傅里叶对抗攻击与最大最小熵算法结合

黎浚

摘要

无监督领域自适应 (Unsupervised domain adaptation, UDA) 包括有标签源域的有监督损失和无标签目标域的无监督损失，由于有监督源域损失具有明显的域间差距，而无监督目标域损失由于缺乏标签而经常有噪声，因此相比经典监督学习通常面临更严重的过拟合。本文复现了一种鲁棒域自适应技术 (Robust Domain Adaptation, RDA)，引入对抗性攻击来缓解 UDA 中的过拟合。RDA 通过一种新的傅里叶对抗攻击 (Fourier Adversarial Attacking, FAA) 方法实现了鲁棒的域适应，该方法允许大幅度的扰动噪声，但对图像语义的修改最小。FAA 方法具有通用性，因此本研究复现了 RDA 算法并将 FAA 方法应用于半监督领域自适应算法最大最小熵中，实验结果表明傅里叶攻击在最大最小熵算法中取得一定的提升。

关键词：无监督领域自适应；傅里叶对抗攻击；半监督领域自适应

1 引言

深度卷积神经网络已经在各种计算机视觉任务中定义了新的技术状态，但其训练过的模型经常过度拟合训练数据，并且由于域间差距的存在，对于来自不同来源的数据会经历明显的性能下降。无监督域适应 (Unsupervised domain adaptation, UDA) 是一种利用无标签目标数据来解决域间差距问题的方法。为此，大多数现有的 UDA 工作涉及源域数据的监督损失和目标域数据的无监督损失，以学习在目标领域表现良好的模型。然而这些方法往往面临更严重的过拟合，因为 UDA 中的有监督源域损失有一个额外的域间差距，而无监督目标域损失由于缺乏注释往往是嘈杂的。几乎所有的深度网络训练都存在过拟合，并在应用于新数据时往往会降低经过训练的深度网络模型的泛化能力。

本文复现了一种鲁棒域适应技术，引入了一种新的傅里叶对抗攻击 (Fourier Adversarial Attacking, FAA) 技术来缓解无监督域适应中的过拟合问题。FAA 通过生成对抗样本来缓解过拟合，防止过度最小化有监督和无监督的 UDA 损失。具体来说，FAA 将训练图像分解为多个频率分量 (FCs)，只扰动捕获很少语义信息的 FCs。与传统攻击限制扰动噪声的大小以保持图像语义完整不同，FAA 允许在其生成的对抗样本中产生较大的扰动，但对图像语义的修改很小。这一特征对于无监督域适应至关重要，无监督域适应通常涉及清晰的域间隙，因此需要具有大扰动的对抗样本。通过在训练中引入 FAA 生成的对抗样本，网络可以继续进行“随机游走”，避免过度拟合和漂移到损失平坦的区域，从而实现更健壮的域适应。

半监督领域自适应方法最大最小熵（Minimax Entropy, MME）通过对域不变原型的对抗性训练，减少源域和目标域间的差异，但是该方法容易过拟合，目标域损失函数收敛效果并不理想，容易产生震荡。因此本研究提出将 FAA 应用于半监督领域自适应方法，利用对抗样本辅助寻找域不变原型，使域不变原型的寻找更准确，减轻模型过拟合的风险。

2 相关工作

2.1 领域自适应

领域自适应已被广泛研究以减轻数据标注的约束。现有的大部分作品大致可以分为三类。第一类是基于对抗训练的，它使用鉴别器来对齐特征、输出或潜在空间中的源域和目标域。第二类是基于图像转换的，它适应图像外观以减轻域间差距。第三类是基于自我训练的，通过预测伪标签或最小化熵来指导目标样本的迭代学习。域自适应涉及两种典型的训练损失，即有标签源域数据上的监督损失和无标签目标域数据上的无监督损失。最新的方法倾向于过度最小化这两种类型的损失，这直接导致模型的偏差和次优自适应。

2.2 在网络中的过拟合

过拟合是深度网络训练中普遍存在的现象，在深度学习和计算机视觉领域得到了广泛的研究。现有的大多数工作通过某些正则化策略来解决过拟合问题，如权重衰减、dropout、l1 正则化、混合、标签平滑、批量归一化、虚拟对抗训练、flooding 等。然而，这些策略大多是为监督或半监督学习设计的，不太适合包含域间差距和无监督损失的领域自适应学习。

