南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

毕业设计说明书

作　　者：罗文水　　学　号：918106840738

学　　院：计算机科学与工程学院

专业(方向)：计算机科学与技术

班　　级：9181062301

题　　目：面向开放环境的标签噪声学习方法研究

指导者：宫辰　教授

评阅者：

2022　年　6　月

声　明

我声明，本毕业设计说明书及其研究工作和所取得的成果是本人在导师的指导下独立完成的。研究过程中利用的所有资料均已在参考文献中列出，其他人员或机构对本毕业设计工作做出的贡献也已在致谢部分说明。

本毕业设计说明书不涉及任何秘密，南京理工大学有权保存其电子和纸质文档，可以借阅或网上公布其部分或全部内容，可以向有关部门或机构送交并授权保存、借阅或网上公布其部分或全部内容。

学生签名：

年　　月　　日

指导教师签名：

年　　月　　日

毕业设计说明书中文摘要

如今，深度学习在实际生活中广泛应用，它的成功依赖于大型数据集的获取和可靠的标注过程。然而，获取高质量的数据集代价高且耗时长，遂通过众包、搜索引擎等途径获得廉价数据集的方式应运而生，这样的数据集难免含有混合的开集和闭集标签噪声，使图像识别准确率严重下降。针对该问题，本文提出了 SelectionToSieve(S2S)模型，该模型构建了一对多分布以外样本（Out-of-Distribution，简称 OOD）检测网络，利用连续性损失以保证 OOD 检测曲线的连续性。在预热阶段之后，该模型将数据集划分为干净样本集、分布以内样本集、分布以外样本集，随后借助半监督学习方法进一步提升分类的准确率。大量的实验证明了 S2S 模型的有效性。在闭集标签噪声和混合标签噪声两种设定的实验中，与 6 个基准模型相比，S2S 模型都获得了较好的分类准确率。此外，消融实验证明了损失函数的权重在一定范围内时，该模型依然保持较好的性能，实验中的各个损失成分都对分类准确率的提升有一定帮助。

关键词　混合标签噪声　开集标签噪声　图像分类　深度学习

毕业设计说明书外文摘要

Title A Study on Open-World Label Noise Learning

Abstract

Nowadays, deep learning is widely applied, which is highly dependent on the well-curated datasets and reliable annotation process. However, it's time-consuming and costly to acquire high-quality datasets. Cheap datasets obtained through crowdsourcing or search engines are alternatives but they will inevitably contain combined Open-set and Closed-set label noise, which seriously deteriorates the classification performance. In this paper, the SelectionToSieve (S2S) model is proposed to address this problem, which uses the OOD detection network for Out-of-Distribution samples detection. Furthermore, S2S develops continuity loss based on the continuity of the OOD detection score. After the warm-up phase, the dataset is divided into three parts: clean sets, In-Distribution sets and Out-of-Distribution sets, based on which Semi-Supervised Learning method is used to further improve the performance of classification. Numerous experiments have proved the effectiveness of the proposed model. Compared with the six baseline models, the S2S model achieves the better or comparable classification accuracy in the settings of Closed-set label noise and combined label noise. In addition, the experimental results of ablation study proves that the model still maintains good performance when the weights of different loss are in certain ranges, and each loss component in the overall loss contributes to the improvement of classification accuracy.

Keywords Combined Label Noise Open-set Label Noise Image Classification Deep Learning

目　次

1 绪论

本章中主要介绍了混合开闭集标签噪声的学习问题在实际中的工程背景和意义、相关研究方法的现状，同时介绍了本文提出的方法的简要技术方案，分析了经济因素和可实现性。最后，本章介绍了论文的结构安排。

1.1 工程背景及意义

如今，机器学习算法在实际中广泛应用，然而机器学习算法非常依赖完备而标注正确的数据集。在大数据时代，如何获得可用于算法训练的可靠标准数据成为一个亟待解决的问题，高质量的数据收集和标注往往消耗过多的人力物力，如 ImageNet[22]大型图像分类数据集的形成对资源和时间的消耗非常巨大。在计算机视觉领域，算法性能的提升与数据集的规模有紧密联系，卷积神经网络[21]的成功很大程度上依赖于数据集的完整性和标注的正确性。研究[23]发现卷积神经网络很容易拟合标签噪声，即使没有任何泛化能力的卷积神经网络仍然很容易训练。现有的大部分模型对噪声数据的鲁棒性较差，因此，标签噪声学习这一课题应运而生。

传统的标签噪声[28]指的是在监督学习中，训练数据的标注存在错误，而潜在的正确标签属于被研究的类别。开放环境下的标签噪声[2]指在监督学习中，训练数据的标签存在标注错误，且潜在的正确标签可能属于被研究的类别，也可能不属于被研究的类别，其中不属于被研究类别的样本又被称为分布以外样本（Out-of-Distribution Sample，简称 OOD）。若企图通过搜索引擎快捷地获取数据集将会获得包含混合开闭集标签噪声的数据集。如图 1 所示，如果想要获得包含“狼”和“狗”的数据集[10]，我们可以在 Google 等搜索引擎中搜索“狼”关键词用于建立“狼”标签对应的训练数据，该训练数据将会同时包含闭集标签噪声和开集标签噪声。图 1 中蓝色标注的图像，属于“狗”类别却被认为是“狼”，故该样本存在闭集标签噪声；此外，橙色边框的图像不属于“狼”类别也不属于“狗”类别，却被标注为“狼”标签，故该样本存在开集标签噪声。所以混合标签噪声在实际的数据集中非常普遍，即使是精心标注的数据集中也可能存在混合标签噪声从而限制了算法性能的提升。

标签噪声学习旨在通过具有标签噪声的数据训练一个鲁棒的分类器，使得该分类器在未知的测试数据上能够得到较高的分类准确率，从而在一定程度上缓解对高质量的数据集的依赖，所以该课题在实际中具有很高的应用价值。

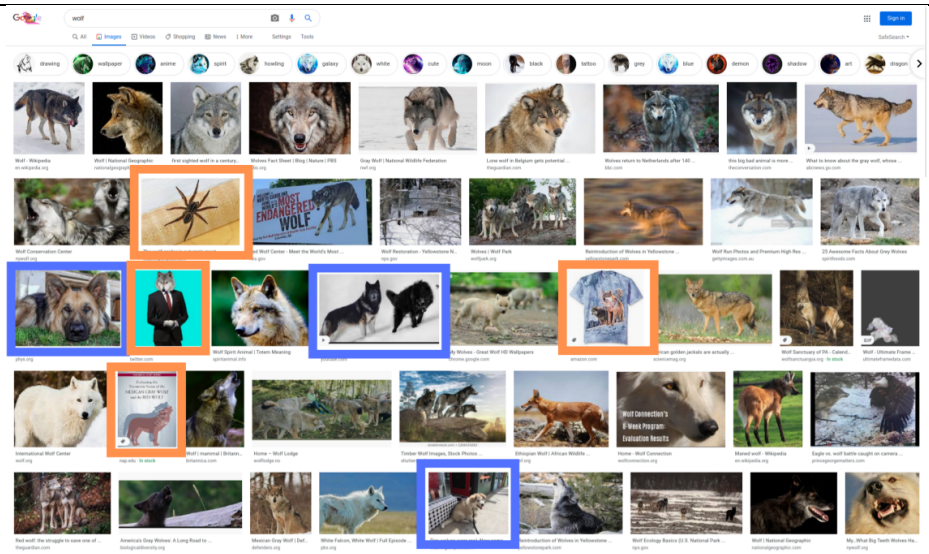

图 1 关键词为“狼”时搜索引擎的搜索结果

1.2 相关技术的现状

根据文献，针对开放环境的标签噪声研究方法主要有如下几个方面：

- **设计具有鲁棒性的网络结构**

Wang 等人[3]在 2018 年首次提出了开放环境下的标签噪声学习问题，在 Wang 的论文[3]中提出了一种迭代的开放环境标签噪声学习框架，此模型使用孪生网络、局部异常因子等方法减小相似样本之间特征表示的距离，扩大噪声样本和干净样本之间的距离，并使用对比损失和加权的 softmax 损失进行梯度反向传播，从而在特征空间中将不同类别的样本分离，并将噪声数据和干净数据分离；Lee 等人[11]提出了 CleanNet 模型，此模型基于注意力机制、自编码器等对样本损失进行加权，并使用余弦相似度评估特征的相似性，但是模型的劣势在于需要额外的验证信息，而在实际情况下额外的验证信息是难以获取的。

- **设计具有鲁棒性的损失函数**

Ghosh 等人[6]提出了多分类噪声学习下一致性损失函数的充分性条件，并从理论和实验上证明了平均绝对损失函数对标签噪声的鲁棒性，此文提出的一致性损失函数设计原则已成为众多一致性损失函数的理论保证；Wang 等人[2]提出了对称的交叉熵损失函数，在分类交叉熵损失的基础上添加对称项，他们从理论上证明了使用该对称项在噪声数据集上学习到的最优分类器等价于干净数据集上的全局最优分类器，同时证明了对称交叉熵损失项与平均绝对损失函数的等价性；Zhang 等人[7]提出了广义的交叉熵损失函数并从理论上证明了该损失函数的一致性，此方法使用交替凸搜索优化方法学习参数，并且在闭集和开集标签噪声设定下均有良好的实验结果。Ma 等人[8]提出了标准化的损失函数，并将损失函数分为激活和非激活两部分，以同时解决鲁棒损失函数下训练的欠拟合和过拟合问题。此外，还有研究提出了其他具有鲁棒性的损失函数[25-27]。

