中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

硕 士 学 位 论 文

基于标签平滑与表征补偿机制的在线知识迁移

Online Label Smoothing and Representation Compensation Mechanism for Online Knowledge Transfer

论文作者 张长彬 指导教师 程明明教授
申请学位 硕士 培养单位 南开大学
学科专业 计算机技术 研究方向 计算机视觉
答辩委员会主席 评 阅 人

南开大学研究生院
二〇二二年三月

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学 (简称“学校”) 研究生学位论文收藏和利用的管理规定，同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权，同意在以下几方面向学校授权。即：

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》，并作为资料在学校图书馆等场所提供阅览，在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务；

2. 学校可以采用影印、缩印或其他复制手段保存学位论文；学校根据规定向教育部指定的收藏和存档单位提交学位论文；

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	基于标签平滑与表征补偿机制的在线知识迁移				
姓　　名	张长彬	学号	2120190467	答辩日期	
论 文 类 别	博士 ☐　学历硕士 ☐　专业学位硕士 ☑　同等学力硕士 ☐　划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称	计算机技术	
联 系 电 话	15505188669		电子邮箱	zhangchbin@mail.nankai.edu.cn	
通讯地址 (邮编)：300000					
非公开论文编号			备注		

注：本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文，须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	☐ 限制 (⩽2 年) ☐ 秘密 (⩽10 年) ☐ 机密 (⩽20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		**批准日期**	20 年 月 日

南开大学学位评定委员会办公室盖章 (有效)

注：限制 ★2 年 (可少于 2 年); 秘密 ★10 年 (可少于 10 年); 机密 ★20 年 (可少于 20 年)

摘要

知识迁移在单次学习以及连续学习场景下都发挥了重要的作用。但是在这两种场景下的现有知识迁移方法仍存在以下两点局限：一方面在单次学习场景下的知识迁移方法的灵活性低，通常依赖于特定的网络结构或者需要预训练的高精度的教师模型，并且这些方法往往具有比较高的训练代价，通常需要多次前向推理；另一方面在连续学习场景下的知识迁移方法难以在保留旧知识和学习新知识之间进行权衡，并且会使得模型规模逐渐增长。

为了解决在单次学习场景下的知识迁移方法的灵活性问题，本文提出了一种在线标签平滑策略来在线地迁移知识。该方法通过统计模型预测的方式考虑了类别间关系，能够以一种在线的方式将之前训练的模型的知识迁移到当前模型中，可以灵活地应用于各个分类模型的训练中。该方法实现的在线知识迁移方式具有极高的灵活性，不会带来额外的训练代价。另一方面，为了解决在连续学习场景下的知识迁移方法难以在新旧知识之间进行权衡的问题，利用重参数化机制，本文提出了一种表征补偿机制，是一种用于解耦新旧知识的在线知识迁移策略。该策略能够将新旧知识在参数空间解耦，能够使得新旧知识以在线方式自适应地进行融合，并且不会带来额外的推理代价。针对单次学习和连续学习这两种场景中知识迁移存在的问题，本文研究内容及贡献总结如下：

- 为了提升单次学习场景下的知识迁移方法的灵活性，本文提出了基于在线标签平滑的在线知识迁移策略，具有极高的灵活性，为模型的训练引入了类内的约束，从而提高模型表现。
- 为了解决在连续学习场景下的知识迁移方法很难在保留旧知识和学习新知识之间进行权衡的问题，本文提出了基于表征补偿机制的自适应的在线知识迁移策略，将新旧知识进行解耦，从而更好地缓解灾难性遗忘的问题。
- 本文在六个常用的公开分类数据集和三个语义分割数据集上进行了广泛的实验。实验结果表明，本文提出的基于在线标签平滑与表征补偿机制的在线知识迁移策略能够显著提升模型的表现。

关键词： 在线知识迁移；在线标签平滑；表征补偿机制；知识蒸馏；连续学习

Abstract

Knowledge transfer plays an important role in both fully supervised and continual learning scenarios. In these two cases, there are still the following situations. On the one hand, the flexibility of knowledge transfer methods in fully-supervised scenarios is low, and usually depend on a specific network structure or a pre-trained teacher model. These methods often have relatively high training costs with multiple forward propagation. On the other hand, knowledge transfer methods in continual learning scenarios are difficult to balance between retaining old knowledge and learning new knowledge, gradually increasing the size of the model.

To improve the flexibility of knowledge transfer methods in fully-supervised learning scenarios, this paper proposes an online label smoothing strategy to transfer knowledge online. The method considers the relationship between categories and transfers the knowledge of the previously trained model to the current model in an online manner. This method can be flexibly applied to the training of each model, without extra training cost. On the other hand, to solve the problem that the knowledge transfer methods in the continual learning are difficult to balance the old and new knowledge, utilizing re-parameterization mechanism, this paper proposes a representation compensation mechanism, which is used to decouple the old and new knowledge. This strategy can decouple the old and new knowledge in the parameter space, and can fuse the new and old knowledge be fused in an online manner, without extra inference cost. Aiming at the problems of knowledge transfer in the two scenarios of fully-supervised learning and continual learning, the research contents and contributions of this paper are summarized as follows:

- To improve the flexibility of the knowledge transfer method in fully-supervised learning scenario, this paper proposes an online knowledge transfer strategy based on online label smoothing, which has extremely high flexibility and introduces intra-class constraints for model training to improve model performance.
- In continual learning scenario, to solve the problem that knowledge transfer meth-

ods are difficult to balance between retaining old knowledge and learning new knowledge, this paper proposes an adaptive online knowledge transfer strategy based on representation compensation mechanism, which decouples old and new knowledge, so as to better alleviate the problem of catastrophic forgetting.