2.3 对抗攻击

对抗攻击在各种安全问题中都有研究。已有工作通过快速梯度符号、最小对抗性扰动、普遍对抗性扰动、无梯度攻击、可转移对抗性样本生成等从不同方面改进了对抗性攻击。对抗性攻击也被应用于其他任务，例如，使用对抗性样本来缓解有监督和半监督学习中的过拟合，生成对抗性样本用于数据增强，增强可转移特征以最小化域间差异。现有的大多数对抗性攻击方法通常限制扰动噪声的大小，以实现对图像语义的最小修改。然而，这种生成的对抗样本不能很好地解决领域自适应学习中的过拟合问题，这通常涉及到相当大的领域差距。本文复现的论文[1]设计了一种创新的傅里叶对抗性攻击技术，允许在没有扰动大小限制的情况下生成对抗性样本，同时对图像语义进行最小的修改。

2.4 半监督领域自适应

半监督领域自适应（Semi-supervised domain adaptation, SSDA）尚未得到充分的探索[2]。半监督领域自适应中 Yves Grandvalet 和 Y oshua Bengio 使用标准熵最小化有标记源域及目标域数据和无标记目标域数据的熵，但未考虑到源域和目标域数据的共同点。Saito K 等人使用最大熵和最小熵的对抗性训练寻找域不变原型，考虑到了域间的共同点，但该方法容易导致过拟合和收敛不佳。因此本研究期望通过傅里叶攻击的方式，允许在不受幅度限制的情况下生成对抗样本，但对图像语义的修改最少，从而限制测试集损失的过度最小化达到阻止过拟合的目的。

3 本文方法

3.1 健壮领域自适应

如图 1 所示，RDA 通过傅里叶对抗性攻击实现了鲁棒域自适应。训练分为两个阶段，即进攻阶段和防守阶段。给定一个训练图像，攻击阶段学习识别具有有限语义信息的正确 FCs，允许大量级的扰动噪声。它还学会了在对图像语义进行最小修改的情况下生成对抗样本。在防御阶段，生成的对抗样本用于通过防止过度最小化训练损失来缓解过拟合。

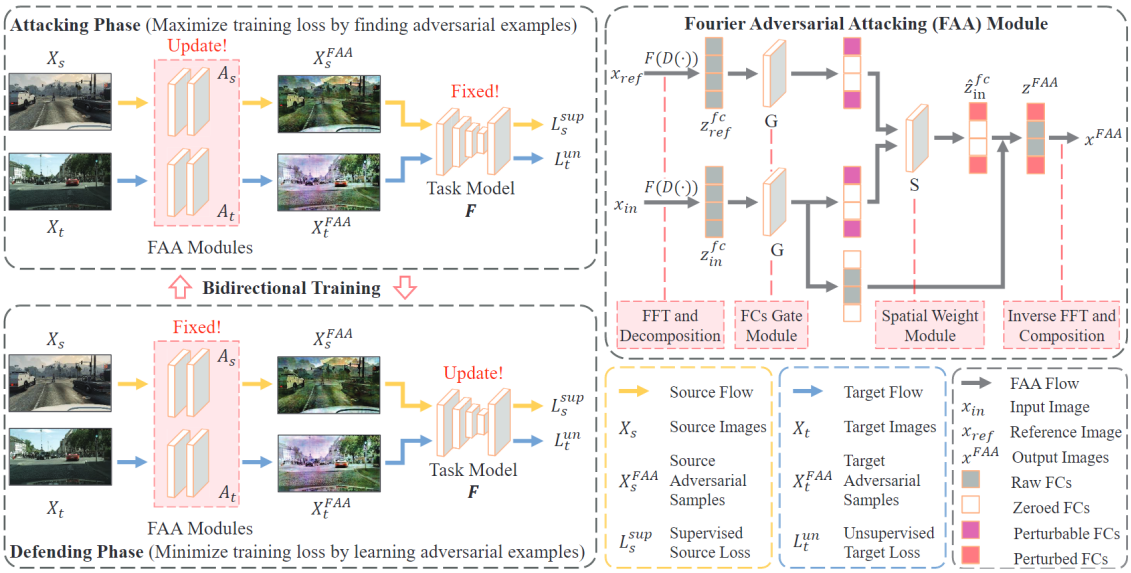

图 1: RDA 算法整体流程图

3.1.1 傅里叶分解

在傅里叶分解阶段将一个图像 x 变换到到频域空间并分解为多个频率分量 FC，从而允许对每个 FC 进行显式地操作和可控地扰动。在此阶段使用傅里叶变换将 x 转换到频域空间，并进一步进行分解为等效带宽的分量。