- **设计样本选择方法**

Han 等人[1] 基于神经网络先学习训练集中的简单模式再学习困难模式的特性[24]提出了 Co-teaching 模型，该模型中使用两个网络，按一定比例分别筛选两个网络中损失较小的样本并交给另一个网络进行梯度反向传递和参数更新，该设计准则简洁有效，是对神经网络记忆能力的合理利用；Yu 等人[4]提出了 Co-teaching+模型，此方法选择损失较小的样本后，再从中选择两个网络预测类别不一致的样本交给对方进行参数更新，利用网络的分歧以达到两个网络互相监督的目的；Li 等人[9]提出了 DivideMix 模型，此模型同时应用了半监督学习的 MixMatch 方法、Co-teaching 策略、高斯混合模型、数据增广等方法，获得了各种噪声场景下分类准确率的大幅度提升；Sachdeva 等人[10]提出了 EvidentialMix 模型，此方法使用高斯混合模型对干净样本集、闭合噪声样本集、开集噪声样本集上的损失进行建模；Wu 等人[12]提出了 NGC 模型，通过标签传递、图筛选等方法过滤噪声样本，此模型具有优越的开放环境噪声学习能力。Xia 等人[17]在样本筛选的基础上利用定向对抗样本进一步增强模型的泛化能力，获得了不错的实验结果。

- **设计具有鲁棒性的正则化方法**

标签噪声学习中的正则化技术可以分为隐式正则化和显式正则化。Wei 等人[13]提出了 ODLN 模型，该模型使用含开集标签噪声的数据对抗固有的闭集标签噪声，在实验上，该模型验证了开集标签噪声样本可以促进模型对闭集标签噪声的学习，以此获得更好的模型泛化能力，另外该方法在分布以外样本检测任务中也获得了较好的结果；Xia 等人[14]提出了具有鲁棒性的早期学习方法，对重要/非重要的样本采用不同的梯度更新策略，获得了较高的测试集分类准确率；Chen 等人[15]对标签引入动态的随机高斯噪声，并从理论上证明了引入噪声后损失对参数的梯度包含不同尺度的信息, 有助于解决网络预测分布尖锐的问题，因此该策略提高了模型的泛化能力，缓解了对训练集噪声的过拟合问题。

- **借助自监督表示学习或对比学习的方法**

近期，开放环境下的标签噪声学习的研究者们格外关注自监督学习方法，例如 Feng 等人[16]提出的 S3 模型，将训练过程分为两个阶段，第一阶段实施样本选择并重置标签，第二阶段同时做干净样本的监督学习和整个数据集的自监督学习，通过特征表示的一致性抵制网络预测的自信偏置，该模型在开放环境的噪声学习中得到了最好的测试准确率，在噪声率高的场景下(90%的对称标签噪声)优势更为显著; Nodet 等人[20]对预训练网络的不同利用方式进行了对比实验，实验验证了通过对比学习预训练的网络在标签噪声学习中的微调有利于提高测试准确率。

除了上述方法外，开放环境下标签噪声的学习方法还有标签转移矩阵法[18]、元学习[19]

等等。

与本课题密切相关的几个研究领域为：开放环境识别问题（Open-set Recognition）、开放环境下的半监督学习问题（Open-set Semi-Supervised Learning）、分布以外检测问题（Out-of-Distribution Detection）、异常值检测问题（Anomaly Detection）、领域适应（Domain Adaptation）等。其中开放环境识别问题旨在使用干净的训练数据集训练分类器，使其能够在测试阶段拒绝分布外样本，并在分布以内样本上获得较好的分类性能。Lu 等人[28]考虑到数据本身的不确定性，提出了新的方法用于筛选高质量且具有多样性的类别原型，从而指导模型对测试集样本进行准确分类并拒绝其中的 OOD 样本；分布以外检测或异常值检测问题旨在以较高的准确率筛选出训练数据中存在异常的样本，对该问题的研究[30-34]与本课题亦密切相关；对于开放环境下的半监督学习，训练数据集中有标记的样本都是干净样本，而未标记的样本中除了待分类的样本还存在分布以外样本，这些分布以外的样本所属的数据流形与分布以内样本有本质不同，因此该设定下学习目标有两个，其一是正确分类指定类别的样本，其二是过滤未标记样本中的分布以外样本。该领域的主要研究方法[35-38]以深度学习为主干，通过设计优化过程或损失以提高分布以内样本的分类精度。

1.3 总体技术方案及其社会影响

本论文提出了名为SelectionToSieve（S2S）的方法用于解决含开放环境的标签噪声的图像分类问题。该方法首先基于神经网络的记忆能力，使用交叉熵损失函数预训练网络，其次使用数据增广技术[39]得到一张图像的两种特征表示，依靠JS散度划分出干净样本，将剩余样本视作无标记样本。然后，该方法设计了一个OOD样本检测网络模块，用一对多损失$\mathcal{L}_{ova}$提升该网络对OOD样本的检测能力；为了使模型对于OOD样本的检测更具确定性，该方法使用了极小化熵正则化损失项$\mathcal{L}_{e}$；另外，为了保证OOD检测网络输出的连续性，该方法设计了对图像弱增广和强增广模式下的一致性（连续性）损失函数$\mathcal{L}_{con}$；为了充分利用无标记数据中的分布以内样本，该方法将OOD检测器预测为ID的样本构建$\mathcal{D}_{id}$数据集。在网络输出足够确定的情况下，该方法使用网络输出标签作为伪标签，构建伪标签和网络输出标签概率分布的交叉熵损失$\mathcal{L}_{id}$，按照一定权重将损失加权后得到最终的损失函数。本论文使用基于梯度的自动微分方法[40]训练网络从而得到最终模型。

该方法在闭集标签噪声和混合开闭集标签噪声实验评估中都展现了优秀的性能，测试集上的分类准确率超过了与之对比的6个基准模型。在实际应用中，该方法可以对具有混合开闭集标签噪声的数据集进行自动清理并自动学习模型，从而减少了人力物力和财力损耗，提高

了计算机利用含标签噪声的数据集自动分类样本的能力。

1.4 技术方案的经济因素分析

该方案在经济上对资源的消耗较小，性价比高。大多数传统数据集都是通过众包[22]、在线标注等方法获取，然而对于爬虫或搜索引擎等自动获取的图像数据集，后续需要大量人力和时间进行标签修正，由于并非所有标注者都是专家，标签的标注错误普遍存在，如医学病灶图像的识别问题，专家的数量少，标注自然存在大量开集和闭集噪声。本方案的实施减少了雇佣标注者的经济成本，使得深度学习模型对大型数据集的正确性依赖得以缓解，节省了构建更加鲁棒的模型的经济成本。同时，更高的分类精度和更好的 OOD 样本剔除表现使得该模型在进一步应用中能够减少更多的成本。

1.5 论文章节安排

本论文首先在第一章分析了研究课题的工程背景和意义，介绍了相关的研究方法和本论文提出的 S2S 模型，并分析了该方案的社会影响和经济优势；在第二章中，本论文符号化定义了开放环境的标签噪声学习问题，提供了算法的直观解释和动机；在第三章中，本论文描述了算法的基本原理，定义了各个损失的数学表达式，并提供了详细的算法流程；在第四章中，本论文列举了实验设定、数据集、基准模型等，对比了模型的实验结果，并列举了大量的消融实验以证明方法的有效性。最后，本论文以简短的结论和展望收尾。

2 问题定义

本章公式化定义了开放环境下的标签噪声学习问题，阐述了本课题的预备知识；其次，本章第二节阐述了算法的动机，为后续提出的算法给出了直观的解释。

2.1 预备知识

标签噪声指在监督学习中，训练样本数据的标签存在标注错误，目前可以分为以下两种情形：1. 闭集标签噪声（Closed-set Label Noise）：样本的标签存在错误，但是潜在的正确标签属于被研究的类别；2. 开集标签噪声（Open-set Label Noise）：样本的标签存在标注错误，且潜在的正确标签不属于被研究的类别。闭集标签噪声可细分为对称标签噪声、非对称标签噪声、成对标签噪声、基于实例的标签噪声等等；开集标签噪声则是来源于训练数据分布以外样本的混入。在实际环境中，训练数据集中往往既存在闭集标签噪声也存在开集标签噪声。

标签噪声普遍存在于开放环境中。根据现有研究，开放环境的设定有如下三个方面：1. 训练数据无噪声，而在测试阶段会遇到分布以外或不属于研究类别的样本，如人脸识别任务，测试阶段会存在不属于被研究类别的人脸；2. 训练数据存在闭集或开集标签噪声，而测试场景不存在分布以外样本，即无需模型在测试阶段有拒绝样本的能力；3. 训练数据存在闭集或开集标签噪声，测试数据集中存在分布以外样本。本论文解决的是第二种开放环境设定下的学习问题。

在传统图像分类问题下，训练样本为$\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{n}\sim\mathcal{D}_{clean}$，其中$\mathcal{D}_{clean}$为无标签噪声数据的分布，标签空间为$\mathcal{Y}=\{1,2,\ldots,C\}$。标签$\mathbf{y}_i$满足$\sum_{j=1}^{C}\mathbf{y}_i^j=1$，是样本的独热标签（one-hot encoding）。而在开放环境的标签噪声设定下，训练样本为$\{(\mathbf{x}_i,\tilde{\mathbf{y}}_i)\}_{i=1}^{n}\sim\tilde{\mathcal{D}}_{noisy}$，真实标签记为$\{\mathbf{y}_i\}_{i=1}^{n}$。对于闭集标签噪声学习问题，假定噪声率函数为$\eta(\mathbf{x}_i)$，则样本$(\mathbf{x}_i,\tilde{\mathbf{y}}_i)$以概率$1-\eta(\mathbf{x}_i)$满足$\tilde{\mathbf{y}}_i=\mathbf{y}_i$，以$\eta(\mathbf{x}_i)$概率满足$\mathbf{y}_i\neq\tilde{\mathbf{y}}_i\sim T(\mathcal{Y},\theta_{noisy})$，其中$T(\mathcal{Y},\theta_{noisy})$是一个依靠一定概率分布随机选择噪声标签的函数，所依靠的概率分布由参数θ_{noisy}控制。对于混合开闭集标签噪声的学习问题，训练样本分为干净的分布以内（ID）样本，即$(\mathbf{x}_i,\tilde{\mathbf{y}}_i)$满足$\tilde{\mathbf{y}}_i=\mathbf{y}_i$、含有闭集标签噪声的 ID 样本，即$(\mathbf{x}_i,\tilde{\mathbf{y}}_i)$满足$\tilde{\mathbf{y}}_i\neq\mathbf{y}_i$且$\mathbf{y}_i$在研究的类别中、含有开集标签噪声的样本$(\mathbf{x}_i,\tilde{\mathbf{y}}_i)$，满足$\tilde{\mathbf{y}}_i\neq\mathbf{y}_i$且$\mathbf{y}_i$不在研究类别中。分类的目标是学习一个深度判别网络$f(\mathbf{x};\Theta)\in[0,1]^C$，表示给定样本$\mathbf{x}_i$，在参数$\Theta$条件下的预测标签概率分布。网络的预测标签为$\hat{\mathbf{y}}_i=\arg\max_j f_j(\mathbf{x}_i;\Theta)$。