- This paper conducts extensive experiments on six commonly used publicly available classification datasets and three semantic segmentation datasets. The experimental results show that the online knowledge transfer strategy based on the online label smoothing and representation compensation mechanism proposed in this paper can significantly improve the performance of the model.

Key Words: online knowledge transfer; online label smoothing; representation compensation mechanism; knowledge distillation; continuous learning

目录

摘要 ····· I
Abstract ····· II
第一章 绪论 ····· 1
第一节 研究背景和意义 1
第二节 国内外研究现状 4
1.2.1 单次学习场景下知识迁移研究现状 4
1.2.2 连续学习场景下知识迁移研究现状 5
第三节 本文研究内容 6
第四节 论文章节安排 8
第二章 相关工作 ····· 10
第一节 图像识别 10
第二节 知识迁移与连续学习 11
第三节 语义分割与连续语义分割 12
第四节 知识蒸馏 14
第三章 基于标签平滑的在线知识迁移 ····· 15
第一节 研究动机以及贡献 16
第二节 在线标签平滑方法 18
3.2.1 在线知识迁移 18
3.2.2 在线生成标签 19
3.2.3 引入类内约束 21
第三节 实验结果对比和分析 21
3.3.1 图像分类 21
3.3.2 对噪声标签的鲁棒性分析 24
3.3.3 对对抗攻击的鲁棒性分析 26
3.3.4 与模型集成的关系 28
3.3.5 消融实验 29

第四节 本章小结 31
第四章 基于表征补偿机制的在线知识迁移 32
第一节 研究动机以及贡献 32
第二节 表征补偿机制 35
4.2.1 基于表征补偿机制的知识迁移机制 35
4.2.2 表征补偿机制的有效性分析 36
4.2.3 知识蒸馏机制 37
4.2.4 平均池化与条带池化的比较 39
第三节 实验结果对比和分析 40
4.3.1 实验设置 40
4.3.2 连续类别语义分割 43
4.3.3 连续领域语义分割 44
4.3.4 消融实验 45
第四节 本章小结 49
第五章 总结展望 50
第一节 本文工作总结 50
第二节 未来工作展望 51
参考文献 52
致谢 63
个人简历 64

第一章 绪论

目前深度学习使得人工智能的许多领域取得了突破性的进展，得到了广泛的研究。对于图像分类、语义分割等任务而言，模型的表现与数据规模、模型规模具有直接的关系。然而，由于边缘设备对模型的参数量、计算量和功耗等有许多限制条件，所以如何将大模型的知识迁移给轻量级的模型，具有重要的研究意义。另一方面，考虑到实际应用中模型可处理的类别和领域的可拓展性，如何有效地将原来训练好的模型的知识迁移给当前训练的模型，具有重要的研究价值。本文从单次学习和连续学习两个场景中来探讨知识迁移方法的应用价值与研究价值。

第一节 研究背景和意义

数据驱动的基于神经网络的模型在包括计算机视觉在内的许多领域取得了许多里程碑式的进展，尤其是图像识别[1, 2]、语义分割[3]、目标检测[4]、深度估计[5] 和三维重建[6] 等等。以图像识别为例，不断进步的神经网络模型结构在公开数据集上的表现取得了巨大的进展。早期模型主要是基于卷积的神经网络结构，包括 ResNet[2]、DenseNet[1] 等等。近期，基于注意力机制的 transformer 结构[7] 受到了许多研究者的关注。在目前的实际应用中，边缘设备上倾向于使用轻量级模型，即具有更小参数量、更小计算量和更小复杂度的模型。然而，如果轻量级模型的网络容量有限，在面对大规模数据量的情况下，很难具有很高的表现。于是，目前有许多技术将大模型学习到的知识迁移给小模型，从而用来提升小模型的表现，比如知识蒸馏[8, 9] 等。对于知识蒸馏来说，通常需要使用大模型来在数据集上进行训练，作为教师模型，之后在小模型的训练过程中，将大模型学习到的知识迁移到小模型上。广泛的研究表明，通过知识蒸馏操作，可以提升小模型的表现。

然而，基于“教师-学生”的知识蒸馏机制需要比较高昂的训练代价。首先，这种蒸馏机制首先需要训练出一个表现比较好的教师模型，教师模型的规模往往要比较大，具有较大的模型容量。然而当模型的容量逐渐变大时，在同等规模的数据的情况下，模型会更容易陷入过拟合，从而导致模型的泛化能力变差[10]，