$$\begin{aligned} z &= F(x) \\ z^{fc} &= D(z; N) \end{aligned} \tag{1}$$

其中 $F(\cdot)$ 表示傅里叶变换，z 为 x 在频域空间的表示，$D(z; N)$ 表示将 z 分解为 N 个等效带宽的频域分量。在分解过程中使用 $z(i, j)$ 表示频域空间的复数坐标而 c_i 和 c_j 表示图像的质心，则等效带宽分量 $z^1, z^2, z^3, z^4, \ldots, z^n = D(z; N)$ 可以被定义为

$$z^n(i,j) = \begin{cases} z(i,j), if \dfrac{n-1}{N} < d((i,j),(c_i,c_j)) < \dfrac{n}{N} \\ 0, otherwise \end{cases}$$

其中 $d(\cdot)$ 为欧式距离，N 为分量数量，n 表示分量序号。

3.1.2 对抗攻击

在分解频域分量后，通过扰动部分频域分量攻击域适应损失。并引入一个可学习的门模块去选择适合用于扰动的频域分量。门模块 G 对每一个分量使用二值评分，“1”表示适合扰动，“0”表示不适合扰动，通过相乘筛选出适合扰动的分量。门模块 G 使用 Gumbel-Softmax 实现，是一种可通过标准反向传播训练的分类变量的可微抽样机制。在目标域随机选择一张图像用频域分量表示为 $z_{ref}^{fc} = z_{ref}^1, z_{ref}^2, z_{ref}^3, \ldots, z_{ref}^n = D(F(x_{ref}); N)$，使用门模块筛选出适合扰动的分量并用于扰动图像 x 的频域分量。

$$z_{ref}^{fc} = (1 - G(z^{fc}))z^{fc} + G(z^{fc})z_{ref}^{fc} \tag{2}$$

对于被识别的输入图像 x，提取参考图像 x_{ref} 的对应分量，并将其用作扰动噪声。与许多现有对抗性样本生成方法中的随机噪声相比，这种方式生成对抗性样本使用来自目标自然图像的扰动噪声更加合理和有意义。此外，使用目标样本的无语义 FC 减少了域间差距，这有助于提高域自适应中的目标域性能。

最后对生成的频域分量合并，进行逆傅里叶变换即可得到生成图像 x^{FAA}。

$$x^{FAA} = F^{-1}(C(z^{fc})) \tag{3}$$

3.2 最大最小熵

MME 架构主要由特征提取器 F 和分类器 C 组成，分类器 C 具有权重向量 W 和温度系数 T，W 被训练为最大化无标记样本的熵而 F 被训练为最小化熵。为了实现对抗性学习，通过梯度反转层翻转无标记样本的熵损失梯度符号。下图是 MME 算法的流程图。