2.2 动机与直观解释

假定已经分离出部分高概率的干净样本和含有噪声的无标记样本，则本论文对于每一个类别$j \in \mathcal{Y}$都设计一个 OOD 检测器$F^j(\cdot)$，表示样本不属于第j个类别的概率，则$F^j(\cdot)$对于输入图像的微小变动应当是具有连续性的，如图 2 所示。对于未知类别的 OOD 样本，其与每个已知类别的特征表示之间不相似，从而在不添加连续性正则化损失之前$F^j(\cdot)$无法学习到平滑的 OOD 概率曲线，而在添加了连续性正则化损失之后，对样本及其增广样本之间的连续性差距变小，从而 OOD 检测器能够获得连续性更好的分布曲线。由图 2 中可见，改进后的 OOD 检测器能够获得与真实的 OOD 概率分布更接近的分布，从而更好地指导开集噪声设定下模型的训练。

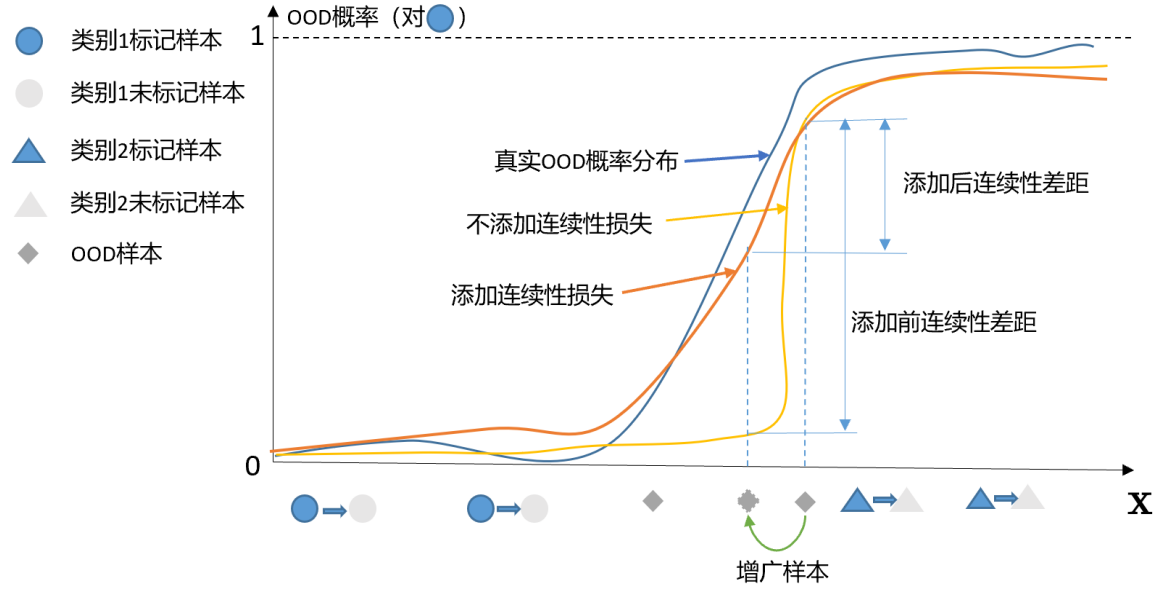

图 2 连续性损失的动机与图例解释

3 算法设计

本章主要介绍了 SelectionToSieve（S2S）模型的原理，解释了模型中使用的各个损失，并提供了详细的算法步骤。

3.1 算法的原理与说明

本论文提出的 S2S 模型中首先使用特征提取网络得到输入图像的特征表示$\mathcal{V}(\mathbf{x};\Theta_1)\in\mathcal{R}^d$，其中$\Theta_1$表示该特征提取网络的参数，$d$ 为特征表示的维度；其次，该模型定义了分类网络模块，该网络模块用$f(\mathcal{V}(\mathbf{x};\Theta_1);\Theta_2)\in\Delta^{C-1}$表示，其中$\Delta^{C-1}$表示$C-1$维度的单位单纯形，该分类网络模块将图像的特征表示作为输入，将$\mathcal{Y}$上的概率分布作为输出。故用于监督分类的交叉熵损失函数如公式(1)所示。

$$\mathcal{L}_{CE}\left(\tilde{\mathcal{D}};\Theta_1,\Theta_2\right)=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\tilde{\mathbf{y}}_i^j\log f_j(\mathcal{V}(\mathbf{x}_i;\Theta_1);\Theta_2), \tag{1}$$

其中$\tilde{\mathbf{y}}_i^j$表示标签$\tilde{\mathbf{y}}_i$的第j个元素。考虑到深度神经网络倾向于在训练早期首先记住训练数据中简单的模式[1]，S2S 模型首先使用交叉熵损失对网络进行预训练（热启动），其次根据 JS 散度衡量网络输出的概率分布与给定的标签分布之间的距离d_i，如公式(2)所示，将网络输出$f(\mathcal{V}(\mathbf{x};\Theta_1);\Theta_2)$记为$\mathbf{q}_{(\Theta_1,\Theta_2)}(\mathbf{x}_i)$，则有：

$$\begin{aligned} d_i &= JS\left(\mathbf{y}_i||q_{(\Theta_1,\Theta_2)}(\mathbf{x}_i)\right) \\ &= KL\left(\mathbf{y}_i||\frac{\mathbf{y}_i+q_{(\Theta_1,\Theta_2)}(\mathbf{x}_i)}{2}\right)+KL\left(q_{(\Theta_1,\Theta_2)}(\mathbf{x}_i)||\frac{\mathbf{y}_i+q_{(\Theta_1,\Theta_2)}(\mathbf{x}_i)}{2}\right), \end{aligned} \tag{2}$$

其中$KL(\mathbf{p}||\mathbf{q})=\sum_{i=1}^{C}p_i\log\frac{p_i}{q_i}$表示离散概率分布$\mathbf{p}$和$\mathbf{q}$之间的 KL 散度。由于 JS 散度的值在 0 和 1 之间，具有概率的特性，所以 S2S 模型考虑将某样本被预测为干净样本的概率定义为：

$$P_{clean}(\mathbf{x}_i)=1-d_i\ \in\ (0,1).$$

接着，通过设定阈值的方式筛选出高概率的干净样本，即：

$$\mathcal{D}_c=\left\{(\mathbf{x}_i,\tilde{\mathbf{y}}_i)\in\tilde{\mathcal{D}}|P_{clean}(\mathbf{x}_i)>\tau_{clean}\right\},$$

从而将训练集划分成$\mathcal{D}_c$和$\mathcal{D}_u=\left\{\mathbf{x}_i|(\mathbf{x}_i,\tilde{\mathbf{y}}_i)\in\tilde{\mathcal{D}}\wedge(\mathbf{x}_i,\tilde{\mathbf{y}}_i)\notin\mathcal{D}_{clean}\right\}$。基于上述样本集合，可以进入后续阶段的网络学习。

第二阶段中，S2S 模型定义了 one-vs-all 网络模块作为 C 路 OOD 样本检测器，用于判断输入样本是否是 OOD 样本。具体地，对于每个类别j，都有预测函数$F^j(\mathcal{V}(\mathbf{x},\Theta_1),\Theta_3^j)\in[0,1]^2$

分别表示样本$\mathbf{x}$被预测为第j个类别的 ID 样本或 OOD 样本的概率，其中Θ_3^j表示第j个类别的 OOD 检测器的参数，本论文以$\Theta_3=\{\Theta_3^j\}_{j=1}^{C}$表示 OOD 检测器的总体参数。为了简化公式，本文将F^j的两个输出维度分别记为$p_\theta^j(ood|\mathbf{x})$和$p_\theta^j(id|\mathbf{x})$，显然$p_\theta^j(ood|\mathbf{x})+p_\theta^j(id|\mathbf{x})=1$成立，其中$\theta=(\Theta_1,\Theta_3)$表示待学习的特征提取器和 OOD 检测器的网络参数。为了让 OOD 检测网络更好地识别 OOD 样本，本模型需要设计合理的损失函数，例如给定第j类的正样本，我们期望在第j类样本的 OOD 检测器上输出$p_\theta^j(id|\mathbf{x})>p_\theta^j(ood|\mathbf{x})$，在其他类别$i\neq j,\ \forall\ i\in\mathcal{Y}$上输出的概率值$p_\theta^j(ood|\mathbf{x})$更大。研究者们在 OOD 检测领域经常使用 one-vs-all 损失[41]（ova）用于训练 OOD 检测器，one-vs-all 损失如公式(3)所示。大量的实验证明了 one-vs-all 损失用于 OOD 检测的优越性[37]。假定网络训练的一个干净数据批次为$\mathcal{B}=\{(\mathbf{x}_i^c,\mathbf{y}_i^c)\}_{i=1}^{|\mathcal{B}|}\subseteq\mathcal{D}_c$，$\mathbf{y}_i^c\in\mathcal{Y}$为实值标签，则本模型中 ova 损失如下所示：

$$\mathcal{L}_{ova}(\mathcal{B};\Theta_1,\Theta_3)=\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}-\log p_\theta^{\mathbf{y}_i^c}(id|\mathbf{x}_i^c)-\min_{k\neq\mathbf{y}_i^c}\log p_\theta^k(ood|\mathbf{x}_i^c). \tag{3}$$