在测试集上具有更差的表现。所以，教师模型往往难以训练，需要考虑如何为模型添加合适的正则化方法，从而提升模型的泛化能力。其次，当获得一个训练好的教师模型之后，在迁移知识给小模型的时候，每张图像会同时通过教师模型和学生模型进行前向传播，从而获得教师模型的若干输出。这样训练时会需要额外的前向传播的计算代价和时间开销。为了减轻知识蒸馏过程中的训练代价，互信息学习[11] 是一种“学生-学生”知识蒸馏框架，不需要一个提前预训练好的教师模型，两个学生模型在知识蒸馏过程中互相约束、互相迁移知识从而提升模型的表现。这样的知识迁移策略给知识蒸馏添加了更多的灵活性。之后进一步，许多研究人员开始关注更加灵活的自知识蒸馏策略，即在整个训练过程中只使用一个需要训练的模型，也不需要提前预训练教师模型。目前一些自知识蒸馏方法通过各种设计来构建出从深层模型到浅层模型的蒸馏，例如BYOT[12] 使用分类网络最后的输出来迁移知识给自身模型的浅层。另外一些自蒸馏方法[13] 通过对数据进行变换，来约束同一训练样本的不同视角在特征空间上的一致性来对模型进行正则化。类似的这样的自蒸馏策略为知识迁移提供了更多的灵活性和便利性。然而，基于结构的自蒸馏方法需要对模型进行特殊的编辑，一定程度上不易用；基于数据的自蒸馏方法在每次训练过程中，都会进行额外的前向传播，增加了训练过程中的计算量。所以，如何进一步提升知识蒸馏方法的灵活性和有效性具有重要的意义。

目前，有许多研究表明，将不同训练轮次的模型进行集成[14]，能够获得更好的效果。所以，一个很直接的思路是将某个训练轮次的模型当作教师模型，将其知识迁移给当前训练轮次的模型，但这样仍然会带来额外的前向传播的计算代价。受到标签平滑的启发，本文设计了一种在时序上进行在线知识迁移的策略，即将不同训练轮次模型之间的知识进行在线知识蒸馏，从而能够带来极大的灵活性和易用性。本文提出的在线知识迁移策略能够在模型进行训练时，将前一轮次的模型的知识在线蒸馏给后一轮次的模型，不需要进行模型结构上的改变，也不需要额外的计算负担。

另一方面，数据驱动的神经网络模型除了面临着怎么得到表现更好的轻量模型外，还需要考虑到模型能够处理的类别或者领域的可拓展性。尽管基于全监督的神经网络模型在许多领域表现出色，但是这些模型往往只能处理固定的包含在训练集中的类别，这对模型的实际应用产生了较大的挑战。比较直接的拓展模型类别的思路是将新增加的数据和原来数据合并后，重新进行训练。但

是，在许多需要频繁更新模型需要处理的类别的场景下，例如商品零售、障碍物识别等，这样做会带来极大的训练代价，并且严重降低生产效率。受到目前在大规模数据集上预训练模型，之后迁移知识到各种下游数据集上的技术的启发，另外一种直接的思路是，直接将在之前的数据上训练好的模型在新增加的数据集上进行微调。然而，许多研究工作指出[15, 16]，这样的微调训练会导致模型发生灾难性遗忘现象，即模型会快速拟合新增加的训练数据，会快速忘记旧知识，在旧的类别或者领域上具有很差的表现。这种如何能够利用知识迁移技术快速有效地能够对模型进行拓展，使其能够同时处理新旧类别的技术被称为连续学习。目前，为了尽量减少对之前旧数据集的依赖，许多连续学习的方法会对重新构建数据集的方法和微调模型的方法进行折中，不会依赖全部的旧数据集的训练图像，而是设置一个有限大小的缓存区[17]，用于存储旧数据集中的若干最重要的图像。这些存储的旧数据集中的图像在缓解模型的灾难性遗忘的现象上起到了重要的作用。由于缓存区大小是有限的，所以如何从旧数据集中筛选出对于维持模型的稳定性的样本是非常重要的。然而，考虑到在许多情况下，由于数据隐私的问题，对于旧数据集的访问是有限制的，所以这种做法变得不可用。所以，如何在给定在之前数据上训练好的模型以及新增数据集的情境下，能够对模型进行拓展，使其能够同时在旧的数据类别和新的数据类别上具有不错的表现，这具有重要的研究意义。

许多研究工作表明，领域迁移[15, 18] 是神经网络面临的一个很大的挑战。目前神经网络模型会假设训练数据和测试数据来源于相同的领域，符合相同的分布，从而能够在测试集上具有较好的泛化性。然而，在实际应用中，模型除了需要经常扩展新的类别，还需要扩展在新的数据领域上的处理能力。例如，研究人员往往将来源于一个城市的数据看作一个领域，对于自动驾驶模型，往往需要不断地增加对新城市的处理能力。所以，除了研究如何拓展模型对新类别的处理能力，如何拓展模型在不同领域上的处理能力也具有重要的研究意义。

