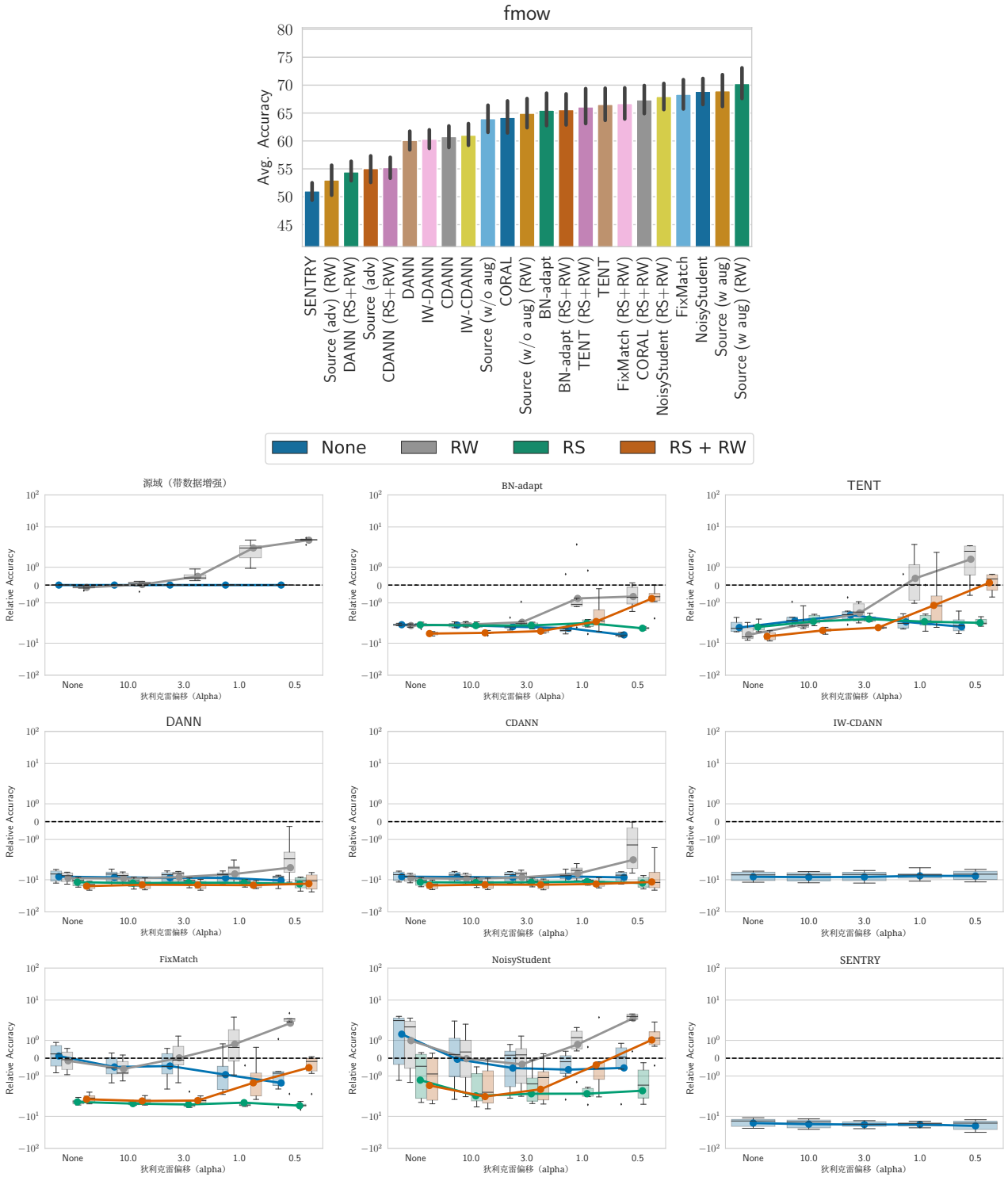


图 *21*。FMoW。 FMoW 中不同域适应算法在各种偏移对上的相对性能和准确率图。

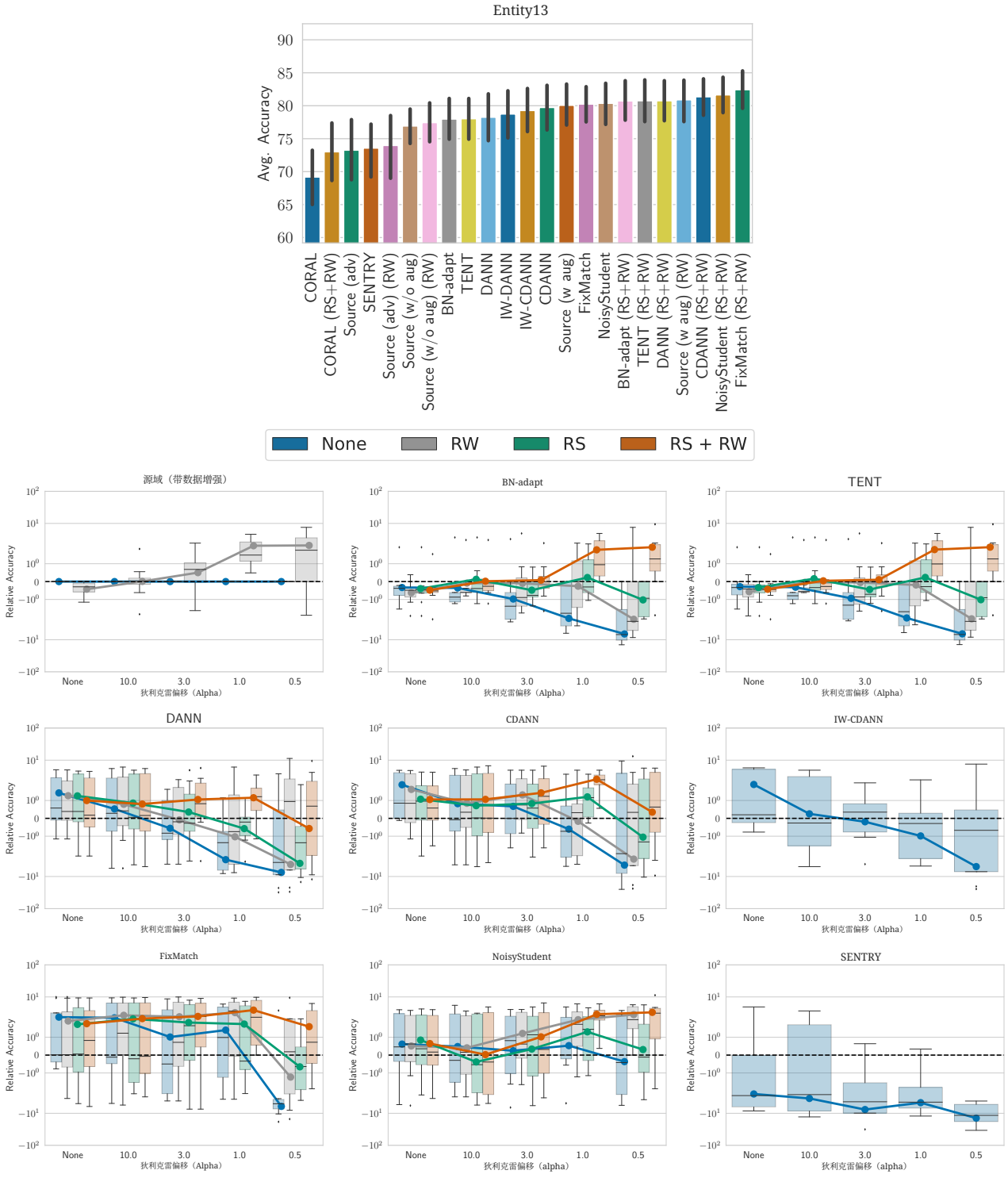


图*22*。Entity13。R 展示 Entity13 中不同域适应算法在各类偏移对下的相对性能和准确率图 .