观察 one-vs-all 损失，我们可以发现：该损失同时优化两个目标，其一是在极大化干净样本被预测为 ID 样本的概率，其二是极大化干净样本在其他类别上被预测为 OOD 样本的概率下界。

此外，为了在预热阶段之后更全面地利用$\mathcal{D}_c$数据集并提高分类器的判别能力，本模型考虑在第二阶段将交叉熵损失加入监督损失项，构成损失$\mathcal{L}_{sup}$，如公式(4)所示：

$$\mathcal{L}_{sup}=\mathcal{L}_{ova}(\mathcal{B};\Theta_1,\Theta_3)+\mathcal{L}_{CE}(\mathcal{B};\Theta_1,\Theta_2). \tag{4}$$

随后，我们需要设计对于$\mathcal{D}_u$的损失以充分利用无标记数据中的有利信息。假定神经网络每个批次的无标记训练数据为$\mathcal{U}=\{\mathbf{x}_i^u\}_{i=1}^{|\mathcal{U}|}\subseteq\mathcal{D}_u$，一般地，在半监督学习中无标记数据批次的样本数大于有标记数据批次的样本数，按照研究设定[37]，有$|\mathcal{U}|=\mu|\mathcal{B}|$，在本模型中固定参数$\mu=2$。其次，本文考虑使用数据增广技术[39]对无标记批次中的图像进行增广，常用的数据增广方法有随机剪裁、水平翻转、分辨率调整、颜色调整等，同时也有强增广方法如 Randaugment[50]、Augmix[39]等。假定数据增广方法为$\mathcal{A}(\cdot)$，按照$\mathbf{x}_{aug}=\mathcal{A}(\mathbf{x})$得到增广后的数据，图像$\mathbf{x}$的不同增广表示为$\mathcal{A}_1(\mathbf{x})$和$\mathcal{A}_2(\mathbf{x})$。理论上，同一张图像在数据增广之后具有类别不变形，故在特征空间中同一张图像的不同增广的特征表示$\mathcal{V}(\mathcal{A}_1(\mathbf{x});\Theta_1)$和$\mathcal{V}(\mathcal{A}_2(\mathbf{x});\Theta_1)$应当具有较小的距离，同时，不同增广得到的标签概率分布$f(\mathcal{V}(\mathcal{A}_1(\mathbf{x});\Theta_1);\Theta_2)$和$f(\mathcal{V}(\mathcal{A}_2(\mathbf{x});\Theta_1);\Theta_2)$应当相似，对应的 OOD 概率分布也应当具有相似性，如图 2 中的真实 OOD 概率曲线所示。所以 S2S 模型中考虑使用连续性损失，如公式(5)所示。期望在不同增

广下的图像能够获得近似的 OOD 概率分布，其中$\theta=(\Theta_1,\Theta_3)$表示可学习的特征提取网络参数和 OOD 检测网络参数。

$$\mathcal{L}_{con}=\frac{1}{|\mathcal{U}|}\sum_{i=1}^{|\mathcal{U}|}\sum_{j=1}^{C}\sum_{k\in\{id,ood\}}|p_\theta^j(k|\mathcal{A}_1(\mathbf{x}_i^u))-p_\theta^j(k|\mathcal{A}_2(\mathbf{x}_i^u))|^2. \tag{5}$$

在连续性损失中，由于 OOD 样本本身和每个类别的样本都具有较小的相似性，所以在无标记数据中，OOD 样本属于容易产生混淆的样本，然而 OOD 检测器的连续性是与标签无关的，故可以使用所有的无标记数据作为连续性正则化损失的输入。

为了使得 OOD 检测器的输出具有确定性，即防止 OOD 检测的结果退化成$p_\theta^j(ood|\mathbf{x})\approx 0.5$的情况，本模型中使用了极小化熵正则化损失，如公式(6)所示。

$$\mathcal{L}_e=-\sum_{i=1}^{|\mathcal{U}|}\sum_{j=1}^{C}p_\theta^j(id|\mathbf{x}_i^u)\log p_\theta^j(id|\mathbf{x}_i^u)+p_\theta^j(ood|\mathbf{x}_i^u)\log p_\theta^j(ood|\mathbf{x}_i^u). \tag{6}$$

于是，总的损失函数如公式(7)所示，其中λ_{con}和λ_e分别表示连续性损失和极小化熵正则化损失的权重。

$$\mathcal{L}(\mathcal{B},\mathcal{U};\Theta_1,\Theta_2,\Theta_3)=\mathcal{L}_{sup}+\lambda_{con}\cdot\mathcal{L}_{con}+\lambda_e\cdot\mathcal{L}_e. \tag{7}$$

为了更好地利用无标记的数据，S2S 模型在使用上述损失训练一定周期之后，在每个训练周期初都从无标记数据$\mathcal{D}_u$中筛选出可能的 ID 样本构成$\mathcal{D}_{id}$，将被认为是分布以外的样本集$\mathcal{D}_u-\mathcal{D}_{id}$从无标记数据中剔除。随后该模型将网络提供的伪标签作为样本 $\mathbf{x}_i^u\in\mathcal{D}_{id}$的实际标签。具体地，假如对于样本$\mathbf{x}_i^u$，网络的输出结果为$\mathbf{q}_i^u=f(\mathcal{V}(\mathbf{x}_i^u;\Theta_1);\Theta_2)$，则其对应的伪标签为$\hat{q}_i^u=arg\max_{j\in\mathcal{Y}}\mathbf{q}_{i,j}^u$，若此时 OOD 检测网络得到$p_\theta^{\hat{q}_i^u}(id|\mathbf{x}_i^u)>p_\theta^{\hat{q}_i^u}(ood|\mathbf{x}_i^u)$，则认为该样本为第$\hat{q}_i^u$个类别的 ID 样本，将其加入$\mathcal{D}_{id}$集合。假定$\delta(C)$表示示性函数，仅当条件$C$为真时满足$\delta(C)=1$，否则$\delta(C)=0$。随后，本模型使用如公式(8)所示的损失函数加入网络的训练，其中 H 表示将$\hat{q}_i^u$转变成独热编码之后与$f(\mathcal{V}(\mathcal{A}(\mathbf{x}_i^u);\Theta_1);\Theta_2)$构成的交叉熵损失。其中示性函数是为了从 ID 样本中去除处于决策边界或难以区分的 ID 样本，这些样本容易对结果产生不利影响，而对于远离决策边界的样本，网络输出的概率分布将会具有较低的熵，更具确定性，从而对应的最大概率值较大，这些样本对分类器性能的提升至关重要。此外，公式(8)中τ表示实验设定的超参数。

$$\mathcal{L}_{id}=\frac{1}{|\mathcal{D}_{id}|}\sum_{\mathbf{x}_i^u\in\mathcal{D}_{id}}\delta(\max_{j\in\mathcal{Y}}\mathbf{q}_{i,j}^u>\tau)\mathrm{H}(\hat{q}_u^i,f(\mathcal{V}(\mathcal{A}(\mathbf{x}_i^u);\Theta_1);\Theta_2)). \tag{8}$$

综合上述，S2S 模型分为三个阶段。第一阶段为预热网络阶段，该模型只使用交叉熵损失

函数训练网络；第二阶段初该模型筛选出高概率的干净样本，使用监督损失、连续性正则化损失、极小化熵损失训练网络；在第三阶段，S2S 模型进一步筛选分布以内样本，在总体损失中加入对置信度高的 ID 样本的监督损失项。

3.3 算法的实现步骤

本论文中提出的 SelectionToSieve（S2S）算法流程如图 3 所示，其中 3-10 行使用交叉熵损失函数预热网络，以充分利用神经网络在训练早期对简单样本的记忆能力；算法中第 11-13 行对干净数据进行筛选，并构成干净数据集和无标记数据集；第 14-16 行在筛选阶段对无标记样本中的 ID 样本进行筛选；第 17-26 行利用筛选过的干净样本和分布以内样本进行深度网络的训练和参数的更新；最终，算法输出各网络模块的参数。

Algorithm 1: Selection2Sieve(S2S)

Input: 噪声数据 $\tilde{\mathcal{D}} = \{(\mathbf{x_i}, \tilde{\mathbf{y}}_\mathbf{i})\}_{i=1}^{n}$ 可学习的特征提取网络 $\mathcal{V}(\cdot;\Theta_1)$、分类网络 $f(\cdot;\Theta_2)$、OOD 检测网络 $F^j(\cdot;\Theta_3^j)\ \forall\, j \in \mathcal{Y}$, 总体网络参数为 θ. 权重 λ_{con}, λ_e, λ_{id}. 训练周期数 E. 预热网络周期数 selection_epochs, 筛选开始周期数 sieve_epochs. 学习率 η. 每个周期迭代次数 iters. 阈值 τ_{clean}, τ.