在连续学习场景下，许多现有知识迁移方法在对模型进行拓展时，往往采用基于知识蒸馏和模型拓展的方法。基于知识蒸馏的方法[15, 16] 能够有效地将旧模型的知识迁移到现在正在训练的模型，从而缓解对旧知识的遗忘。这些知识蒸馏的方法在设计的时候，要遵循的原则是在模型的稳定性与可塑性之间寻求一种平衡。模型的稳定性是指模型在新数据集上进行训练时，对于旧知识的模式不会发生崩塌。模型的可塑性是指模型除了要维持对于旧知识的识别能力，

还具有学习新知识的能力。当知识蒸馏策略设计的过强，会限制模型学习新知识的能力，降低模型的可塑性，而当知识蒸馏策略设计的太弱的时候，无法保证模型的稳定性，模型对于旧知识的遗忘会更多。然而，模型对于旧知识的记忆和新知识的学习是耦合在一起的，所以另外一些方法通过采用模型拓展[19]的方式来增加模型的学习新知识的能力。这些基于模型拓展的方式往往会在通道上拓宽模型，这样的优点是拓展了模型学习新知识的能力，并且学习新知识会对原有的参数的破坏更小。但是，这种方式在每次进行连续学习时会使得模型规模逐渐变大，实用性会降低。为了避免模型逐渐变大，一些研究工作[20]受到剪枝的启发，在每次学习新知识的时候，使用稀疏性约束来限制神经网络中的某些层激活更少的神经元，从而为后续潜在的连续学习过程保留更多的模型容量。然而，许多研究工作表明，神经网络中需要有一些冗余，才能保证其表现，而稀疏性约束会限制模型的学习能力，损害模型的性能。为了解决解耦新旧知识并保持模型规模，本文提出了基于表征补偿机制的在线知识迁移方式，能够在参数空间上对旧知识的记忆和对新知识的学习解耦。并且在连续学习的过程中具有相同的模型容量，不会限制模型的学习能力，所以具有更好的表现。

第二节 国内外研究现状

1.2.1 单次学习场景下知识迁移研究现状

许多研究工作表明，神经网络模型需要一定的冗余度来具有更好的拟合能力[20]，所以通过增加模型的层数、宽度以及参数量能够获得更高精度的模型。当模型逐渐变大，数据规模不足以支撑模型的训练时，模型会陷入过拟合，具有更差的泛化能力。另一方面，实际应用中的边缘设备都会对模型的大小和计算负载有具体的限制要求，所以如何对模型进行压缩，在保证其相对性能的前提下，使其变得轻量化这一问题具有重大的研究意义。知识蒸馏[8, 21]是目前国内外研究人员通常采用的一种技术方案。这种技术具有很高的灵活性和易用行，不会像模型量化和剪枝一样对硬件具有较高的要求，所以得到了目前国内外研究人员的广泛研究[22]。

目前，知识蒸馏主要分为三种架构，“教师模型-学生模型”架构、“学生模型-学生模型”架构以及自蒸馏架构。对于“教师模型-学生模型”架构，需要一个预训练好的具有较高精度的模型作为“教师模型”，将该模型的知识迁移给轻量级的“学生模型”。这样的架构对训练一个精度较高的教师模型带来了较高的

要求。而“学生模型-学生模型”结构不需要预训练好的教师模型，只需要若干未训练的模型从零开始一起训练，但这种机制会依赖于多种模型结构之间的正则化，所以会同时训练许多模型，具有较高的训练代价。自蒸馏架构解决了上述两种蒸馏架构的问题，目前国内外研究人员主要通过设计特定的模型结构从而形成从深层到浅层的知识蒸馏[12]，或者对训练图像进行不同的数据增强得到相同样本的不同视图[13]，从而通过约束两个视图之间的一致性来对模型进行正则化。不过，现有的自蒸馏方案仍然存在两个限制：一是依赖于特定的模型结构的方法的灵活性，二是依赖于同一样本的不同视图的方法具有较高的训练成本，每次训练要进行多次的前向传播。

1.2.2 连续学习场景下知识迁移研究现状

数据驱动的神经网络模型假设训练数据和测试数据具有相同的分布，基于单次学习的深度学习模型则具有更强的先验假设，即要求模型在进行测试时要具有与训练集相同的类别或来源于相同的领域。然而在实际应用中经常需要处理新的类别或者新的领域的数据，需要对现在的模型进行拓展，使其能够处理新的类别或者领域。一种比较简单的解决思路是将新增加的训练集和之前的数据集合并重构数据集，再重新训练一个模型，使其能够同时处理新增加的类别和旧的类别。但是面对需要频繁更新模型所要处理的类别的场景，这种方法会严重降低效率。另外一种思路是将之前训练好的模型直接在新增加的数据集上进行微调，这样具有非常高的效率，最低的训练代价以及极高的灵活性。但是这样模型会发生灾难性遗忘[23] 的问题，即模型会快速拟合到新增加的数据上，而丧失对旧类别或旧领域的判别能力。所以，目前国内外的研究人员会着眼于探讨连续学习这种综合考量模型精度和训练效率的机器学习方法。连续学习主要关注如何在拓展模型处理新的类别或者领域时，能够有效地将原来训练好的模型的知识迁移过来，而尽可能减少对之前使用的训练数据的依赖。出于简单设置的考虑，目前许多研究人员会使用一个有限大小的缓存区来保存若干训练样本，使用这些样本来复习旧知识，仍然依赖于对旧数据的访问。出于数据隐私政策等原因，数据的访问受到了限制，最近越来越多的研究者开始关注一个最具有挑战性的情景，即完全不使用旧的数据集，也不需要缓存区来存储任何旧数据集中的样本。