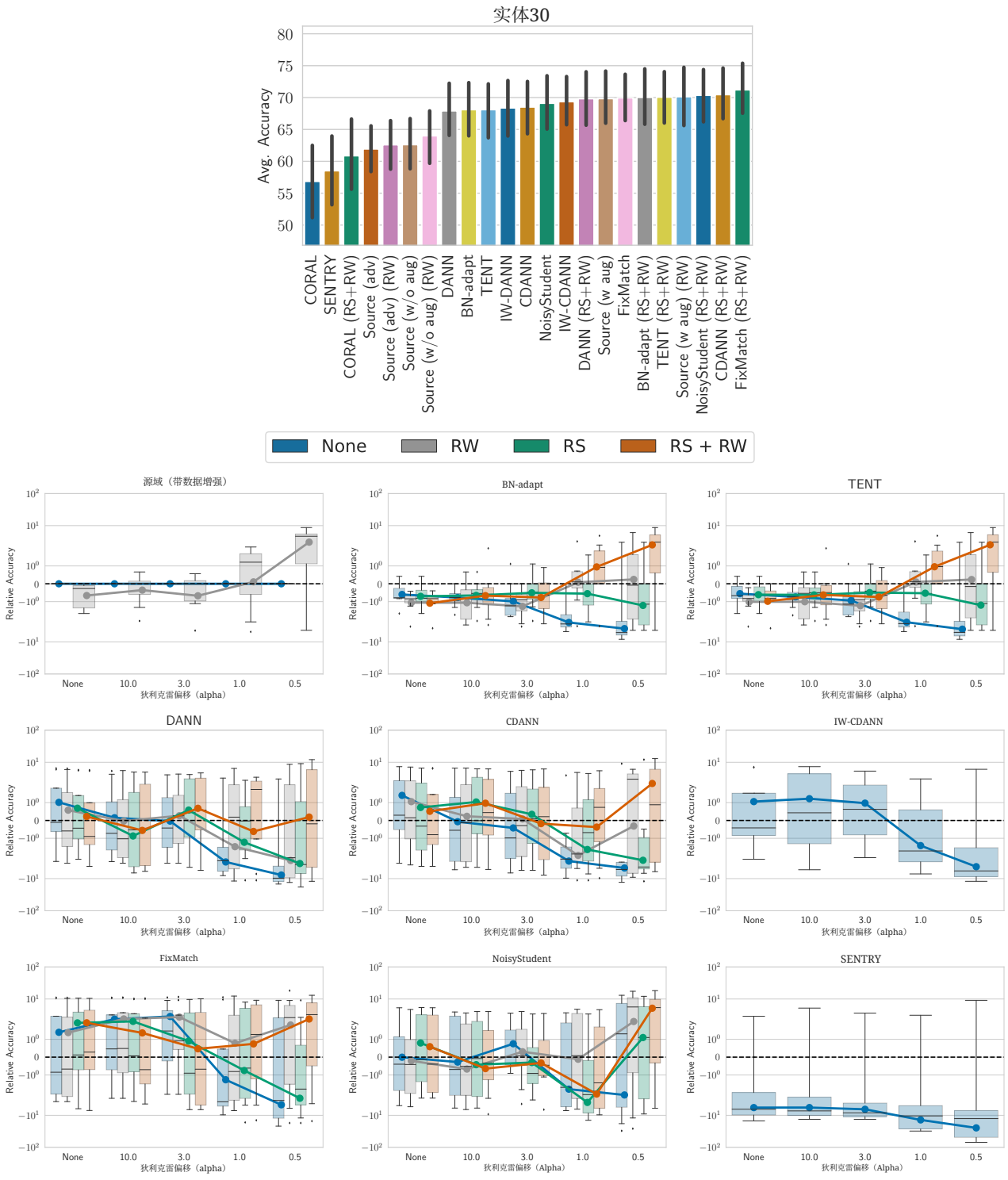


图*23*。实体30。R　实体30中不同领域自适应算法在各种偏移对上的相对性能和准确率图　　.

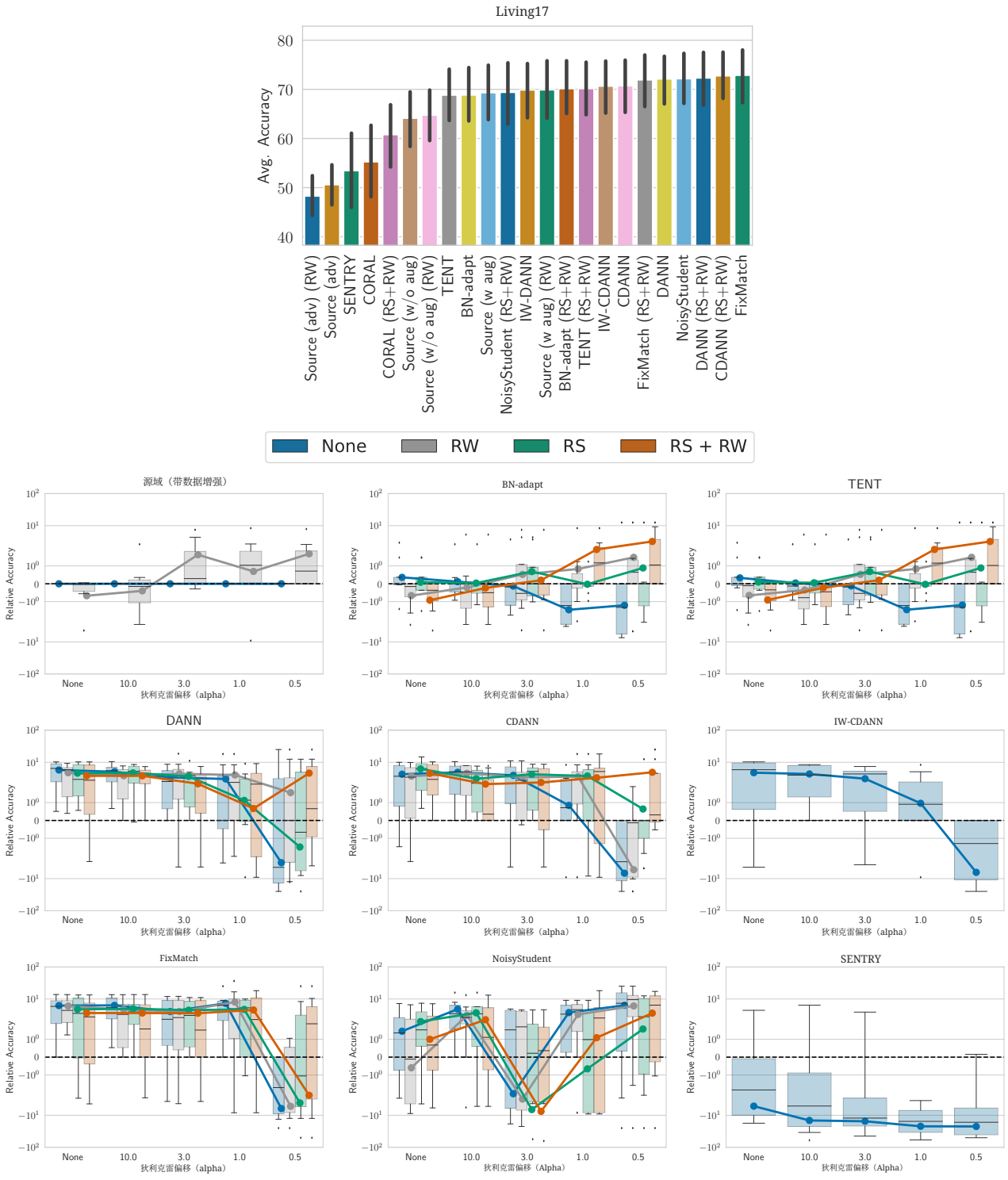


图*24.*Living 17. Living1 中不同领域自适应算法在各种偏移对上的相对性能和准确率图 7.

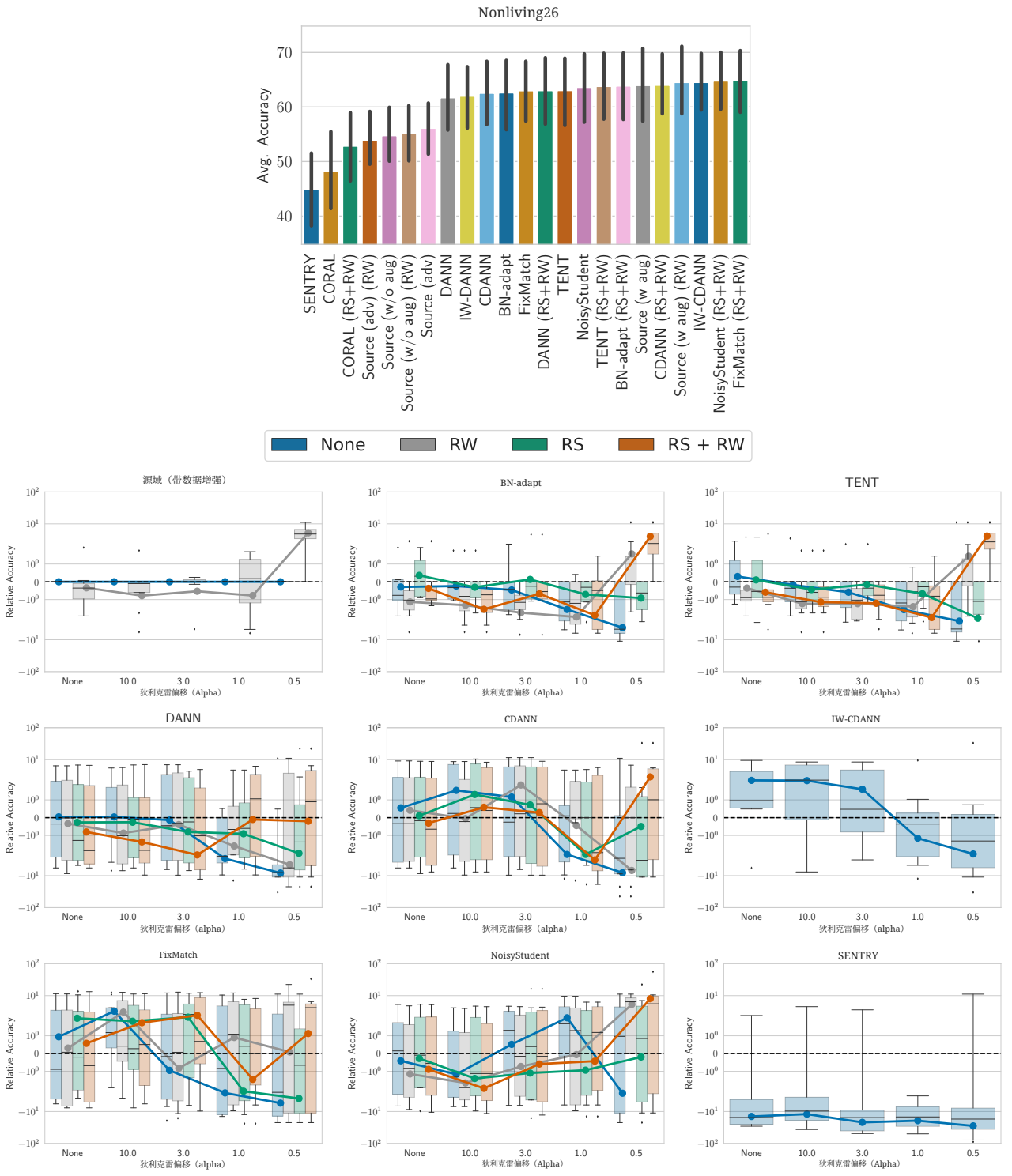


图 *25*。非生物 26。Rel　Nonliving2 中不同域适应算法在各偏移对上的相对性能和准确率图　6.