Output: 网络参数 $\theta = (\Theta_1, \Theta_2, \Theta_3)$.

for e ***from*** *1* ***to*** E **do**
  $step = 0$
  // 预热网络
  if e<*selection_epochs* **then**
    while $step < iters$ **do**
      采样 $\mathcal{B} \subseteq \tilde{\mathcal{D}}$
      计算 $\mathcal{L}(\mathcal{B};\theta) = \mathcal{L}_{CE}(\mathcal{B};\Theta_1,\Theta_2)$
      $\theta = \theta - \eta \cdot \nabla_\theta \mathcal{L}(\mathcal{B};\theta)$
      $step = step + 1$
    end
  else
    // 划分干净样本集和无标记样本集 $\mathcal{D}_c, \mathcal{D}_u$
    if e=*selection_epochs* **then**
      划分 $\mathcal{D}_c \subseteq \tilde{\mathcal{D}}$; $\mathcal{D}_u = \tilde{\mathcal{D}} - \mathcal{D}_c$
    end
    // 筛选 ID 样本
    if e>*sieve_epochs* **then**
      筛选 $\mathcal{D}_{id} \subseteq \mathcal{D}_u$; $\mathcal{D}_{ood} = \mathcal{D}_u - \mathcal{D}_{id}$
    end
    while $step < iters$ **do**
      采样 $\mathcal{B} \subseteq \mathcal{D}_c$; $\mathcal{U} \subseteq \mathcal{D}_u$
      $\mathcal{L}(\mathcal{B},\mathcal{U};\theta) = \mathcal{L}_{sup} + \lambda_{con}\mathcal{L}_{con} + \lambda_e\mathcal{L}_e$
      if e>*sieve_epochs* **then**
        采样 $\mathcal{B}_{id} \subseteq \mathcal{D}_{id}$
        $\mathcal{L} += \lambda_{id}\mathcal{L}_{id}(\mathcal{B}_{id};\theta)$
      end
      $\theta = \theta - \eta \cdot \nabla_\theta \mathcal{L}(\mathcal{B},\mathcal{B}_{id},\mathcal{U};\theta)$
      $step = step + 1$
    end
  end
end

图 3　S2S 算法步骤

4 实现与测试

本章中详细介绍了实验的设定与基准评估方法，包括闭集标签噪声的人为生成方法、混合开闭集标签噪声的生成方法、模型的评估标准。其次，第二节中介绍了所使用的数据集和与 S2S 模型对比的基准模型；随后，第三节中介绍了在两种实验设定下的各模型的实验结果。最后，本章用大量的消融实验深入探讨了各部分损失模块的重要性、不同网络架构的影响以及不同超参数对实验结果的影响。

本论文所有的程序都基于 Pytorch[40]实现，所有实验都在两张 NVDIA GeForce RTX 2080Ti 显卡上完成，程序已经在 github 上开源：https://github.com/randydkx/S2S 。

4.1 混合开闭集标签噪声的评估方法

根据文献，仅依赖于标签而不依赖于实例的的人造标签噪声是通过设计标签转移矩阵[42-43]实现的，即设计转移矩阵T，使得$T_{ij}(\mathbf{x})=T_{ij}=P(\tilde{y}=j|y=i)$。随后，根据该转移矩阵我们可以得到含标签噪声的训练数据集。标签噪声的类型可能是对称的、非对称的、成对的、依赖于实例[1]的，由于本文中考虑仅依赖于标签的标签噪声，所以实验仅在对称和非对称标签噪声两种设定下进行。此外，虽然本论文中提出的算法是专门针对混合开闭集标签噪声学习问题的，但是该算法依然能够处理仅含闭集标签噪声的问题。

在闭集标签噪声设定下，噪声转移矩阵如图 4 所示（仅展示十个类别的转移矩阵），其中左侧图像表示标签含 40%对称噪声的转移矩阵，右侧图像表示标签含 40%非对称噪声时的转移矩阵。具体地，假设噪声率为η，则在对称标签噪声设定下，样本以概率$1-\eta$保持原来的标签，并以均匀概率$\eta/(C-1)$转变成$\{1,2,\ldots,C\}$中的任意一个非正确标签。在非对称标签噪声情形下，根据通常的实验设定[44]，本论文考虑使用成对翻转转移矩阵，按照η比率筛选样本并将其标签设置成类别集合中的下一个标签，若该样本的标签为C，则将其标签设置成 1。

对于混合开闭集标签噪声设定，噪声转移矩阵如图 5 所示，其中左右两侧图像分别表示生成 40%对称闭集标签噪声和 20%开集标签噪声的转移矩阵、生成 40%非对称闭集标签噪声和 20%开集标签噪声的转移矩阵。假定开集噪声率为η，类别数量为C，则本文考虑将数据集中的前$(1-\eta)\cdot C$个类别作为 ID 类别，即实际需要分类的类别有$(1-\eta)\cdot C$个，其余$\eta\cdot C$个类别表示 OOD 类别。假定闭集标签噪声率为ξ，则标签转移矩阵中的左上角分块与闭集标签噪声生成方法相同，分块中闭集标签噪声为ξ。对于属于 OOD 类别的样本，每个样本的标签

以均匀的概率转变成$\{1,2,\ldots,(1-\eta)\cdot C\}$中的任意一个标签，故经过噪声污染的数据集包含$(1-\eta)\cdot C$个类别，总体的噪声率为$\eta+(1-\eta)\times\xi$。本论文仅考虑对数据集施加 20%的开集标签噪声，所以噪声率为$0.2+0.8\times\xi$。

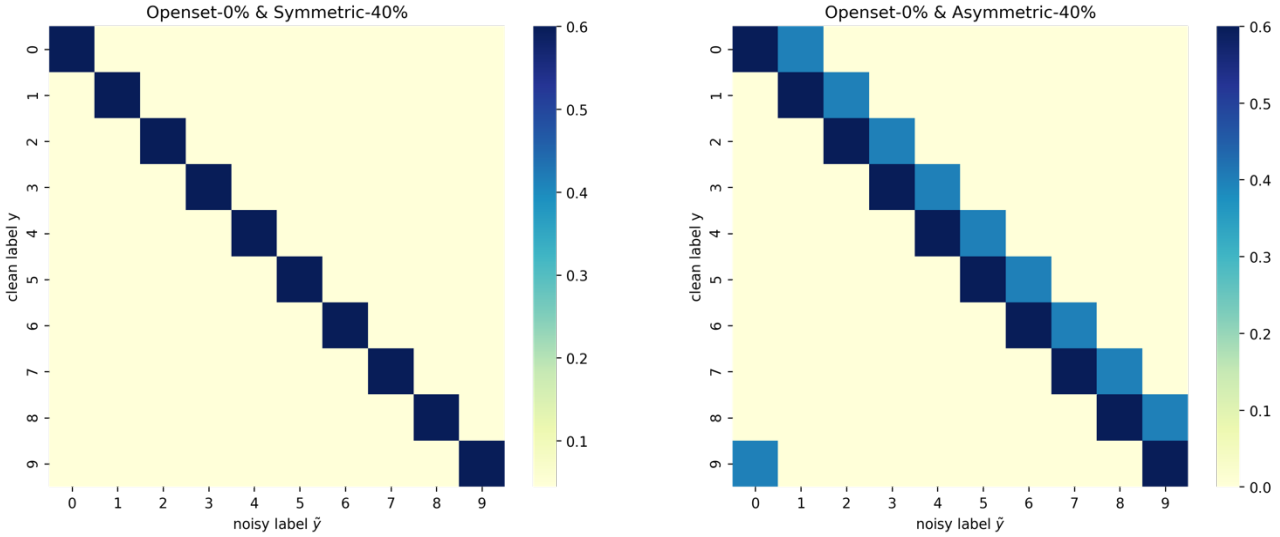

图 4 对称与非对称闭集标签噪声转移矩阵

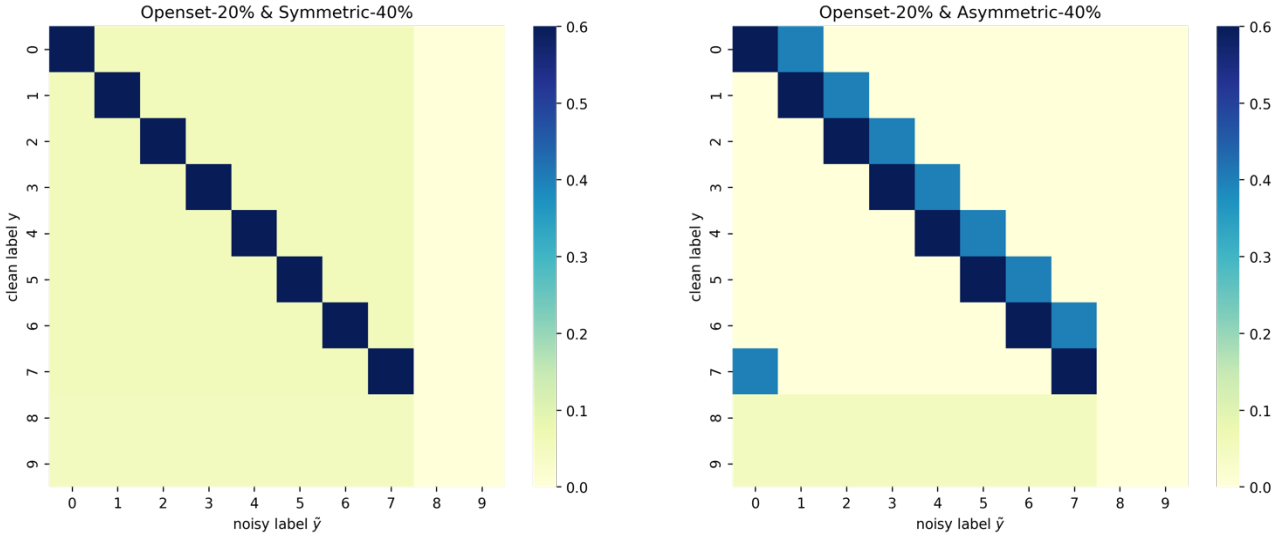

图 5 对称与非对称混合标签噪声转移矩阵

根据文献[1, 44]，在闭集标签噪声和混合标签噪声的设定下，评估准则是在测试集上对已知类别的分类准确率，故在混合开闭集标签噪声设定下，评估方式是在测试集中前$(1-\eta)\cdot C$个类别上的分类准确率。