为了解决这一挑战性的设置，目前国内外的许多研究工作主要分为两个方面，一是利用知识蒸馏[15, 24] 来增强模型的稳定性，缓解其对旧知识的遗忘，二

是对模型进行拓展[19]，使其具有更大的容量来学习新的知识，尽可能减少学习新知识给模型带来的对旧知识的破坏。作为一种强大的知识迁移手段，知识蒸馏能够有效地减少对旧知识的遗忘，但是对于旧知识的记忆和对于新知识的学习是耦合在一起的。具体来说，网络的稳定性和可塑性都要依靠同一套参数，当知识蒸馏的约束比较强时，网络会着重于维持模型对旧知识的记忆，而学习新知识的容量会受到限制。当知识蒸馏对教师模型和学生模型之间的一致性约束比较弱时，模型具有更强的学习新知识的能力，但是会弱化对旧知识的记忆。所以，基于知识蒸馏的连续学习策略需要在记忆旧知识和学习新知识两者之间进行权衡以维持其脆弱的平衡状态。另外一种基于拓展模型的策略，会在每次增加新类别或者新领域时，对模型在宽度上进行拓展，使其具有更大的模型容量来学习新的知识。

第三节　本文研究内容

知识迁移具有重要的应用价值和研究意义，本文主要围绕两种主要场景，单次学习和连续学习场景，来研究更高效率、更高精度的标签平滑方法。如图 1.1所示，本文展示了主要的研究内容、在两种场景下面临的挑战、提出的方法以及实验结果和产生的影响。

首先，在单次学习场景下，为了解决现有知识蒸馏方法需要预训练高精度的教师模型且训练效率低的问题，本文提出了一种在线标签平滑的方法。不同于现有的手工设计的平滑方法，本文提出的方法通过将同类预测正确的样本的预测概率进行累积得到相应类别的软标签。该软标签会被用来监督模型的训练。为了不增加额外的训练代价，本文使用提出了一种在线训练的方式。具体来说，该方法会使用一个极小的缓存区用来存储每个类别对应的软标签，同时使用该软标签和硬标签对模型进行监督。为了得到这个软标签，该方法设置了一个固定大小的训练迭代次数的滑动窗口，在某个窗口内对所有预测正确的训练样本的预测概率进行累积。在窗口结束时，累积的预测概率会进行归一化，从而得到一个可以用作监督信号的概率向量，之后在新的窗口内使用该概率向量来监督模型的学习。同时在新的窗口内，该方法会累积新的概率向量，从而用于下一个训练窗口内的模型的监督信号。这一方法的动机来源于现有的标签平滑方法不具有类别之间的关系，而知识蒸馏得益于隐藏在模型预测的概率向量之间的类别关系。这一方法是一种基于训练样本的正则化方法，实现了在时序上将不同