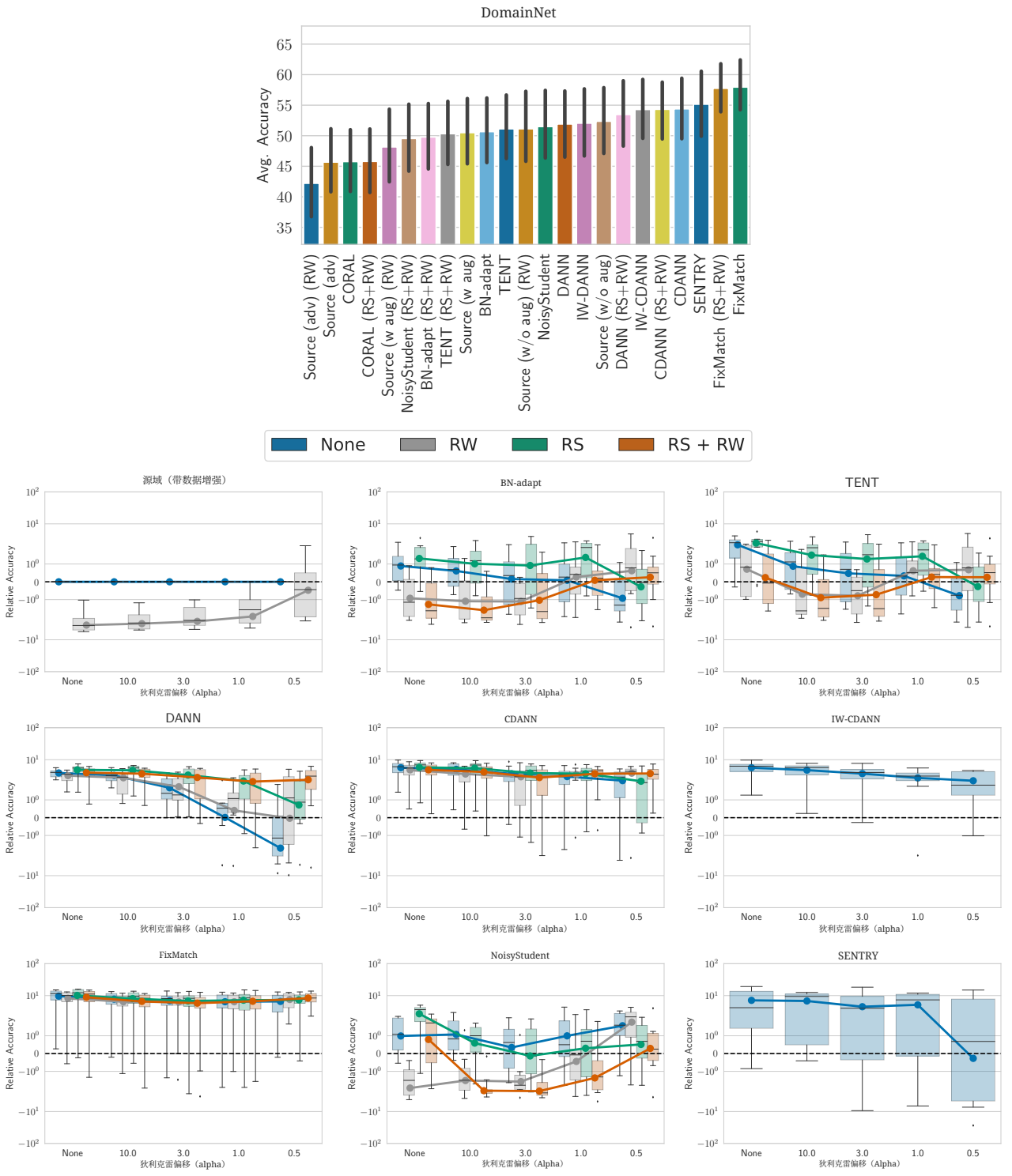


图 *26*。DomainNet。R　针对 DomainN 中不同偏移对的各类域适应算法的相对性能和准确率图　et.