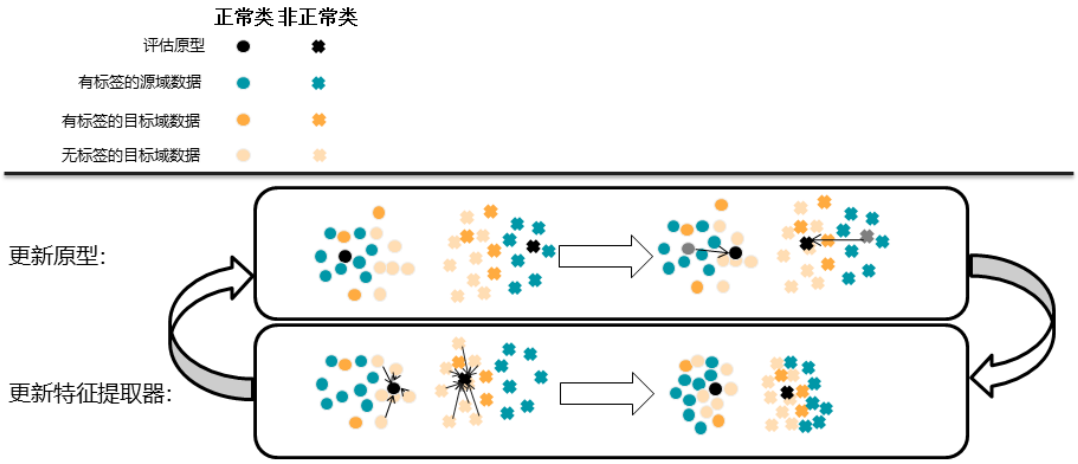

图 2: MME 算法流程图

3.2.1 估计原型

MME 假设每个标签都有一个域不变原型，首先使用特征提取器 F 和分类器 C 估计原型，特征提取器使用深度卷积神经网络作为特征提取器的神经网络结构，并且 F 和 C 使用标准交叉熵损失作为损失函数：

$$L = E_{(x,y)\in D_s,D_t} L_{ce}(p(x), y) \tag{4}$$

通过这种方式提取出鉴别特征只生成了源域与有标记目标域样本相关的鉴别特征，并没有学习到整个目标域的鉴别特征，因此使用无标记的目标域样本进行最大最小熵的训练。

3.2.2 最大最小熵训练

算法目标是让原型向域不变原型靠拢，而估计得到的原型更偏向于源域，因此使用无标记的目标域样本的进行训练，使估计原型靠近每一个无标记目标域样本的特征，因此引入熵 H 来实现上述目

标:

$$H = -E_{(x,y)\in D_u} \sum_{i=1}^{k} p(y=i|x) log p(y=i|x) \tag{5}$$

其中 K 是分类的数量，而 $p(y=i|x)$ 表示预测到第 i 个类的概率，即 $p(x) = (\frac{1}{T}\frac{W^T F(x)}{||F(x)||})$，取得更高的熵意味着每个鉴别特征都与目标域特征相似从而估计得到域不变原型。为了在无标记目标域样本中获得鉴别特征，还需围绕估计得到的原型对无标记的目标域样本进行聚类。算法使用特征提取器 F 降低无标记目标域样本的熵，通过降低原型与无标记样本的熵可以得到鉴别特征。重复这个最大熵和最小熵的过程就可以获得所需要的鉴别特征。总之 MME 算法可以表述为分类器 C 和特征选择器 F 之间的对抗性学习。分类器训练最大熵，特征选择器训练最小熵，整体的对抗性学习目标函数为：

$$\theta_F = argminL + \lambda H \tag{6}$$

$$\theta_C = argminL - \lambda H \tag{7}$$

其中 λ 是用来控制最大最小熵训练和样本分类之间的平衡。

3.3 基于傅里叶对抗攻击的最大最小熵算法

基于上述的最大最小熵半监督领域自适应框架以及傅里叶对抗生成图像的方法，本研究提出将二者进行结合，通过傅里叶攻击对目标域损失函数进行限制，防止其过度最小化。模型整体流程图如图 3 所示：