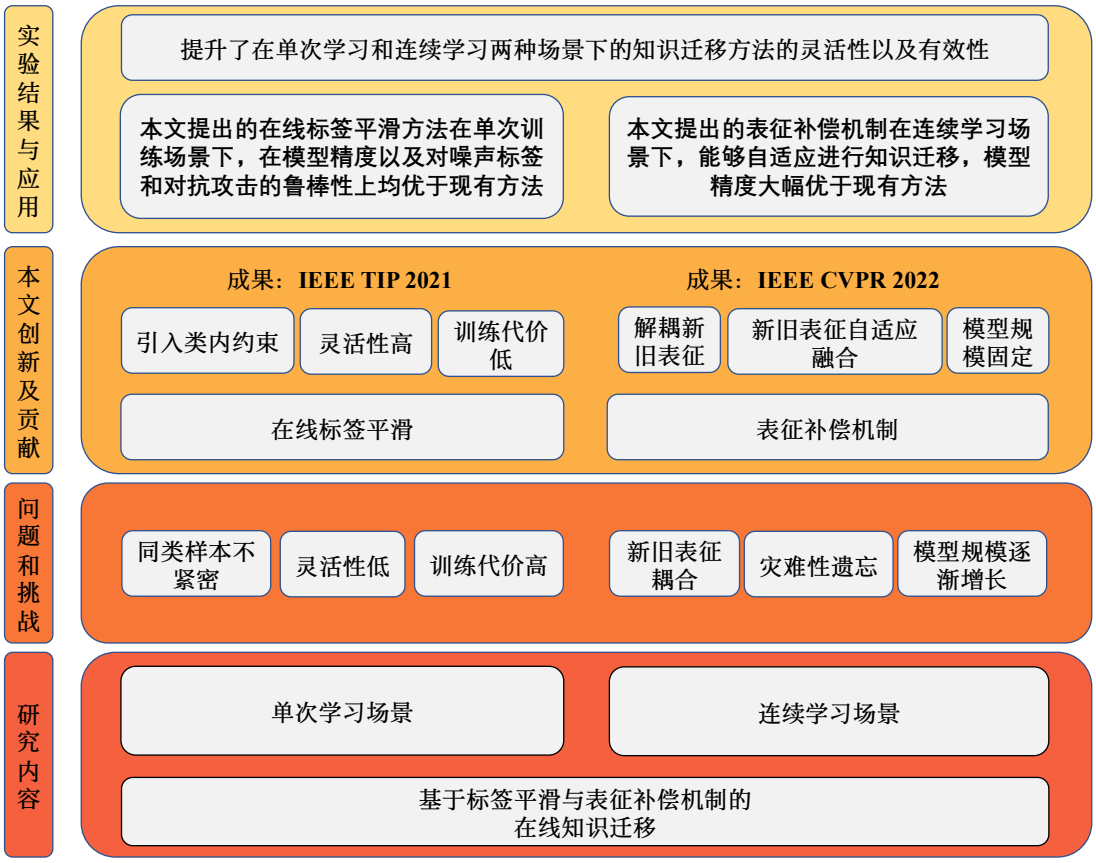


图 1.1 本文研究内容

训练轮次的模型的知识迁移到当前模型。该方法考虑到了不同实例之间的关系，会约束同类之间的样本之间更加紧密，能够使得模型在许多场景下获得更好的鲁棒性，比如噪声标签以及对抗攻击等。本文提出的这种在线标签平滑的策略，既具有现有标签平滑方法的高度的灵活性，又考虑到了不同类别之间的相关性，从而对模型引入了类内的一致性约束，能够使得同类样本距离更加紧密。

另一方面，在连续学习的场景中，针对现有只利用知识蒸馏的方法无法将对旧知识的记忆和对新知识的学习解耦的问题，本文提出了表征补偿机制将参数进行解耦，从而缓解对旧知识的遗忘和增强对新知识的学习能力，实现了将旧知识自适应迁移到新模型。具体来说，该方法将模型中的单个卷积结构拓展为并行的线性相加的结构，从而保证两个卷积可以进行等价融合为一个卷积层。在每次拓展模型时，会首先将训练好的模型的并行卷积层合并为一个，将其参数初始化给现在要训练的模型的其中一个分支，并将其固定，用于保持对旧知识的记忆。而另外一个分支仍然是可训练的，主要用于学习新的知识。在最终

优化目标的约束下，可训练的分支不仅要负责学习新的知识，还要与旧的知识进行自适应。这一自适应的机制能够将训练好的模型中的旧知识迁移到当前模型中，通过自适应的方式实现了在保持旧知识和学习新知识之间寻求平衡，从而更好地缓解灾难性遗忘的问题。

第四节　论文章节安排

本文主要基于在线标签平滑和表征补偿机制两种方式来构建在线知识迁移方法，来提升目前知识蒸馏方法的灵活性以及解决在连续学习过程中难以平衡新旧知识的问题。首先针对目前标签平滑方法没有考虑到类别之间的关系的问题，本文提出了一种基于标签在线累积的方式得到用于监督模型训练的软标签。这种数据相关的标签既考虑到了隐藏在模型预测概率中的知识，又具有知识蒸馏方法难以比拟的灵活性，可以有效地将之前训练轮次的模型的知识迁移到当前模型。另一方面，知识蒸馏在连续学习中得到了广泛的应用，但是其难以在保持旧知识和学习新知识之间寻求一个良好的平衡。针对这一问题，本文设计了一种基于表征补偿机制的训练策略，将网络中的单个卷积层拓展成为并行的线性相加的两个卷积层，其中一个分支是不可训练的，用于记忆旧的模式，另外一个分支是可训练的，用于学习新的知识。在进行连续学习的过程中，原有的训练好的模型的两个分支会被等价合并为一个卷积层，并将其参数初始化给当前模型的的不可训练分支，包含了旧模型的所有知识，而另外一个分支仍然是可训练的，所以整个模型仍然具有相同的可优化参数。这样的一个解耦机制能够通过特征自适应的方式来缓解模型的灾难性遗忘的问题以及学习新知识的能力不足的问题。为了验证提出的两种在线知识迁移策略的有效性，本文分别在分类和分割数据集上进行了实验并详细分析了两种策略。本文共分为五章，其主要内容为：

第一章为绪论，首先介绍了在单次学习和连续学习两种场景下来设计在线知识迁移策略的研究背景和意义，之后讨论了国内外的研究现状，最后对本文的主要研究内容进行了概括和介绍。

第二章为相关工作，由于本文是在图像识别和连续语义分割上进行实验，所以分别讨论了图像识别、语义分割、连续学习和知识蒸馏的相关研究工作。

第三章为基于标签平滑的在线知识迁移，在探讨单次学习场景下的在线知识迁移方法。针对目前标签平滑方法没有考虑类别间关系的问题，提出了一种

数据相关的在线标签平滑方式，能够将之前训练轮次的模型的知识迁移给当前模型。本文首先在该章的第一小节中介绍了研究动机和贡献，即现有的标签平滑方法是基于手工设计的平滑规则，没有考虑到类别之间的关系，而知识蒸馏则是得益于其模型预测概率中隐藏的类别之间的关系。为了解决这一问题，在该章第二小节本文提出了一种新的在线标签平滑方法。之后在第三小节中，本文在图像分类上进行了广泛的实验。本文分析这种数据相关的在线标签平滑方法能够使得同类样本之间更加紧密，所以之后分析了该方法对噪声标签以及对抗攻击的鲁棒性。由于在线标签平滑方法起到了将之前训练轮次的模型知识迁移到当前模型，所以在该小节中进一步讨论了与模型集成的关系。最后，该小节进行了消融实验，来进一步验证策略的有效性。