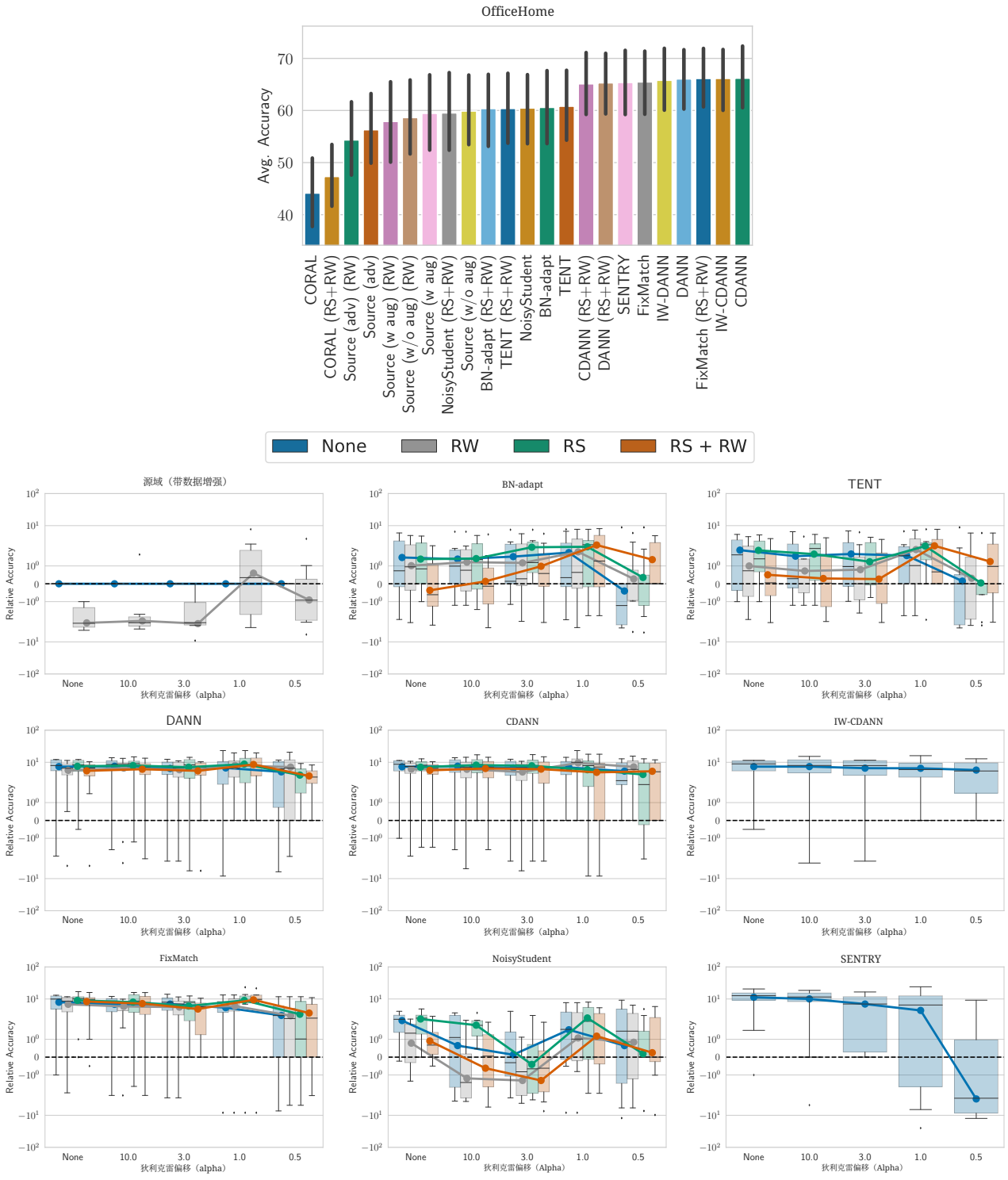


图 *27*。OfficeHome。相对 OfficeHome 中不同领域自适应算法在各个偏移对上的相对性能和准确率图。

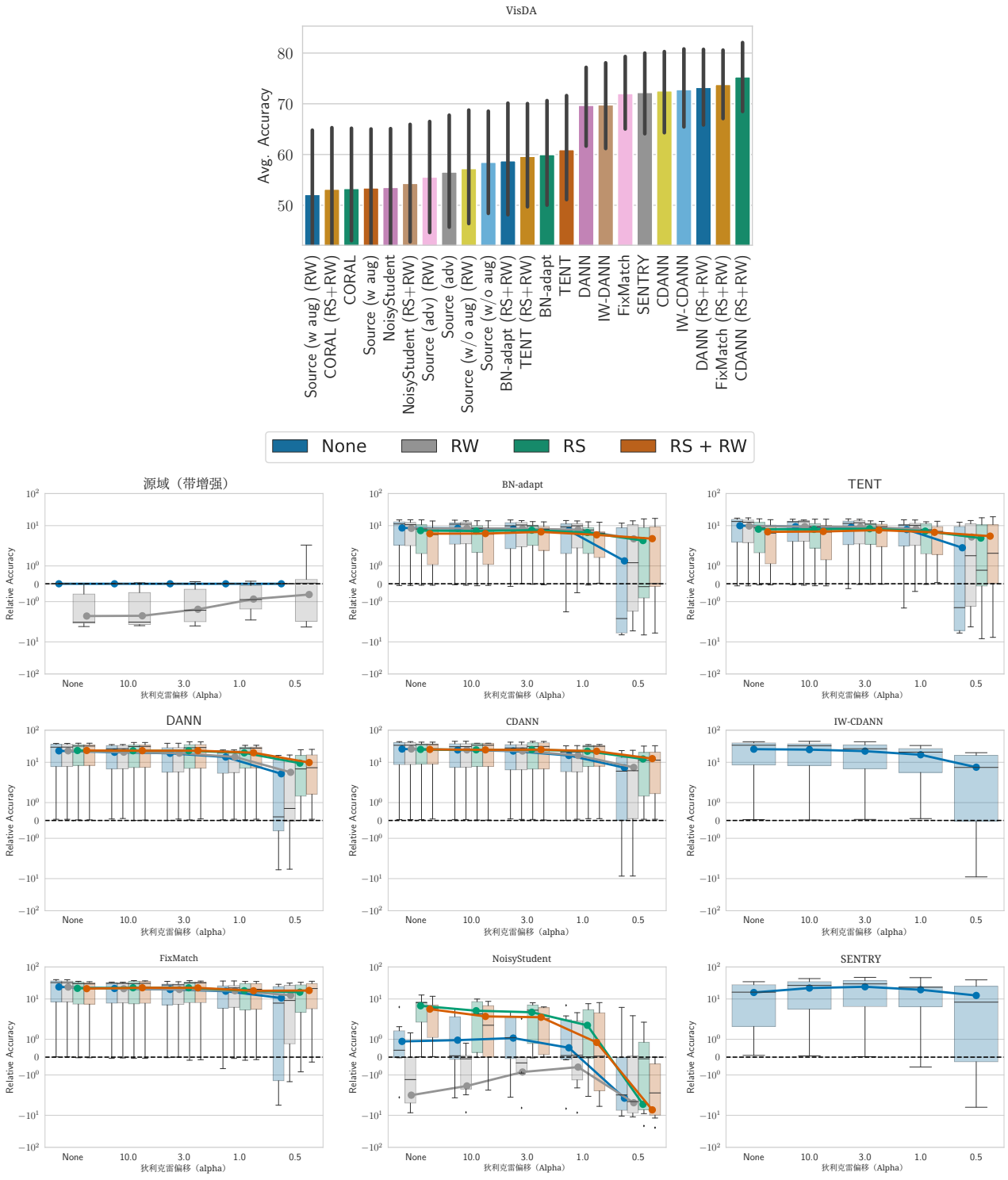


图*28.* Visd　　a. Visd 中不同域适应算法在各种偏移对上的相对性能和准确率图　　a.