4.2 数据集与基准模型简介

根据相关文献[44]设定，本论文考虑使用 CIFAR100 数据集作为训练数据集。CIFAR100 数据集中包含 100 个类别，每个类别有 600 张大小为 32×32 的彩色图像，其中 500 张作为训练数据，100 张作为测试数据。该数据集具有层次化的标签，即每张图像都具有细粒度和粗粒度的标签，本论文仅考虑使用 100 个细粒度的类别标签。

对于闭集标签噪声，本文将噪声数据集记为 CIFAR100N。对于混合开闭集标签噪声，本文考虑开集标签噪声率为 20%，故构建的噪声数据集记为 CIFAR80N，其中 80 表示具有 80 个已知类别，N 表示该数据集是经过噪声处理的数据集，将最后 20 个类别作为 OOD 类别。

由于本论文中提出的 S2S 模型属于样本筛选方法，故本文考虑的对比模型为：Standard、Decoupling[48]、Co-teaching[1]、Co-teaching+[4]、JoCoR[49]、Jo-SRC[44]。值得注意的是 Decoupling、Co-teaching、Co-teaching+、JoCoR、Jo-SRC 模型都使用孪生网络共同学习参数，而 S2S 模型仅使用一个网络学习参数，且 Co-teaching、Co-teaching+模型都假设噪声率是已知的，从而比本论文提出的 S2S 模型具有更多的先验信息。

基准模型的介绍分别如下：

- ✓ Standard

  该模型不考虑标签噪声的存在，仅仅使用交叉熵损失函数对网络参数进行学习，其学习率、优化器、网络架构等都与 S2S 模型的设定一致。

- ✓ Decoupling[48]

  Eran Malach 等人提出的 Decoupling 方法同时训练两个不同初始化的孪生网络，他们认为初始权重不同的两个网络应当对干净的样本预测结果一致，预测不一致的样本对分类效果的影响更大，故需要着重利用。具体地，Decoupling 方法在每个批次中，根据两个网络的分歧筛选样本，只使用两个网络预测结果不同的样本进行参数更新。

- ✓ Co-teaching[1]

  Han 等人提出的 Co-teaching 方法同时训练两个架构一致的网络。为了充分利用神经网络在训练早期的记忆能力[24]，即神经网络在早期会首先记住训练集中的简单模式再记住困难模式的能力，Han 等人提出了只利用损失小的样本的更新方法。具体地，在每个周期 Co-teaching 方法用同样批次的数据训练两个网络，每个网络都按一定比率将自身预测的损失小的样本交予另一个网络进行梯度传递和参数更新。

- ✓ Co-teaching+[4]

  Yu 等人提出的 Co-teaching+模型在 Co-teaching 模型基础上融合了 Decoupling 方法的优点，即在 Co-teaching 模型的每个 epoch 中都提前筛选两个网络产生分歧的样本，再基于这些样本和 Co-teaching 模型的小损失更新策略进行模型的训练。

- ✓ JoCoR[49]

  Wei 等人提出了 JoCoR 模型，他们认为一致性才是攻克标签噪声问题的关键。同样的，JoCoR 模型也训练两个神经网络，该模型在交叉熵损失的基础上加上了对称的 KL 散度以迫使两个网络在同一个样本上预测的概率分布相似，并采用了 Co-teaching 中提出的小损

失更新方法。每个网络仅将自身预测的损失小的样本交予另一个网络进行参数更新，即着重考虑预测一致性更好的样本。

- ✓ Jo-SRC[44]

Jo-SRC 模型是专门针对混合开闭集标签噪声问题提出的算法，该模型关注同一个样本在不同网络中的预测一致性，在每个周期分离出干净样本。对于剩余样本，该模型根据预测概率分布的一致性，将预测标签分布一致性更好的样本认为是 ID 样本，其余样本认为是 OOD 样本。该模型通过设计一致性损失函数，迫使 ID 样本通过不同网络预测的概率分布趋于一致，同时迫使 OOD 样本通过不同网络预测的结果不一致。

4.3 实验结果与分析

本小节重点介绍了 S2S 模型在闭集标签噪声和混合的开闭集标签噪声上的实验结果。

在实验设定上，本节中所有的实验都设定训练周期数为 300 个周期，每个周期 512 次迭代，其中筛选干净样本开始于第 10 个周期（预热 10 个周期），筛选分布以内样本开始于第 20 个周期。学习率为 0.02，使用余弦下降策略改变学习率，使之从预热阶段之后，按照余弦曲线调整学习率，使其从 0.02 开始在每次迭代后衰减，最终变为$0.02\times\cos(7\pi/16)$。此外，本模型使用带动量的 SGD 优化器在每个周期完成参数更新，权重衰减系数为 5e-4，并按照其他研究的设定[44]设置动量系数为 0.9。S2S 模型在每次实验中都固定超参数$\mathcal{L}_e=0.1,\mathcal{L}_{con}=1,\mathcal{L}_{id}=1$。实验发现使用固定的权重超参数已经能够获得较好的实验结果。另外，本实验设定$\tau_{clean}=0.8$，$\tau=0.02$。参考 Jo-SRC 模型[44]，本论文中使用与之相同的卷积神经网络 Base(CNN)以进行公平地对比，事实上，任何卷积神经网络架构都能够替代本论文中使用 Base(CNN)网络，包括 Resnet[45]、Wide-Resnet[46]、Preact-Resnet[47]等等。同样地，也可以任意更换分类网络和 OOD 检测网络的架构，不失一般性，本节的实验中仅仅采用多层感知机构建分类网络和 OOD 检测网络。具体地，本节使用的卷积神经网络特征提取网络 Base(CNN)架构如附录 A 所示；分类网络使用两层的感知机，其中隐含层神经元数量是输入维度（图像特征表示的维度）的两倍；对于每个 OOD 检测器，都使用两层的感知机作为网络架构。其次，为了保证算法的稳定性，参考其他研究者的研究[44]，本论文对 CIFAR100 数据集中每个样本的标签进行柔化。对于基准模型，除算法以外所有模型的实验设定均和本文提出的 S2S 模型一致，其中每个模型的超参数都使用原论文中的设定以保证公平性。

4.3.1 闭集噪声上的分类效果

在闭集标签噪声设定下，S2S 模型和基准模型的实验结果如表 1 所示。其中 Standard、Decoupling[48]、Co-teaching[1]、Co-teaching+[4]、JoCoR[49]、Jo-SRC[44]模型的实验结果摘自文

献[44]。

表 1 CIFAR100N 数据集上个各模型最后 10 个周期的平均测试准确率(%)

方法-噪声率	Standard	Decoupling	Co-teaching	Co-teaching+
Symmetric-20%	35.14	33.10	43.73	49.27
Symmetric-50%	16.97	15.25	34.96	40.04
Symmetric-80%	4.41	3.89	15.15	13.44
Asymmetric-40%	27.29	26.11	28.35	33.62
方法-噪声率	Jo-CoR	Jo-SRC	S2S	
Symmetric-20%	53.01	58.15	**58.41**	
Symmetric-50%	43.49	**51.26**	42.51	
Symmetric-80%	15.49	23.80	**30.89**	
Asymmetric-40%	32.70	38.52	**49.60**	

从表 1 中可见 S2S 模型在大多数情况下能够超过所有的基准模型，特别地，在 40%的非对称标签噪声设定的实验中，S2S 模型的测试集准确率相比其他模型高了 10%以上。该模型仅仅在 50%的对称标签噪声情形下得到了稍差的结果，但在该设定下该模型也超过了基准模型 Standard、Decoupling、Co-teaching、Co-teaching+。在该设定下性能下降的原因可能是 Jo-CoR 模型和 Jo-SRC 模型本身使用两个网络进行交互学习，而本论文中提出的 S2S 模型仅训练一个深度网络。该实验证明了 S2S 模型在闭集标签噪声设定下相比于基准模型的优越性。值得注意的是 Decoupling、Co-teaching、Co-teaching+模型在训练过程中都依赖与噪声率的先验信息，比 S2S 模型有更多的应用限制，而在实际中很难获取准确的噪声率信息，故 S2S 模型有更强的实用性。

闭集噪声的分类准确率曲线如图 6（左图）所示。我们容易观察到，在不同标签噪声率的设定下，该模型都没有对标签噪声产生过拟合。即使在 80%的闭集标签噪声的设定下，S2S 模型依然得到了较高的分类准确率，这是该模型能够在极端闭集标签噪声设定下超过基准模型的关键原因。