第四章为基于表征补偿机制的在线知识迁移，在探讨连续学习场景下的在线知识迁移策略。目前连续学习中依靠优化目标来同时保持模型的稳定性和增强模型的可塑性，但是仅依靠优化目标很难在这两者之间找到一个较好的平衡。针对这一挑战，该章提出了一种基于重参数化策略的连续学习策略，能够以自适应的方式，将旧知识迁移到现有模型中。本文首先在该章的第一小节中介绍了研究动机与贡献，即解耦了模型对旧知识的记忆和模型对新知识的学习，使其以一种自适应的方式进行知识迁移。在第二小节中详细介绍了本文提出的基于表征补偿机制的方法。之后，在第三小节中在连续类别语义分割和连续领域语义分割两个任务上进行了实验和分析。最后，该章进行了消融实验来进一步分析所提出方法的有效性，并且对本章内容进行了总结。

第五章为总结展望，该章节对本文的工作进行了总结。单次学习场景与连续学习场景是应用和研究中最重要的两个场景，所以这两个场景下的知识迁移策略在实际应用中具有重大的价值。本文在最后对未来的工作进行了展望。

第二章 相关工作

第一节 图像识别

正则化技术。使用硬标签（将 1 分配给目标类别，将 0 分配给非目标类别）训练深度神经网络通常会导致模型过于自信。增强标签是一种简单而有效的方法，可以缓解过拟合问题并提高深度神经网络的准确性和鲁棒性。Bootstrapping[25] 提供了两个方法 Bootsoft 和 bootshard，分别使用预测的分布和预测的类别来平滑硬标签。Xie 等人[26] 随机扰动样本批的一些标签，以此对网络正则化。为了进一步避免训练模型对某些特定样本的过度拟合，Dubey 等人[27] 在训练中对属于不同类别的样本的输出概率向量添加了混淆，以便模型可以学习到特定样本的微小区别的特征。Li 等人[28] 使用两个网络将图像和标签嵌入到一个高纬空间中，并通过这些嵌入之间的距离对网络进行正则化。Christian 等人[29] 利用软标签进行训练，其中软标签是通过在硬标签和标签上的均匀分布之间取平均值来生成的。继 AET[30, 31] 和 AVT[32] 之后，Wang 等人[33] 提出了一个结合半监督和自监督训练的框架 EnAET[33]，它通过预测非空间和空间变换参数来学习特征表达。EnAET[33] 通过积累多个样本的预测来获得软标签的。具体来说，这些样本是通过不同的变换函数对同一样本的增广视图进行累积，从而得到每个样本的软标签。这种利用数据增强进行的一致性约束也经常用于自蒸馏。

噪声标签。由于人类注释中存在的不正确性，当前数据集中的噪声标签是不可避免的，即数据集中某些样本的标签是错误的。为了解决这个问题，许多研究者从模型[34, 35]，数据[36, 37] 和训练策略[38–40] 方面来探索解决这个问题的方法。一个典型的想法[41–43] 是对不同的样本加权，以减少噪声样本对训练的影响。Ren 等人[41] 在验证集上进行验证，以动态调整样本批中每个样本的权重。MetaWeightNet[42] 也利用验证集，通过多层感知器学习样本的权重。此外，一些研究者从优化[44, 45] 的角度来解决这个问题。Wang 等人[44] 通过用对称交叉熵函数替换标准交叉熵函数，提高了对噪声标签的鲁棒性。Arazo 等人[46] 观察到，在训练的早期阶段，噪声样本通常比干净样本有更高的损失。基于这一观察，他们建议使用 β 混合模型来表示干净样本和噪声样本，并采用该模型来提

供噪声样本的估计的真实类别。另一种想法[47, 48] 是只使用正确的标签来训练网络。PENCIL[49] 提出了一个新的框架，可以同时学习正确的标签和模型的权重。这种方法为每个样本保留了一个可学习的标签。Han 等人[40] 设计了标签校正阶段，并迭代地执行了训练阶段和标签校正阶段。他们为每个类别做了多个原型，并重新定义了所有样本的标签。

重参数化机制。结构化重参数机制首先在 ACNet[50] 中被提出，该工作利用线性结构的可加性来对模型结构进行简化，从而节省模型推理阶段的参数量和时间。在训练过程中，ACNet[50] 设计了并行的几个不同形状的卷积核结构来代替原来的单个卷积层，在推理阶段，这些不同形状的卷积核会被等价地合并为单个卷积层。于是，在推理阶段，模型会保持与普通模型相同的结构，不会带来额外的推理代价和计算代价。之后，RepVGG[51] 对这一结构化重参数策略进行了拓展，设计了不同的卷积核结构。本文受到结构化重参数的启发，设计了一种进行连续学习的策略。