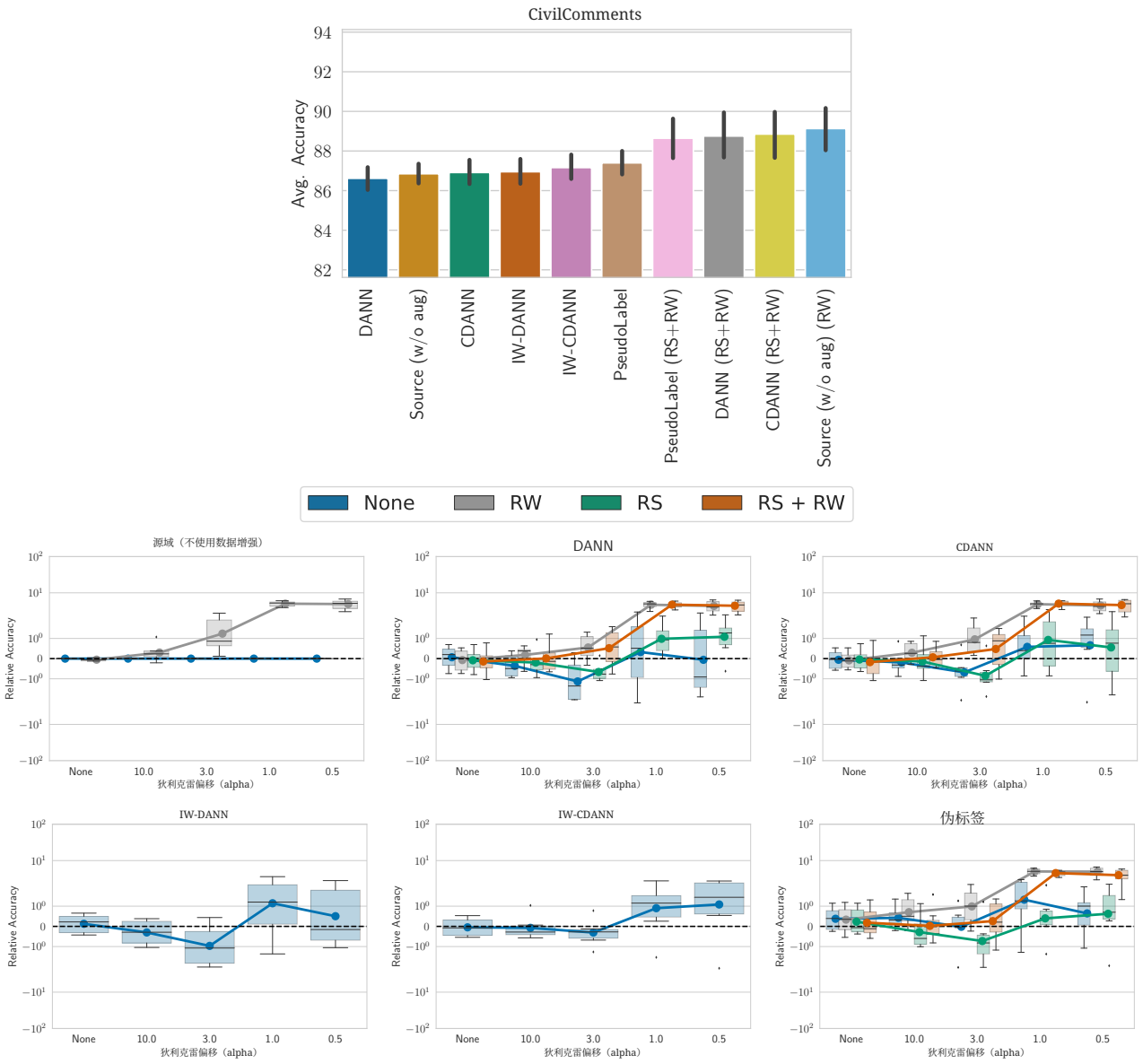


图 *29*。CivilComments。在 CivilComments 的不同偏移对下，各种领域自适应算法的相对性能和准确率图。

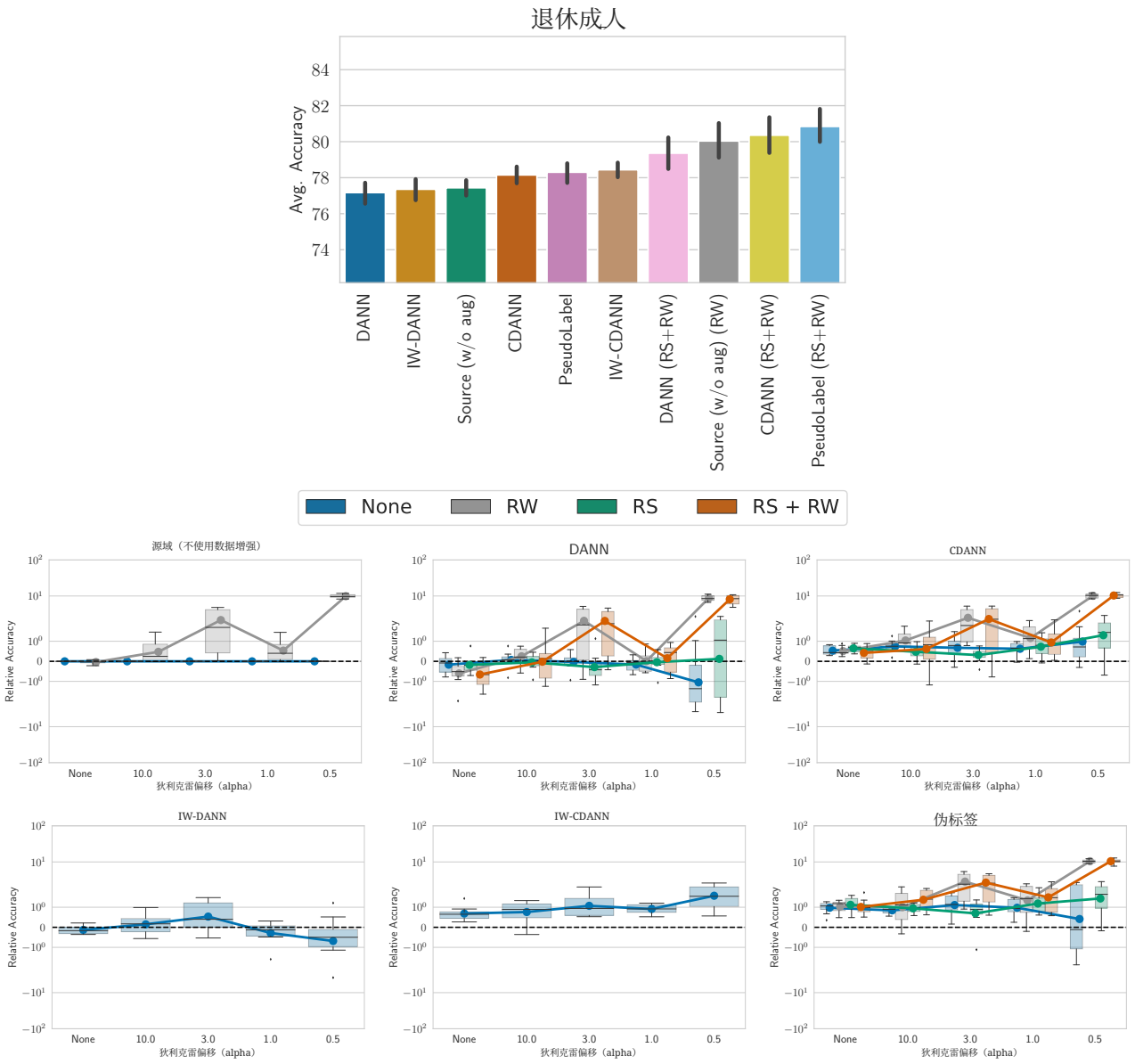


图*30.* 退休成年人。在退休成年人数据集中，不同域适应算法在各种偏移对上的相对性能和准确率图。

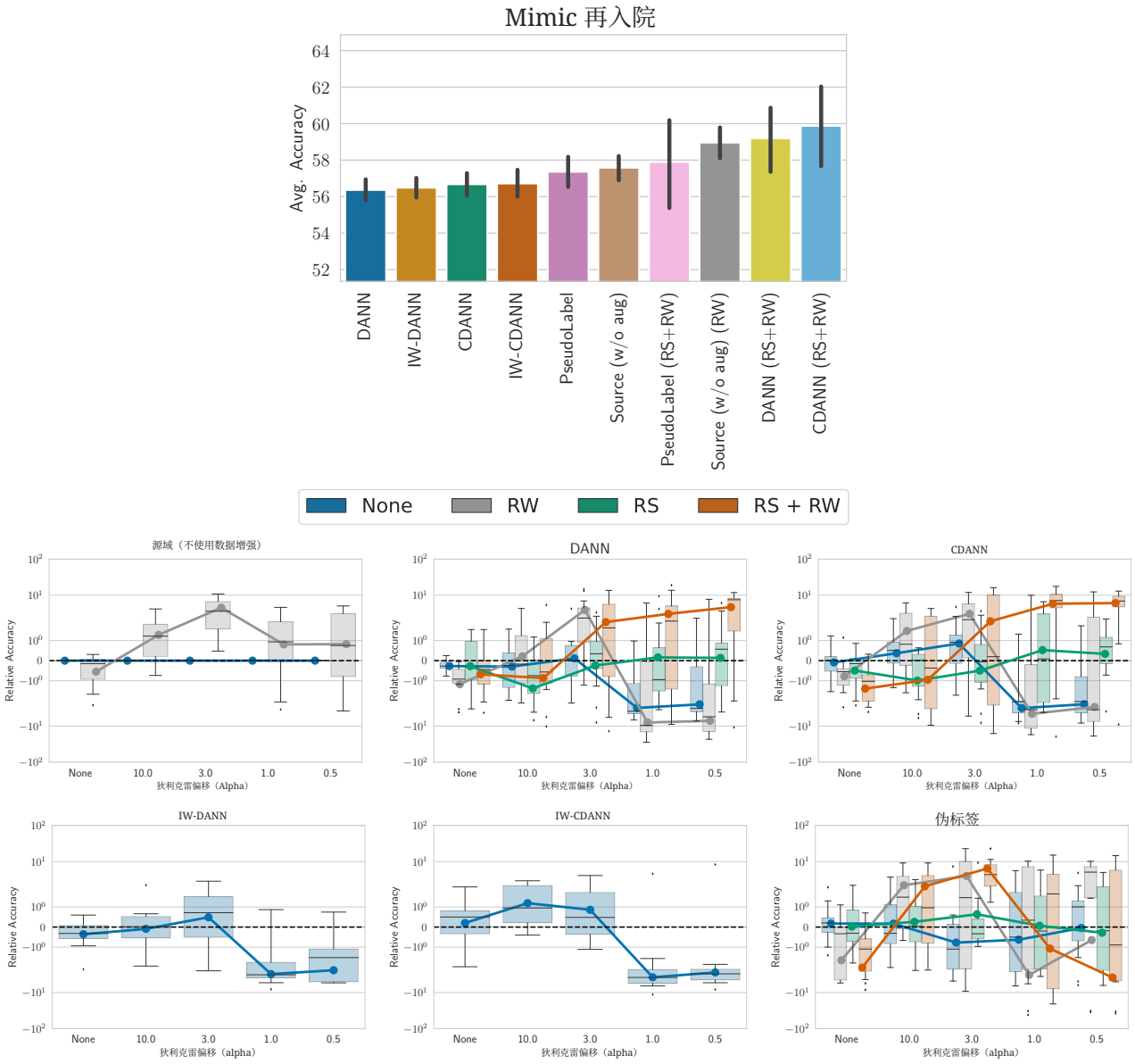


图*31.* MIMIC 再入院。MIMIC 再入院中，不同域适应算法在各类偏移对上的相对性能和准确率图。

K. 各数据集上不同领域自适应方法的总体准确率

数据集	源域	DANN	IW-DANN	CDANN	IW-CDANN	伪标签
CivilComments	86.85	86.62	86.95	86.91	87.16	87.4

数据集	源域	DANN			CDANN			伪标签		
	无 RW	无 RW	RS	RS+ RW	无 RW	RS	RS+ RW	无 RW	RS	RS+ RW
CivilComments	86.8 89.1	86.6 88.8	87.1 88.8		86.9 89.0	86.9 88.9		87.4 89.3	86.9 88.6	

表*2.* 在自然语言处理数据集上采用不同领域自适应方法的结果（跨目标标签边缘分布偏移汇总）。

数据集	源域	DANN	IW-DANN	CDANN	IW-CDANN	伪标签
退休成人	77.44	77.17	77.35	78.15	78.44	78.30
Mimic 再入院	57.57	56.36	56.48	56.67	56.71	57.35

数据集	源域	DANN			CDANN			伪标签		
	无 RW	无 RW	RS	RS+ RW	无 RW	RS	RS+ RW	无 RW	RS	RS+ RW
退休成年人	77.4 80.0	77.2 79.5	77.4 79.4		78.1 80.5	78.1 80.4		78.3 80.8	78.5 80.8	
MIMIC 再入院	57.6 59.0	56.4 55.1	57.3 59.2		56.7 56.8	57.4 59.9		57.4 57.7	57.7 57.9	

表*3.* 在目标标签边缘分布偏移下汇总的表格数据集上，不同领域自适应方法的结果。

数据集	源域 (wo aug)	源域（带数据增强）	BN-适应	TENT	DANN	IW-DAN	CDAN	IW-CDAN	Fix-Match	Noisy-Student	Sentry
CIFAR-10	89.69	89.14	89.21	89.20	90.86	90.78	90.00	89.93	91.87	90.72	91.83
CIFAR-100	65.99	76.69	77.57	77.58	74.80	74.81	74.57	74.66	79.03	77.60	74.74
FMoW	64.00	68.99	65.52	66.55	60.11	60.33	60.79	61.05	68.37	68.90	51.06
Camelyon	77.42	76.95	85.70	82.48	86.66	85.89	85.45	84.27	86.29	79.29	86.81
DomainNet	52.37	50.50	50.66	51.12	51.91	52.05	54.40	54.29	57.96	51.49	55.16
Entity13	76.93	80.07	77.99	78.04	78.26	78.75	79.74	79.28	80.25	80.37	73.58
实体30	62.61	69.83	68.09	68.09	67.90	68.36	68.51	69.34	69.95	69.10	58.51
Living17	64.13	69.30	68.84	68.82	72.12	69.87	70.72	70.65	72.86	72.16	53.44
Nonliving26	54.75	63.95	62.60	63.02	61.69	61.99	62.53	64.51	62.98	63.60	44.82
OfficeHome	59.89	59.45	60.59	60.82	66.05	65.79	66.19	66.15	65.48	60.47	65.37
VisDA	58.47	53.41	59.98	60.96	69.69	69.79	72.55	72.80	72.02	53.51	72.23
Avg	66.02	68.94	69.70	69.70	70.92	70.77	71.40	71.54	73.37	69.75	66.14

表*4*。 在视觉数据集上采用不同领域自适应方法的结果（跨目标标签边缘分布偏移汇总）。尽管没有任何一种领域自适应方法能在不同数据集上始终保持一致的表现，但在我们的测试平台中，FixMatch似乎能够相较于仅源域分类器提供最高的总体提升。

数据集	源域		BN-adapt				CDANN				FixMatch			
	无	RW	无	RW	RS	RS+ RW	无	RW	RS	RS+ RW	无	RW	RS	RS+ RW
CIFAR-10	89.1	89.4	89.2	91.4	92.1	92.9	90.0	91.3	91.4	92.5	91.9	93.1	93.6	94.1
CIFAR-100	76.7	77.5	77.6	78.8	77.9	79.0	74.6	75.8	74.1	75.3	79.0	79.6	79.1	79.8
FMoW	69.0	70.3	65.5	67.2	66.2	65.6	60.8	61.9	57.0	55.2	68.4	69.4	64.9	66.7
Camelyon	77.0	77.9	85.7	85.9	88.5	89.3	85.5	85.8	87.9	88.5	86.3	87.0	86.6	86.8
DomainNet	50.5	48.2	50.7	50.1	51.4	49.8	54.4	54.2	54.7	54.3	58.0	57.5	58.4	57.8
Entity13	80.1	80.9	78.0	79.4	79.8	80.7	79.7	80.2	80.6	81.4	80.3	81.9	81.4	82.4
实体30	69.8	70.1	68.1	69.2	69.1	70.0	68.5	69.6	69.4	70.5	70.0	71.6	70.1	71.2
Living17	69.3	69.9	68.8	69.7	69.6	70.1	70.7	71.3	72.9	72.7	72.9	72.8	72.3	71.9
Nonliving26	63.9	64.5	62.6	63.0	63.7	63.9	62.5	62.9	63.8	64.0	63.0	64.7	63.9	64.8
OfficeHome	59.4	57.9	60.6	60.5	60.9	60.4	66.2	66.3	66.1	65.1	65.5	64.9	66.5	66.1
VisDA	53.4	52.1	60.0	60.6	59.5	58.8	72.6	72.6	75.3	75.3	72.0	72.5	73.5	73.8
Avg	68.9	69.0	69.7	70.5	70.8	70.9	71.4	72.0	72.1	72.3	73.4	74.1	73.7	74.1