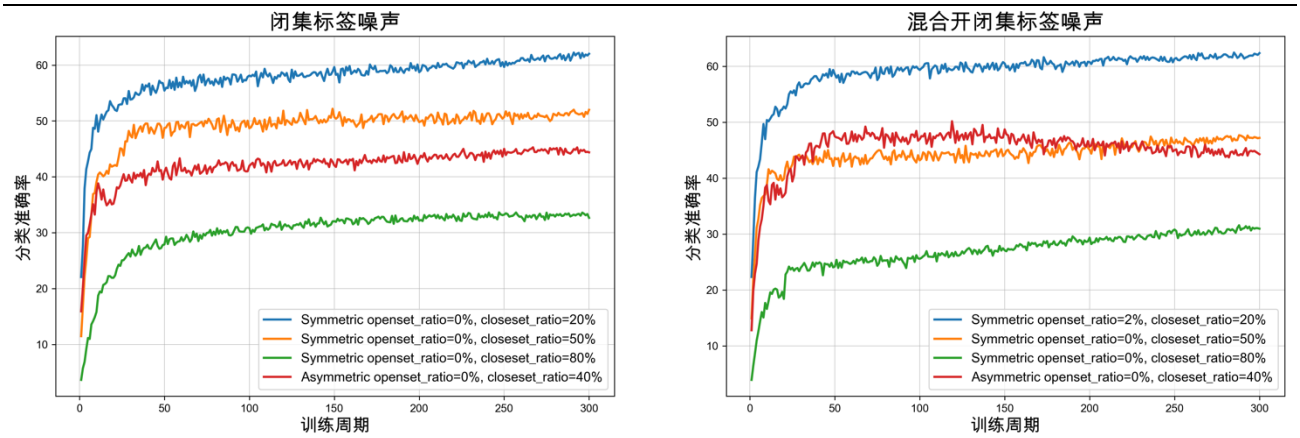

图 6 闭集和混合标签噪声设定下的分类准确率曲线

4.3.2 混合开闭集噪声上的分类效果

S2S 模型和基准模型在 CIFAR80N 上的测试集分类准确率如表 2 所示。从表 2 中可见，S2S 模型在大多数情况下能够超过其他的基准模型，其中在 80%的对称标签噪声设定下，该模型大幅度超过另外五个模型。然而，S2S 模型在 50%对称标签噪声设定下的分类准确率低于 Co-teaching+模型和 Jo-CoR 模型，该现象的产生很大程度上是因为 Co-teaching+模型和 Jo-CoR 模型同时训练两个神经网络，其次，Co-teaching+模型训练中人为提供了噪声率的先验信息，而 S2S 模型训练过程无需噪声率的先验信息。S2S 模型在 CIFAR80N 数据集上的分类准确率曲线如图 6（右图）所示，从图中可见，在大多数情况下 S2S 模型都没有对噪声产生过拟合（仅仅在 40%对称标签噪声设定下产生了轻微的过拟合）。特别地，在 80%的对称标签噪声设定下，该模型没有出现过拟合，而其他基准模型最多只能得到 12.85%的测试集准确率，可见基准模型对噪声数据的过拟合现象比 S2S 模型更严重。

表 2 CIFAR80N 数据集上个各模型最后 10 个周期的平均测试准确率(%)

方法-噪声率	Standard	Decoupling	Co-teaching	Co-teaching+
Symmetric-20%	29.37	43.49	60.38	53.97
Symmetric-50%	13.87	28.22	52.42	46.75
Symmetric-80%	4.20	10.01	16.59	12.29
Asymmetric-40%	22.25	33.74	42.42	43.01
方法-噪声率	Jo-CoR	S2S		
Symmetric-20%	59.99	**60.14**		
Symmetric-50%	**50.61**	45.81		
Symmetric-80%	12.85	**27.30**		
Asymmetric-40%	39.37	**44.74**		

4.4 消融实验

本小节中重点分析了 S2S 模型的优势，大量的实验探索了不同网络架构对模型的影响、各个损失函数对分类准确率的影响以及超参数对模型的影响，证明了 S2S 模型的每个模块都是不可或缺的，同时也为 S2S 模型的使用提供了经验指导。

4.4.1 不同网络架构的影响

本小节着重研究不同的网络架构对实验结果的影响。具体的，S2S 模型中的特征提取网络可以用一些更深层或者更宽的网络代替。故本节中考虑将原始的 S2S 模型中的特征提取网络替换为 Resnet-18[45]、Preact-Resnet-18[46]、Wide-Resnet-28-2[47]，并对比了不同网络架构下的实验结果。Resnet-18 表示使用残差连接构成的 18 层深度神经网络；Preact-Resnet-18 表示采用了不同激活方式的 Resnet-18 网络；Wide-Resnet-28-2 表示 28 层的深度神经网络，其中每个残差连接中特征映射的数量变为 Resnet-18 架构中的两倍。实验结果如表 3 所示（其中 Preact-Resnet-18 和 Wide-Resnet-28-2 没有使用预训练的网络权重），表中 Avg10 表示最后 10 个周期的平均测试集分类准确率，Best 表示训练过程中最优的测试集准确率。观察表 3，我们可以发现在使用参数量更大的网络架构时，分类准确率可以获得约 2%-7%的提升，Resnet-18、Preact-Resnet-18、Wide-Resnet-28-2 三个架构得到的最优分类准确率相当，然而 Wide-Resnet-28-2 架构在最后 10 个周期的平均测试准确率要显著高于另外两者，可见该模型的结果更加稳定且对噪声数据产生的过拟合问题不显著。值得注意的是，Wide-Resnet-28-2 的参数量仅有 147 万，约是 Base(CNN)的三倍，Preact-Resnet-18 架构的参数量是它的 7.6 倍，然而 Wide-Resnet-28-2 依然在更短的训练时间内获得了更高的分类准确率。为了和基准模型在同样的网络设定下进行评估，本文中采用了较为简单的基础网络 Base(CNN)。

此外，本节测试了在使用 ImageNet 预训练的权重初始化网络后 Resnet-18 的分类结果，在表中标注为 Resnet-18(P)，将 Resnet-18(P)和 Resnet-18 进行对比后发现，无论是最后 10 个周期的平均测试集分类准确率还是最优测试集分类准确率都获得了大幅度的提升（约 4%），该现象可能是因为预训练的网络得到的特征映射本身具有一定的鉴别性，从而在网络的预热阶段已经能够获得获得较好的特征表示，有利于干净样本、分布以内样本的筛选。

表 3 CIFAR80N 数据集上不同架构网络的测试结果

网络架构	Base(CNN)	Resnet-18	Preact-Resnet-18	Wide-Resnet-28-2	Resnet-18(P)
参数量(M)	**0.52**	11.17	11.17	1.47	11.17
训练时长(h)	**2.33**	5.35	7.07	5.97	5.11
Avg10(%)	60.14	62.94	62.94	64.25	**67.35**
Best(%)	62.44	65.28	65.63	65.90	**69.64**

4.4.2 算法各组件的重要性分析

本小节着重分析了构成 S2S 模型的各个损失的重要性。具体地，本实验考虑将损失项 $\mathcal{L}_{sup}$，$\mathcal{L}_{e}$,$\mathcal{L}_{con}$,$\mathcal{L}_{id}$分别从原损失函数中移除，分别记为 w/o sup、w/o e、w/o con、w/o id，并重新进行实验，结果如表 4 所示。从中可见在去除任何一个模块之后模型的性能都会下降。特别地，在去除了监督损失项之后，S2S 的分类准确率仅有 22.23%，对数据集的拟合程度远远不够。此外，本文提出的连续性损失项能够使分类精度提升约 2%，而筛选的 ID 样本构成的损失项使得性能提升约 4%。因此每个损失模块都具有足够的重要性。

表 4 CIFAR80N 数据集上个各模型最后 10 个周期的平均测试准确率(%)

方法	w/o sup	w/o e	w/o con	w/o id	S2S
分类准确率	22.23	58.63	58.61	56.63	**60.14**

4.4.3 超参数的敏感性分析

本小节着重探讨了 S2S 模型对于超参数的敏感性。具体地，本小节中$\mathcal{L}_{e}$,$\mathcal{L}_{con}$,$\mathcal{L}_{id}$损失的权重系数分别从集合$\{0.1,0.5,1.0,2.0\}$中取值，随后权重不同的模型在 20%开集标签噪声和 20%闭集标签噪声设定下进行了实验，得到的实验结果如图 7 所示（其中红色标注的为本论文中采取的超参数和对应的分类准确率值）。

从图 7 中可见，ID 损失项的权重对分类准确率的影响较为明显，说明该模型对于 ID 损失项较为敏感。此外，从图 7 中可见，ID 损失项设置为 0.5、1.0、2.0 时实验结果相差约 0.5%，而该项损失设置成 0.1 时分类精度将会下降 2%，故该项损失能够容忍的权重区间可以认定为[0.5,2.0]；连续性正则化损失项的权重对实验结果的影响不大，采用不同的超参数值几乎得到了一样的分类准确率；极小化熵正则化损失在权重为 1.0 时分类准确率降低了约 1.5%，故该项损失应当赋予较低的权重，如 0.1、0.5，从图中可见，将其损失权重设置为 0.1 或 0.5 时该损失对分类准确率的影响较小。

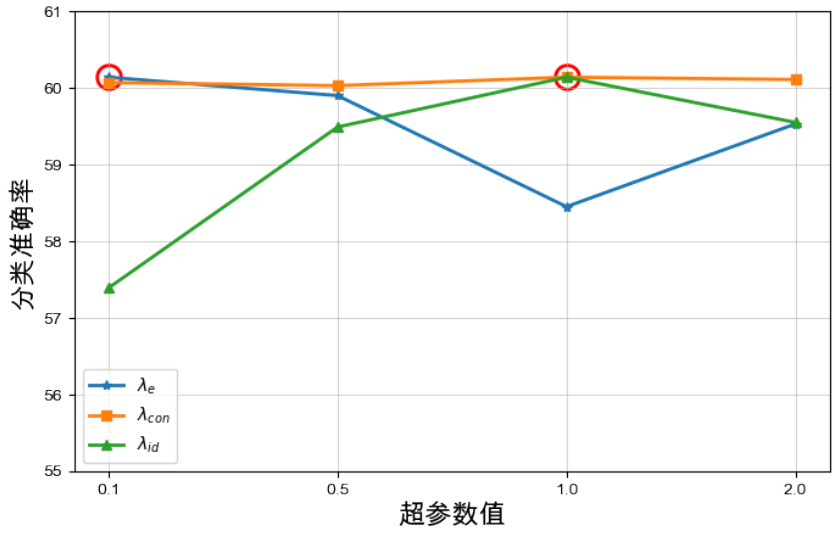

图 7 超参数的敏感性

4.5 改进与未来的研究方向

尽管本论文中提出的 S2S 模型能够在混合开闭集标签噪声学习的实验中获得较高的分类准确率，但该模型仍然有很多不足之处和可以改进的部分。例如，S2S 模型中筛选干净样本的步骤可以在预热网络之后的每个周期前都执行，从而获得动态的干净数据集。此外，可以考虑在预热网络之前添加自监督学习[51]过程或者使用大型数据集中通过自监督学习方法得到的权重来初始化网络。