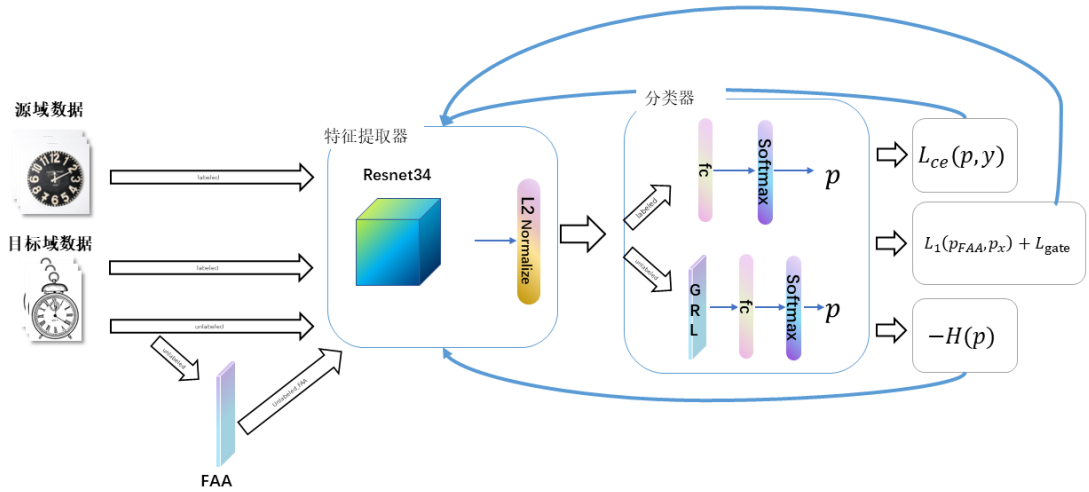

图 3: 算法总体流程图

在输入阶段将无标签的目标域图像进行傅里叶对抗攻击，生成的对抗图像输入特征提取器和分类器中进行训练，得到攻击图像的预测概率 p_{FAA}, 并与原图像分类概率 p 求损失作为重建损失，最后与门损失求和反向传播，在 MME 中加入 FAA 的损失函数进行训练，FAA 损失函数如下式：

$$L = L_{rec} + L_{gate} \tag{8}$$

4 复现细节

4.1 与已有开源代码对比

本次复现参考了 RDA 算法和 MME 算法的开源代码。本研究的主要工作在于在使用 RDA 算法开源代码整体框架的同时对其核心 FAA 方法进行复现，并将 FAA 方法应用到 MME 算法中。工作量主要体现在 FAA 方法的复现，并与 MME 算法结合，在 MME 算法中加入傅里叶对抗攻击环节，重写 FAA 方法的结构使之适配 DomainNet 数据。

4.2 实验环境搭建

本次复现 RDA 算法中使用 **GTA5⟶Cityscapes** 进行语义分割。**GTA5** 数据集包含 24966 张合成图片并包含了 19 个和 **Cityscapes** 共同的类。在将 FAA 方法与 MME 算法结合的实验中使用 **DomainNet** 进行域适应训练，一个最近的大规模领域自适应基准数据集，具有 345 个类和 6 个域。由于某些域和类的标签非常杂乱，我们选择了其中两个域作为源域和目标域，Real 作为源域，Sketch 作为目标域，目标域中每个类随机选取 3 个样本作为有标签的目标域数据进行半监督领域自适应。

本实验采用 Python 语言的 Pytorch 框架编程实现本文所有深度学习算法。其中使用的平台工具为：Python3.6、Pytorch-GPU-1.10.2、机器的 CPU 配置为 Intel 公司的 Intel(R) Xeon(R) Platinum 8255C 2.50GHz, 内存为 43.0G, 显卡为 NVIDIA 公司的 RTX 2080 Ti 显存为 12.0G, 操作系统为 Ubuntu。

在 RDA 算法复现实验中，使用 DeepLabv2 作为预训练模型，学习率设置为 $2.5e^{-4}$，最大迭代次数设置为 80000 次，FAA 学习率设置为 e^{-6}。

在 MME 算法实验中以 ResNet34 作为特征提取器的神经网络结构并移除最后一层全连接层并用构建的分类器 F 替代。在每次迭代中，准备了两个 minibatch，一个由已标记的样本组成，另一个由未标记的目标样本组成。一半的标注样本来自源域，一半来自标注的目标域。采用动量为 0.9 的 SGD 优化器，温度系数 T 设置为 0.005，权衡参数 λ 设置为 0.1。

4.3 创新点

本次研究创新点在于，提出将应用于无监督领域自适应的 FAA 方法与半监督领域自适应方法 MME 进行结合，使用 FAA 生成对抗图像辅助训练，本人认为生成的对抗图像可以防止目标域损失过度最小化，同样适用于大多数标签未知的半监督领域自适应，可以解决 MME 算法收敛不理想，容易过拟合的问题。

5 实验结果分析

5.1 评价指标

在 RDA 语义分割实验中，使用交并比和均交并比作为评价指标。而在 MME 与 FAA 结合的半监督领域自适应分类实验中，使用准确率作为评价指标。

5.1.1 交并比

交并比 IoU 计算的是“预测的边框”和“真实的边框”的交叠率，即它们的交集和并集的比值。最理想情况是完全重叠，即比值为 1。

$$IoU = \frac{target \cap prediction}{target \cup prediction} \tag{9}$$

5.1.2 均交并比

均交并比 mIoU 是数据集中每个类交并比的平均。

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} IoU_i \tag{10}$$

5.2 实验结果

复现 RDA 语义分割结果如表 1 所示，复现效果与原文复现效果相比，表现较差，推测是在复现过程中图像傅里叶变换过程中舍弃了虚部导致精度下降，并且最大迭代次数设置过小。实验结果如表 2 所示，验证集损失曲线图如图 4 所示，可见本次实验提出的 MME_FAA 算法相比现有的算法具有一定优势，表现为准确率更高和收敛速度更快。