数据集	TENT				DANN				NoisyStudent			
	无	RW	RS	RS+ RW	无	RW	RS	RS+ RW	无	RW	RS	RS+ RW
CIFAR-10	89.2	91.4	92.1	92.9	90.9	92.3	91.5	92.6	90.7	90.8	90.6	90.7
CIFAR-100	77.6	78.8	78.0	79.0	74.8	75.9	74.8	76.1	77.6	78.0	77.9	78.0
FMoW	66.6	67.4	66.7	66.1	60.1	61.6	56.4	54.5	68.9	69.8	67.1	68.0
Camelyon	82.5	82.7	87.8	88.9	86.7	87.3	88.4	88.8	79.3	79.1	79.2	79.3
DomainNet	51.1	50.6	51.8	50.3	51.9	52.1	53.6	53.5	51.5	49.8	51.3	49.5
Entity13	78.0	79.5	79.8	80.8	78.3	79.4	79.7	80.8	80.4	81.5	80.6	81.7
实体30	68.1	69.2	69.1	70.1	67.9	69.2	69.0	69.8	69.1	70.1	69.3	70.3
Living17	68.8	69.7	69.6	70.1	72.1	73.0	71.8	72.3	72.2	71.1	69.3	69.4
Nonliving26	63.0	63.4	63.3	63.8	61.7	62.4	63.1	63.0	63.6	64.3	63.2	64.8
OfficeHome	60.8	60.4	60.9	60.4	66.1	66.1	66.5	65.3	60.5	59.5	60.8	59.5
VisDA	61.0	61.5	60.3	59.6	69.7	69.9	73.1	73.2	53.5	51.5	55.7	54.3
Avg	69.7	70.4	70.8	71.1	70.9	71.7	71.6	71.8	69.7	69.6	69.5	69.6

表*5.* 对视觉数据集的目标标签边缘分布偏移进行汇总后，采用重采样（*RS*）和重新加权（*RW*）校正（使用 *RLLS* 估计）的域适应方法结果。RS 和 RW 似乎对所有数据集都有帮助，并且两者结合使用时，能显著提升所有域适应方法相较于不进行校正时的总体性能。

L. 深度领域自适应方法说明

本节总结了我们在 RLSbench 测试平台中比较的深度域自适应方法，并讨论了各方法如何与我们的元算法结合，以处理类别比例偏移。

L.1. 仅源域训练

我们将基于有标签源数据的经验风险最小化作为基线。由于这种方法完全忽略无标签目标数据，我们称之为仅源域训练。正如正文所述，我们分别在使用和不使用数据增强的情况下进行仅源域训练。形式上，我们最小化以下经验风险最小化损失：

$$L_{\text{source only}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(T(x_i), y_i)) \,, \tag{4}$$

其中，T 是用于视觉数据集的随机数据增强操作，ℓ 是损失函数。对于自然语言处理和表格数据集，T 是恒等函数。本文始终采用交叉熵损失最小化。除非另有说明，我们对视觉数据集使用强数据增强作为数据增强技术。对于自然语言处理和表格数据集，我们不使用任何数据增强。

如正文所述，我们不报告仅源域模型的重采样结果，因为该模型仅使用源数据进行训练；在我们的实验中，我们发现，仅平衡源数据并无差异（大多数数据集的源域本身已是平衡的）。获得分类器 f 后，我们可以先估计目标标签边缘分布，然后使用重要性比率 $w_t(y) = \widehat{p}_t(y)/\widehat{p}_s(y)$ 对分类器 f 进行事后重加权。

仅源域模型的对抗训练除了采用数据增强对仅源域模型进行标准训练外，我们还测试了对抗鲁棒模型（Madry 等人，2017年）。为了训练对抗鲁棒模型，我们用鲁棒风险最小化目标替代标准的经验风险最小化目标：

$$L_{\text{source only (adv)}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(R(T(x_i), y_i), y_i) \,, \tag{5}$$

其中，$R(\cdot)$ 执行对抗增强。在本文中，我们使用基于 ℓ_2 扰动模型的定向投影梯度下降（PGD）攻击。

L.2. 领域对抗训练方法

域对抗训练方法旨在学习领域不变的特征表示。这些方法针对支撑集不重叠的实际问题，并受到理论结果的启发：理论结果表明，分布内与分布外性能之间的差距取决于源分布和目标分布之间某种散度度量（Ben-David 等人，2010年a；Ganin 等人，2016年）。在同时最小化源域误差的同时，这些方法会对齐源分布与目标分布的表示。为实现对齐，这些方法会惩罚跨领域特征表示之间的散度，促使模型生成在不同领域中相似的特征表示。

在介绍这些方法之前，我们先定义一些符号。考虑一个模型 $f = g \circ h$，其中 $h:\ \mathcal{X} \to \mathbb{R}^d$ 是将输入映射到某个 d 维特征空间的特征提取器，而预测头 $g:\ \mathbb{R}^d \to \Delta^{k-1}$ 则将特征映射到预测空间。遵循Sagawa 等人（2021年），对于所有领域不变方法，我们都会对视觉数据集中的源域和目标数据使用强数据增强。对于自然语言处理和表格数据集，我们不使用任何数据增强。

DANN DANN 由 Ganin 等人 (2016年) 提出。DANN 利用域判别分类器来近似源域和目标域特征表示之间的散度。域判别器 f_d 旨在区分源域和目标域。给定来自源域和目标域的一批输入后，该深度网络 f_d 会判断样本来自源数据还是目标数据。具体而言，使用以下损失函数：

$$L_{\text{domain disc.}}(f_d) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_d(h(T(x_i))), 0) + \frac{1}{m}\sum_{i=n+1}^{n+m} \ell(f_d(h(T(x_i))), 1) \,, \tag{6}$$

其中 $\{x_1, x_2, \ldots\ldots, x_n\}$ 是 n 源域样本，$\{x_{n+1}, \ldots\ldots, x_{m+n}\}$ 是 m 目标域样本。总体而言，使用以下损失函数来优化采用 DANN 的模型：

$$L_{\text{DANN}}(h, g, f_d) = L_{\text{source only}}(g \circ h) - \lambda L_{\text{domain disc.}}(f_d)\,. \tag{7}$$

$L_{\text{DANN}}(h, g, f_d)$ 相对于域判别分类器最大化，并且 $L_{\text{DANN}}(h, g, f_d)$ 相对于底层特征提取器和源分类器最小化。实践中通过梯度反转层实现这一点。训练时，我们使用三个不同的学习率 η_f，η_g，以及 η_{f_d}。我们在附录 M 中讨论这些超参数的具体细节。我们的 DANN 实现改编自 Sagawa 等人(2021年) 和迁移学习库（Jiang 等，2022年）。

CDANN 条件域对抗神经网络是 DANN 的一种变体（Long 等人，2018年）。这里，域判别器以分类器 g 的预测为条件。具体而言，这些方法不使用 h 的表征输出来训练域判别器，而是对输入 $h(x)$ 处的特征表示 x 与分类器的概率预测 $f = g \circ h(x)$（即 $h(x) \otimes f(x)$）之间的外积进行运算。因此，它们不是在 f_d 的 d 维输入空间中训练域判别分类器，而是在 $d \times k$ 维空间中进行训练。具体使用以下损失函数：

$$L_{\text{CDAN domain disc.}}(f_d, g, h) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_d(f \otimes h(T(x_i))), 0) + \frac{1}{n}\sum_{i=n+1}^{n+m} \ell(f_d(f \otimes h(T(x_i))), 1)\,, \tag{8}$$

其中 $\{x_1, x_2, \ldots\ldots, x_n\}$ 是 n 源域样本，$\{x_{n+1}, \ldots\ldots, x_{m+n}\}$ 是 m 目标域样本。整体损失与 DANN 相同，其中 $L_{\text{domain 判别器}}(f_d)$ 被替换为$L_{\text{CDAN 域判别器}}(f_d, g, h)$。

我们基于迁移学习库（Jiang 等，2022年）改编了 CDANN 的实现。

为了将 DANN 和 CDANN 适配到我们的元算法中，在每个训练轮次，我们可以按照算法 1 的步骤 1 和 4 对源域数据和目标数据进行重新平衡。获得分类器 f后，我们可以使用该分类器先估计目标标签边缘分布，然后利用所得估计值执行重加权调整。

IW-DANN 和 IW-CDANN Tachet 等人（2020年）提出了采用重要性重加权校正的训练方法，通过使用 DANN 和 CDANN 目标函数来适应目标标签比例的变化。具体而言，在每个训练轮次中，他们首先估计重要性比率 $\widehat{w}_t$（使用 BBSE 基于训练源域数据和训练目标域数据进行估计），然后对域判别器目标和经验风险最小化目标进行重新加权。具体而言，IW-DANN 的域判别器损失可写为：

$$L^{\widehat{w}}_{\text{domain disc.}}(f_d) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(y_i)\ell(f_d(h(T(x_i))), 0) + \frac{1}{n}\sum_{i=n+1}^{n+m} \ell(f_d(h(T(x_i))), 1)\,, \tag{9}$$

其中，我们将源域损失乘以重要性权重。同样，我们可以通过重要性重加权将仅使用源域训练的目标函数改写如下：

$$L^{\widehat{w}}_{\text{source only}}(f) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(y_i)\ell(f(T(x_i), y_i))\,. \tag{10}$$

总体而言，使用以下目标函数来优化 IW-DANN 模型：

$$L_{\text{IW-DANN}}(h, g, f_d) = L^{\widehat{w}}_{\text{source only}}(g \circ h) - \lambda L^{\widehat{w}}_{\text{domain disc.}}(f_d)\,, \tag{11}$$

其中，重要性权重会在每个训练轮次结束后，使用上一步得到的分类器进行更新。同样，对 CDANN 目标函数使用重要性重加权后，我们得到 IW-CDANN目标函数。

在总体层面，IW-CDANN 和 IW-DANN 校正与我们针对 DANN 和 CDANN 的元算法校正相匹配。然而，对于具有有限数据的过参数化模型（Byrd & Lipton，2019年），这种重要性重加权校正的行为可能与我们的元算法不同。近期的实证和理论研究结果表明，对于训练多个训练轮次的过参数化模型，重要性重加权几乎没有影响，甚至没有影响（Byrd &

Lipton，2019年；Xu 等人，2021年）。另一方面，在有限样本条件下，重采样（类别标签可用时）已展现出不同且颇具前景的实证表现（An 等人，2020年；Idrissietal.，2022年）。这可能凸显出 IW-CDANN（或 IW-DANN）与我们在 CDANN（或 DANN）上使用的元算法之间的行为差异。