目前国内外对混合的开闭集标签噪声的研究尚且处于初期，由于问题本身的复杂性，样本特征表示的可鉴别性对分类的影响十分重要，故未来的研究可能考虑通过自监督学习方法得到更可靠的特征表示并基于该特征表示设计算法。另外，目前对开集标签噪声的研究中有学习增广的标签转移矩阵的方法[18,52]，设计有效的算法或优化过程来估计增广的标签转移矩阵具有研究价值。此外，通过原型学习[29]来攻克混合噪声学习也是一个值得研究的方向。

结 论

针对开放环境下的标签噪声学习问题，本文提出了 Selection2Sieve（S2S）方法，首先从样本集中划分出干净样本集；其次采用一对多分布以外检测网络在每个训练周期之初动态划分出分布以内的样本集，将分布以外（OOD）的样本从训练集中剔除。在损失函数层面，该模型采用 one-vs-all 损失函数训练 OOD 检测网络，考虑到 OOD 检测网络所得概率分布的连续性，S2S 模型在总体损失中加入了连续性损失项。实验证明了本文提出的 S2S 方法能够超越 Decoupling、Co-teaching、Co-teaching+、JoCoR、Jo-SRC 等样本筛选方法，在闭集标签噪声和混合的开闭集标签噪声设定下都获得了最好的或次好的的分类准确率。另外，消融实验分析了不同特征提取网络架构下的实验结果，给出了超参数的非敏感区间，同时证明了各个损失函数对实验的结果都是不可或缺的。

然而，该模型依然有可以改进的地方，例如将预热网络阶段替换成自监督学习阶段，或者动态地筛选干净样本集等。对于开放环境下的标签噪声学习问题，研究尚处于初期，未来的研究可能运用自监督学习、原型学习、增广转移矩阵学习等方法提升分类准确率。

致 谢

本次毕业设计的圆满完成离不开我的导师宫辰教授的悉心指导，没有宫教授从 21 年 10 月开始至今的耐心指导，这篇毕业论文将不会成形，同时他也是将我引入科研环境的人生导师，所以他是我首要感谢的人。其次，我要感谢课题组的师兄师姐们，他们与我有共同的目标，给我指引了方向，在研究的内容、实验甚至未来的规划上都给我提供了无可替代的帮助，大家对我这位初入科研的新人充满包容、耐心与期待。此外，我要感谢理学院、计算机学院、设传学院的各位老师、同学对我本科期间的无私帮助，他们的帮助扫清了我学习路上的障碍，使我在本科期间接触了众多新鲜而有趣的领域。最后，我要着重感谢我的父母，在大学四年中给了我经济上、生活上的帮助，圆满的毕业设计是对他们的一种回报，未来，我希望以更高的成就回馈家庭以表达对他们的感激之情。

参 考 文 献

[1] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” in Proc. NeurIPS, 2018, pp. 8527–8537.

[2] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric crossentropy for robust learning with noisy labels. In Proc. ICCV, 2019, pp. 322–330.

[3] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia, “Iterative learning with open-set noisy labels,” in Proc. CVPR, 2018, pp. 8688–8696.

[4] X. Yu, B. Han, J. Yao, G. Niu, I. W. Tsang, and M. Sugiyama, “How does disagreement help generalization against label corruption?” in Proc. ICML, 2019.

[5] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in Proc. NeurIPS, 2018, pp. 8778–8788.

[6] A. Ghosh, H. Kumar, and P. Sastry, “Robust loss functions under label noise for deep neural networks,” in Proc. AAAI, 2017.

[7] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in Proc. NeurIPS, 2018, pp. 8778–8788.

[8] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, “Normalized loss functions for deep learning with noisy labels,” in Proc. ICML, 2020, pp. 6543–6553.

[9] J. Li, R. Socher, and S. C. Hoi, “DivideMix: Learning with noisy labels as semi-supervised learning,” in Proc. ICLR, 2020.

[10] Ragav Sachdeva, Filipe R. Cordeiro, Vasileios Belagiannis, Ian D. Reid, and Gustavo Carneiro. Evidentialmix: Learning with combined open-set and closed-set noisy labels. 2020.

[11] K.-H. Lee, X. He, L. Zhang, and L. Yang, “CleanNet: Transfer learning for scalable image classifier training with label noise,” in Proc. CVPR, 2018, pp. 5447–5456.

[12] Z.-F. Wu, T. Wei, J. Jiang, C. Mao, M. Tang, and Y.-F. Li, “NGC: A unified framework for learning with open-world noisy data,” in Proc. ICCV, 2021, pp. 62–71.

[13] H. Wei, L. Tao, R. Xie, and B. An, “Open-set label noise can improve robustness against inherent label noise,” in Proc. NeurIPS, 2021.

[14] X. Xia, T. Liu, B. Han, C. Gong, N. Wang, Z. Ge, and Y. Chang, “Robust early-learning: Hindering the memorization of noisy labels,” in Proc. ICLR, 2021.

[15] P. Chen, G. Chen, J. Ye, jingwei zhao, and P.-A. Heng, “Noise against noise: stochastic label noise helps combat inherent label noise,” in Proc. ICLR, 2021.

[16] Chen Feng, Georgios Tzimiropoulos, and Ioannis Patras. S3: supervised self-supervised

learningunder label noise.CoRR, abs/2111.11288, 2021. URL https://arxiv.org/abs/2111.11288.

[17] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama.Instance correction for learning with open-set noisy labels.arXiv preprint arXiv:2106.00455, 2021.

[18] X. Xia, T. Liu, B. Han, N. Wang, J. Deng, J. Li, and Y. Mao, “Extended T: Learning with mixed closed-set and open-set noisy labels,” arXiv preprint arXiv:2012.00932, 2020.

[19] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, “Meta-Weight-Net: Learning an explicit mapping for sample weighting,” in Proc. NeurIPS, 2019, pp. 1917–1928.

[20] Pierre Nodet, Vincent Lemaire, Alexis Bondu, and Antoine Cornu´ejols. Contrastive representations for label noise require fine-tuning. CoRR, abs/2108.09154, 2021. URL https://arxiv.org/abs/2108.09154.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012

[22]. Deng, W. Dong, R. Socher, J. Li, Li, K. Li, and L. Fei Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[23] Zhang, Chiyuan, et al. "Understanding deep learning (still) requires rethinking generalization. " Communications of the ACM 64.3 (2021): 107-115.

[24] D. Arpit, S. Jastrz¸ebski, N. Ballas, D. Krueger, E. Bengio, M. Kanwal, T. Maharaj, A. Fischer, A. Courville, and Y. Bengio. A closer look at memorization in deep networks. In ICML, 2017.

[25] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems, 33:20331–20342, 2020.

[26] Erik Englesson and Hossein Azizpour. Consistency regularization can improve robustness to label noise. arXiv preprint arXiv:2110.01242, 2021a.

[27] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. Advances in Neural Information Processing Systems, 34, 2021b.

[28] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. Advances in neural information processing systems, 26, 2013.

[29] Jing Lu, Yunxu Xu, Hao Li, Zhanzhan Cheng, and Yi Niu. Pmal: Open set recognition via robust prototype mining, 2022. URL https://arxiv.org/abs/2203.08569

[30] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606, 2018.

[31] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 2016. doi: 10.48550/ARXIV.1610.02136.

[32] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In International Conference on Learning Representations, 2018.

[33] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020

[34] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum

classifier discrepancy. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9518–9526, 2019

[35] Jongjin Park, Sukmin Yun, Jongheon Jeong, and Jinwoo Shin. Opencos: Contrastive semi-supervised learning for handling open-set unlabeled data, 2021. URL https://openreview.net/forum?id=lJgbDxGhJ4r

[36] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. volume 119 of Proceedings of Machine Learning Research, pp. 3897–3906, 2020.

[37] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In Advances in Neural Information Processing Systems, 2021

[38] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning, 2021. URL https://arxiv.org/abs/2102.03526.

[39] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. Proceedings of the International Conference on Learning Representations (ICLR), 2020

[40] Adam Paszke and Gross et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pp. 8024–8035. 2019.

[41] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.9000–9009, 2021.

[42] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. Advances in neural information processing systems, 33:7260–7271, 2020.

[43] Xiaobo Xia, Tongliang Liu, N. Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? ArXiv, abs/1906.00189, 2019

[44] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-SRC: A contrastive approach for combating noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5192–5201, 2021

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[46] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In BMVC, 2016

[47] He, Kaiming, et al. "Identity mappings in deep residual networks." European conference on computer vision. Springer, Cham, 2016.

[48] Eran Malach and Shai Shalev-Shwartz. Decoupling ”when to update” from ”how to update”. In NIPS, 2017.

[49] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In CVPR, pages 13726–13735, 2020.

[50] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. arXiv preprint arXiv:1909.13719, 2019

[51] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709

[52] Xiaoqing Guo, Jie Liu, Tongliang Liu, and Yixuan Yuan. Simt: Handling open-set noise for domain adaptive semantic segmentation. In CVPR, 2022.

附录 A 网络架构

本论文中使用的特征提取网络架构如表 5 所示。此外，OOD 检测网络和分类网络模块都是用两层的感知机网络，其隐含层神经元数量是输入层的两倍，在 CIFAR80N 和 CIFAR100 数据集上输入特征数量都为 512，OOD 检测网络的输出层神经元数量为 160，分类网络模块的输出层神经元数量为 80。

表 5 特征提取网络架构

Base 网络架构
3 x 3, 64 BN, ReLU
3 x 3, 64 BN, ReLU
2 x 2 Max-Pooling
3 x 3, 128 BN, ReLU
3 x 3, 128 BN, ReLU
2 x 2 Max-Pooling
flatten