Method	Road	SW	build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
baseline	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
Baseline+FFA_T	92.59	52.16	82.13	27.60	24.28	40.60	38.31	20.57	85.91	42.15	73.29	64.39	21.03	87.63	49.76	53.07	0.59	37.45	4.98	47.29
Recurrence	79.19	11.44	78.11	18.33	21.12	19.71	32.99	12.62	78.11	6.08	74.36	58.68	29.51	82.68	21.85	38.33	1.50	17.53	12.11	36.54

表 1: RDA 复现语义分割结果

模型名称	**准确率**
MME	61.2
DANN	54.9
ADR	51.1
CDAN	59
MME_FAA[本文]	**63.4**

表 2: MME_FAA 实验结果

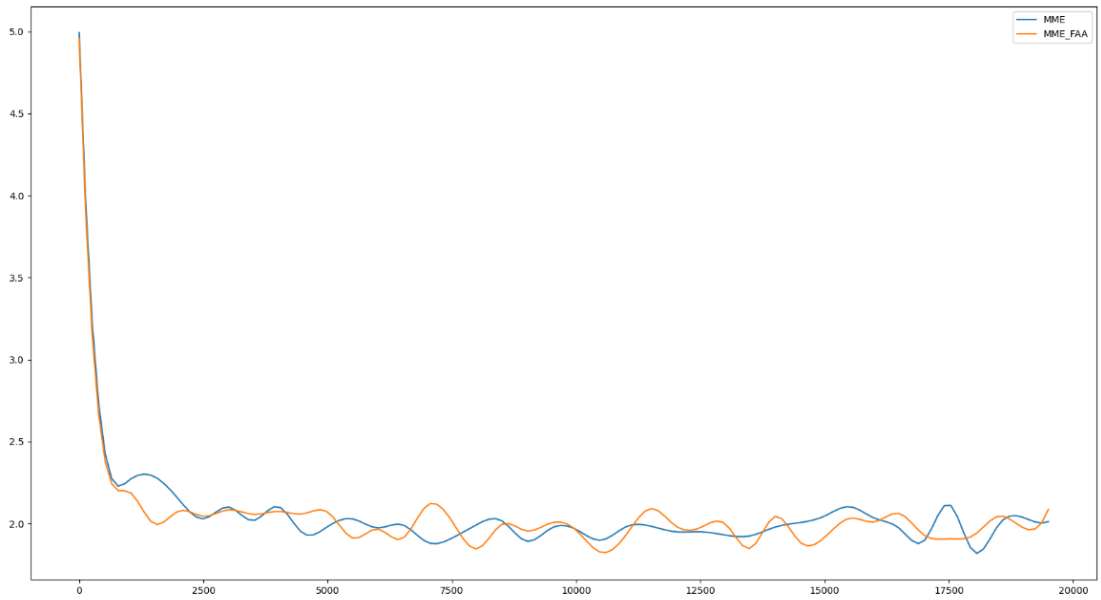

图 4: MME 收敛效果图

6 总结与展望

本次复现工作复现了 RDA 算法框架中的核心 FAA 方法，并将无监督领域自适应的 FAA 方法应用在半监督领域自适应算法 MME 中，得到了一定的提升。本次复现仍然有值得改进的地方：在 RDA 复现的模型训练过程中，参数设置不当，导致模型收敛不完全导致效果弱于论文中的指标；在与 MME 算法结合的时候参数选取还不完全，可以进行更细致的调参以取得更好的效果。

参考文献

[1] HUANG J. RDA: Robust Domain Adaptation via Fourier Adversarial Attacking[J]. Proceedings of the IEEE/CVF international conference on computer vision, 2021: 8988-8999.

[2] K S. Semi-supervised domain adaptation via minimax entropy[J]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8050-8058.