我们参考了作者提供的实现（Tachet 等人，2020年）。

L.3. 自训练方法

自训练方法通过使用分类器自身的预测结果为无标签样本生成“伪标签”，并将其视为有标签样本进行训练，从而利用无标签数据。近期的自训练方法通常还会采用一致性正则化，例如鼓励模型对无标签样本的增强版本作出相似的预测。在我们的工作中，我们对以下方法进行了实验：

伪标签 （Lee 等人，2013年）提出了伪标签方法，该方法利用分类器对无标签样本作出的预测。该算法动态生成伪标签，并在每个批次中对其进行过拟合。具体而言，在无标签样本上生成伪标签后，损失会针对同一标签进行计算。仅当预测的置信度高于某个阈值时，伪标签方法才会对分配的标签进行过拟合 τ。

将 T 称为数据增强技术（即，对自然语言处理和表格数据集使用恒等变换，对视觉数据集使用强增强）。随后，PseudoLabel 使用以下损失函数：

$$L_{\text{PseudoLabel}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(T(x_i), y_i)) + \frac{\lambda_t}{m}\sum_{i=n+1}^{m+n} \ell(f(T(x_i), \widetilde{y}_i)) \cdot \mathbb{I}\left[\max_y f_y(T(x_i)) \geqslant \tau\right],$$

其中 $\widetilde{y}_i = \arg\max_y f_y(T(x_i))$。PseudoLabel 会随着训练轮数增加 λ_t 有标签损失与无标签损失之间的权重，最初将无标签损失的权重设为 0，然后线性增加无标签损失的权重，直至达到超参数 λ 的完整值，并在某个阈值步数达到该值。我们将 λ_t 达到最大值 λ 的步数固定为总训练步数的 40%，与（Sohn 等人，2020年；Sagawa 等人，2021年）中的实现一致。

FixMatch Sohn 等人 （2020年）提出了 FixMatch，作为更简单的伪标签方法（Lee 等人，2013年）的一种变体。该算法会在每个批次中动态生成伪标签，并针对这些伪标签进行过拟合。FixMatch 对无标签数据采用一致性正则化。具体而言，伪标签是在无标签样本的弱增强视图上生成的，而损失则根据强增强视图上的预测计算。这种更新背后的直觉是，促使模型对弱增强数据做出的预测与对强增强样本做出的预测保持一致。此外，只有当强增强预测的置信度高于某个阈值 τ时，FixMatch 才会针对弱增强的有标签数据进行过拟合。

将 T_{weak} 称为弱增强函数，将 T_{strong} 称为强增强函数。然后，FixMatch 使用以下损失函数：

$$\begin{aligned} L_{\text{FixMatch}}(f) = &\frac{1}{n}\sum_{i=1}^{n} \ell(f(T_{\text{strong}}(x_i), y_i)) \\ &+ \frac{\lambda}{m}\sum_{i=n+1}^{m+n} \ell(f(T_{\text{strong}}(x_i), \widetilde{y}_i)) \cdot \mathbb{I}\left[\max_y f_y(T_{\text{strong}}(x_i)) \geqslant \tau\right], \end{aligned}$$

其中 $\widetilde{y}_i = \operatorname{argmax}_y f_y(T_{\text{weak}}(x_i))$。我们的实现改编自 Sagawa 等人（2021年），与 Sohn 等人（2020年）的实现一致，但有一处细节不同。Sohn 等人（2020年）对有标签样本进行弱增强，而 Sagawa 等人（2021年）提出对有标签源样本进行强增强。

NoisyStudent Xie 等人 (2020年b）提出了一种不同的伪标签方法变体。Noisy Student 会生成伪标签并将其固定，然后从头开始训练模型直至收敛，再生成新的伪标签。与 FixMatch 和 PseudoLabel 动态生成伪标签不同，Noisy Student 的第一组伪标签是通过仅使用源有标签数据训练初始教师模型获得的。随后在每次迭代中，随机初始化的模型使用上一轮迭代中已收敛模型分配的伪标签，拟合有标签源数据和伪标签目标数据。Noisy Student 的目标函数可概括为：

$$L_{\text{NoisyStudent}}(f^N) = \frac{1}{n}\sum_{i=1}^{n} \ell(f^N(T_{\text{strong}}(x_i), y_i)) + \frac{1}{m}\sum_{i=n+1}^{m+n} \ell(f^N(T_{\text{strong}}(x_i), \widetilde{y}_i)),$$

其中 $\widetilde{y}_i = \arg\max_y f_y^{N-1}(T_{\text{weak}}(x_i))$ 是使用在 $N-1$ 步获得的分类器计算的。请注意，每次迭代中随机初始化的模型在倒数第二层使用了 $p = 0.5$ 的随机失活。我们采用了 Sagawa 等人 (2021年) 的 NoisyStudent 实现。为初始化初始教师模型，我们使用了通过强数据增强训练且未使用随机失活的仅源模型。

SENTRY Prabhu 等人 (2021年) 提出了一种不同的伪标签方法变体，旨在解决宽松标签偏移情境下的域适应问题。a SENTRY 根据目标实例在由强图像变换组成的集成委员会中的预测一致性，将其纳入训练。具体而言，SENTRY 对一个无标签目标样本进行 N 次强增强，并对增强后的样本进行预测。如果集成委员会的多数投票与对弱增强样本的预测一致，则最小化该样本的熵；否则，最大化其熵。此外，作者还采用了一个“信息熵”目标，旨在使每个样本的预测与估计的目标标签边缘分布相匹配。总体而言，SENTRY 的目标函数定义如下：

$$\begin{aligned} L_{\text{SENTRY}}(f) = &\frac{1}{n}\sum_{i=1}^{n} \ell(f(T_{\text{strong}}(x_i), y_i)) + \frac{1}{m}\sum_{i=n+1}^{m+n}\sum_{j=1}^{k} f_k(y=j|x_i)\log(\widetilde{p}_t(y=j)) \\ &+ \lambda_{\text{unsup}}\frac{1}{m}\sum_{i=n+1}^{m+n}\sum_{j=1}^{k} -f_k(y=j|x_i)\log(f_k(y=j|x_i))\cdot(2l(x)-1), \end{aligned}$$

其中 $l(x) \in \{0,\ 1\}$ 是委员会一致性的多数投票结果。更多详情请参阅 Prabhu 等人 (2021年)。此外，在每个训练轮次中，SENTRY 会对源数据进行重采样，并对目标数据进行伪平衡。我们的实现参考了 Prabhu 等人 (2021年) 的官方实现，并做了少量修改。具体而言，为了与所有其他领域自适应方法保持实现一致，我们从头开始使用上述目标函数进行训练，而不是像原论文（Prabhu 等人，2021年）那样先用仅源域分类器初始化，再进行顺序训练。

由于 Fix-Match、NoisyStuent 和 SENTRY 在实现中使用了强数据增强，这些算法仅适用于视觉数据集。对于自然语言处理和表格数据集，我们只使用伪标签训练模型，因为它不依赖任何增强技术。

L.4. 测试时训练方法

这些方法以已经训练好的源模型为基础，在无标签目标数据上调整少量参数（例如批归一化参数、批归一化统计量），以提升目标性能。因此，我们将这些方法限定在使用批归一化的架构对应的视觉数据集上。由于这些方法会即时调整分类器，其计算开销低于套件中的其他领域自适应方法。我们的实验套件包含以下方法：

BN-adapt Li 等人(2016年) 提出了批归一化自适应。近期，Schneider 等人(2020年) 展示了 BN-adapt 在常见腐蚀基准上的效果提升。批归一化自适应适用于具有批归一化参数的深度模型。使用此方法时，我们只需调整批归一化统计量，具体而言，就是每个批归一化层的均值和标准差。

TENT Wang 等人 (2021年) 提出通过优化批归一化参数，最小化预测器在无标签目标数据上的熵。在我们的 TENT 实现中，我们先执行 BN-adapt，再学习批归一化参数。

CORAL Sun 等人 (2016年) 提出 CORAL，通过对白化特征表示，将在源域上训练的模型适配到目标域 $\widehat{}$表示。具体而言，设 Σ_s 为目标数据表示的经验协方差，Σ_s 为源数据表示的经验协方差，CORAL 会通过在目标域上重新训练最终层，使线性层 g 根据以下输出进行调整：$\Sigma_t^{1/2}\Sigma_s^{-1/2}h(x)$。DARE（Rosenfeld 等人，2022年）简化了这一过程，并指出这等价于训练一个线性头 h，使其作用于 $\Sigma_s^{-1/2}h(x)$，并在将目标数据表示输入分类器之前，使用 $\Sigma_t^{-1/2}h(x)$ 对白化这些表示。我们选择实现后一种方法，因为在多域数据集中训练单个分类器成本较低。

使用我们的元算法，在采用测试时适应方法适应仅源域分类器之前，我们先用它执行重采样校正。获得适应后的分类器后，我们估计目标标签边缘分布，并用它通过重新加权来调整分类器。

M. 超参数和架构详情

M.1. 架构和预训练详情

对于所有数据集，我们在不同算法中使用相同的架构：

- CIFAR-10：ResNet-18（He 等人，2016年）在 ImageNet 上预训练
- CIFAR-100：ResNet-18（He 等人，2016年）在 ImageNet 上预训练
- Camelyon：DenseNet-121（Huang 等人，2017年）未按照（Koh 等人，2021年）中的建议在 ImageNet 上预训练
- FMoW：DenseNet-121（Huang 等人，2017年），在 ImageNet 上预训练
- BREEDs（Entity13、实体30、Living17、Nonliving26）：ResNet-18（He 等人，2016年）未在 ImageNet 上预训练，这遵循了（Santurkar 等人，2021年）中的建议。主要原因是避免在超集数据集上预训练，因为我们在该数据集上模拟子群体偏移。
- OfficeHome：ResNet-50（He 等人，2016年），在 ImageNet 上预训练
- DomainNet：ResNet-50（He 等人，2016年），在 ImageNet 上预训练
- VisDA：ResNet-50（He 等人，2016年），在 ImageNet 上预训练
- CivilComments：预训练的 DistilBERT-base-uncased（Sanh 等人，2019年）
- 退休成年人：我们使用一个包含 2 个隐藏层的多层感知机，每个隐藏层包含 100 个隐藏单元
- MIMIC 再入院：我们使用 Yao 等人 （2022年）[2]

除 CIFAR 数据集上的 ResNet 外，我们使用标准的 PyTorch 实现（Gardner 等人，2018年）。对于 CIFAR 上的 ResNet，我们参考此处的实现：https://github.com/kuangliu/pytorch-cifar。对于所有架构，在适用的情况下，我们都会添加抗锯齿（Zhang，2019年）。我们使用随论文发布的官方库。
对于采用标准架构的 ImageNet 预训练模型，我们使用此处公开提供的模型：
https://pytorch.org/vision/stable/models.html。对于在缩小输入尺寸图像（例如 CIFAR-10）上使用的 ImageNet 预训练模型，我们从头开始在缩小输入尺寸的 ImageNet 上训练模型。我们将该模型随我们的公开代码仓库一同提供。对于基于 BERT 的模型，我们使用此处公开提供的模型：https://huggingface.co/docs/transformers/。

M.2. 超参数

首先，我们调整学习率和 ℓ_2 正则化参数，并为每个数据集固定批量大小，使其达到我们能在 15GB GPU 显存中容纳的最大值。训练轮数则根据相应基准测试作者的建议设定。请注意，我们将训练轮数定义为完整遍历一次带标签的训练源数据。我们在表 ℓ_2 6 中汇总了本研究所用的学习率、批量大小、训练轮数和正则化参数。

对于每种算法，我们采用初始论文中报告的超参数。对于领域对抗方法（DANN 和 CDANN），我们参考迁移学习库（Jiang 等，2022年）中的建议。以下列出了各算法的超参数：

- **DANN、CDANN、IW-CDANN 和 IW-DANN** 根据迁移学习库的建议，使用预训练网络进行初始化时，特征提取器的学习率乘数设为 0.1，否则设为 1.0。对于所有采用预训练初始化的数据集，惩罚权重默认设为 1.0。
- **FixMatch** 我们将 lambda 设为 1.0，并将阈值 τ 设为 0.9。

[2]https://github.com/huaxiuyao/Wild-Time/。

数据集	训练轮次	批量大小	ℓ_2 正则化	学习率
CIFAR10	50	200	0.0001（从 {0.0001，0.001，1e-5，0.0} 中选取）	0.01（从 {0.001，0.01，0.0001} 中选取）
CIFAR100	50	200	0.0001（从 {0.0001，0.001，1e-5，0.0} 中选取）	0.01（从 {0.001，0.01，0.0001} 中选取）
Camelyon	10	96	0.01（从 {0.01、0.001、0.0001、0.0} 中选取）	0.03（从 {0.003、0.3、0.0003、0.03} 中选取）
FMoW	30	64	0.0（从 {0.0001、0.001、1e-5、0.0} 中选取）	0.0001（从 {0.001、0.01、0.0001} 中选取）
Entity13	40	256	5e-5（从 {5e-5、5e-4、1e-4、1e-5}中选择）	0.2（从 {0.1、0.5、0.2、0.01、0.0}中选择）
实体30	40	256	5e-5（从 {5e-5、5e-4、1e-4、1e-5}中选择）	0.2（从 {0.1、0.5、0.2、0.01、0.0}中选择）
Living17	40	256	5e-5（从 {5e-5、5e-4、1e-4、1e-5}中选择）	0.2（从 {0.1、0.5、0.2、0.01、0.0} 中选择）
Nonliving26	40	256	0 5e-5（从 {5e-5、5e-4、1e-4、1e-5}中选择）	0.2（从 {0.1、0.5、0.2、0.01、0.0} 中选择）
OfficeHome	50	96	0.0001（从 {0.0001、0。001、1e-5、0.0} 中选取）	0.01（从 {0.001、0。01、0.0001} 中选取）
DomainNet	15	96	0.0001（从 {0.0001、0。001、1e-5、0.0} 中选取）	0.01（从 {0.001、0。01、0.0001} 中选取）
VisDA	10	96	0.0001（从 {0.0001、0.001、1e-5、0.0} 中选择）	0.01（从 {0.001、0.01、0.0001} 中选择）
CivilComments	5	32	0.01（从 {0.01、0.001、0.0001、0.0} 中选择）	2e-5（从 $\{2e-4, 2e-5\}$ 中选择）
退休成年人	50	200	0.0001（从 {0.01、0.001、0.0001、0.0} 中选择）	0.01（从 {0.001、0.01、0.0001} 中选择）
MIMIC 再入院	100	128	0.0（从 {0.01、0.001、0.0001、0.0} 中选择）	5e-4（从 {0.005、0.00010.0005} 中选择）

表 6。 RLSBENCH 中所采用的学习率和批量大小的详细信息

- **NoisyStudent** 我们重复该流程 2 次迭代，并使用 $p = 0.5$的丢弃级别。
- **SENTRY** 我们使用 $\lambda_{\text{src}} = 1.0$、 $\lambda_{\text{ent}} = 1.0$和 $\lambda_{\text{unsup}} = 0.1$。我们使用了一个规模为 3 的集成委员会。
- **PsuedoLabel** 我们将 lambda 设为 1.0，并将阈值 τ 设为 0.9。

近期工作（Deng 和 Zheng，2021年；Guillory 等，2021年；Chen 等，2021年；Jiang 等，2021年；Baek 等，2022年；Garg 等，2022b）提出了许多启发式方法，用于预测分布偏移下的分类器性能。分析这些启发式方法对超参数选择的实用性，是未来研究的一个有趣方向。

M.3. 计算基础设施

我们的实验使用 Nvidia T4、A6000、P100 和 V100 图形处理器组合完成。总体而言，要在配备 8 个 CPU 核心的 T4 图形处理器机器上运行完整的 RLSBENCH 套件，我们大约需要 $70k$ GPU 小时的计算时间。

M.4. 数据增强

在实验中，我们采用数据增强技术，以增强视觉数据集对领域间某些差异的鲁棒性。

对于弱增强，我们采用随机水平翻转和预定义尺寸的随机裁剪。对于强增强，我们依次应用以下变换：随机水平翻转、预定义尺寸的随机裁剪、使用 Cutout 进行增强（DeVries与Taylor，2017年），以及 RandAugment（Cubuk等人，2020年）。RandAugment 的具体实现直接采用Sohn 等人(2020年) 的实现。操作池包括：自动对比度、亮度、颜色抖动、对比度、直方图均衡、色调分离、旋转、锐度、水平和垂直错切、曝光反转，以及水平和垂直平移。所有实验中，我们均应用 N = 2 个随机操作。

N. 在 OfficeHome 数据集上使用不同超参数与 SENTRY 的比较

本节将进一步探讨 SENTRY 的原始论文（Prabhu 等人，2021年）中报告的结果与我们的实现之间存在差异的原因。

需要指出的是，在 OfficeHome 数据集的主要实验中，我们对所有方法（包括 SENTRY）都使用了 96 的批量大小。然而，SENTRY 在其研究中报告的批量大小为 16。因此，我们使用批量大小 16 重新运行 SENTRY 算法。为考察批量大小减小带来的影响，我们将其与 FixMatch（在我们的实验中，OfficeHome 上表现最佳的算法）进行比较，并以减小后的批量大小重新运行 FixMatch。

在表 7 中，我们报告了 OfficeHome 中各个偏移对上的结果。我们观察到，对于 OfficeHome 数据集标签分布的默认轻微偏移，SENTRY 的表现优于 FixMatch。然而，随着偏移程度加剧，SENTRY 的性能有所下降。总体而言，在 OfficeHome 的 4 个偏移对中，RS-FixMatch 有 3 个表现与 SENTRY 相当或优于 SENTRY。

算法	Alpha = 无	Alpha = 10.0	Alpha = 3.0	Alpha = 1.0	Alpha = 0.5	Avg
FixMatch	92.5	95.2	98.0	100.0	100.0	97.1
RS-FixMatch	92.5	96.4	98.0	100.0	100.0	97.4
SENTRY	93.0	94.0	98.0	83.3	87.5	91.2

(a) 产品到产品（分布内）

算法	Alpha = 无	Alpha = 10.0	Alpha = 3.0	Alpha = 1.0	Alpha = 0.5	Avg
FixMatch	71.4	71.5	70.7	73.1	75.5	72.4
RS-FixMatch	74.7	74.0	72.1	73.1	70.4	72.9
SENTRY	78.1	78.0	75.1	71.7	65.3	73.6

（b）从乘积到实数

算法	Alpha = 无	Alpha = 10.0	Alpha = 3.0	Alpha = 1.0	Alpha = 0.5	Avg
FixMatch	41.5	44.0	44.2	48.4	39.4	43.5
RS-FixMatch	45.5	44.8	43.6	50.0	37.4	44.2
SENTRY	45.8	46.5	41.4	40.3	27.3	40.3

(c) 产品到剪贴画

算法	Alpha = 无	Alpha = 10.0	Alpha = 3.0	Alpha = 1.0	Alpha = 0.5	Avg
FixMatch	54.4	51.3	54.7	57.3	55.9	54.7
RS-FixMatch	57.2	53.6	55.9	57.3	58.8	56.6
SENTRY	63.7	62.0	62.1	65.3	55.9	61.8

(d) 产品到艺术

表7. 批量大小为16（而非我们所有实验中使用的96）的OfficeHome结果。

更普遍地说，在我们的多次运行中，我们还观察到使用 SENTRY 进行模型训练时存在不稳定性。进一步调查后，我们发现用于强制实现一致性的最大化目标会导致不稳定。这种现象在未使用预训练权重初始化基础模型的实验中尤为普遍（例如，在 BREEDs 数据集上）。