第二节　知识迁移与连续学习

在大量数据的驱动下，基于全监督的深度学习模型取得了巨大的应用进展。然而在实际应用中，基于单次学习的深度模型只能处理预先设置的固定类别，这严重降低了系统的可拓展性。所以，如何有效地拓展模型，使其能够处理新的类别，并且仍然保持对之前类别的处理能力，就变得极其重要。连续学习主要关注的问题是缓解训练过程中的灾难性遗忘问题，同时保持对新学习的类别的判别能力。为了解决这一问题，许多工作[52–56] 使用基于回放机制的技术来复习旧类别的知识。之前许多方法通过不同的形式来对旧类别的知识的复习和回顾，例如训练样本[53, 54, 57–60]，特征原型[61–63] 和生成模型[64] 等等。尽管这些基于回放的方法通常能够实现非常高的表现，但是往往需要存储数据的代价和访问数据的权限。所以一些研究者在探索更加具有挑战性的设置，不再使用额外的存储空间来存储样本和进行样本的回放训练。在这种设置下，需要设计有效的知识迁移的策略，来保留模型学习到的旧知识。许多方法探索使用正则化技术来保留学习到的旧类的知识，比如使用知识蒸馏[24, 65–70], 对抗训练[71, 72]，以及显式的正则化[73–80]。一些方法会关注在不断拓展网络要学习的数据所带来的网络模型容量不够的问题。这些方法[19, 81–85] 会在学习新类别的时候拓展网络模型的结构。

另外一些方法[20, 86] 会在网络模型参数上施加稀疏化约束，目的在于使得每次连续学习的步骤都激活尽量少的神经元，这样就能留出来尽量多的神经元给后续的学习使用。但是，这样的一个稀疏性约束减少了网络的冗余性，会对每次学习的时候，限制了模型的学习容量。一些工作提出通过将自监督学习用于特征提取器上来使其学到更好和更加通用的表征[58, 87]。连续学习过程中另外可能会出现的一个问题是类别不平衡的问题，这是因为每个连续学习步骤中新增加的类别的数据规模很难保证与之前训练的旧的类别的规模相似。所以，一些方法主要关注在连续学习过程中出现的类别不平衡[88–92] 的问题。

第三节 语义分割与连续语义分割

语义分割。语义分割是一个非常重要的基础任务，在许多下游任务中具有重要的应用，比如用于自动驾驶、场景理解、三维重建和计算摄影等。早期的分割方法主要关注于建模图像的上下文关系[93–95]。后来一些研究工作表明，由于图像中物体尺寸大小不一，所以特征的多尺度信息对于语义分割是非常重要的。所以目前很多方法[96–103] 将注意力放在多尺度特征融合上。这些方法着眼于设计不同的网络结构设计来更有效地融合不同尺度的特征，从而获得更高质量的分割结果。注意力机制在深度学习中获得了极大的关注，许多工作也在设计语义分割上的更有效的注意力机制。最早的显示注意力机制是 Non-local[104] 中提出的，之后许多方法[105–111] 设计不同的注意力机制来建立图像上下文的关系。注意力机制能够起到两种作用，一是注意力图对于不同样本是不同的，所以能够增大网络的容量，具有更强的学习能力，二是有些工作独特的设计能够增大模型的感受野范围，使其具有更强的处理大物体的能力。另外还有一些研究工作着眼于将网络拓宽，在不同的宽度上使用不同的感受野范围来处理特征，之后将不同感受野的特征进行融合，从而获得具有更丰富表征能力的特征。这样做能够使得模型对不同尺度物体更加鲁棒。最近，基于 transformer 的模型结构[112–117] 在语义分割中发挥了重要的作用，获得了很大的性能提升。一些基于 transformer 的模型也会着眼于设计多尺度融合[7, 118–120] 的技术。另一些模型[121, 122] 会关注在设计网络的上下文特征聚合。

连续语义分割。连续语义分割仍然是一个亟待解决的问题, 这一任务主要关注语义分割任务中出现的灾难性遗忘问题。在这一领域中，例如，一些工作[123, 124] 探索使用基于回放技术的算法来复习回顾旧知识。回放技术指的是，

在每一个连续学习的训练步骤中，都会设置一个固定大小的存储空间（比如最多可以存储 100 个样本），之后设计策略挑选最具有代表性的样本，存储在固定空间中。在之后的连续学习步骤中使用这些存储的样本来复习回顾旧知识，以避免模型学习到的旧知识的模式发生崩塌，导致发生灾难性遗忘的问题。一些方法[124] 从对抗样本中借鉴思路，利用教师模型对输入的随机噪声梯度下降，从而生成一些符合教师模型学习的旧知识的模式的样本。利用这些样本作为回放的样本来增强模型对旧知识的记忆。而另外一种设置是完全不需要存储空间来存储之前的训练样本，这种设置更加具有挑战性，只能利用当前新增加的训练数据和上一步训练好的模型（称为教师模型）来设计算法策略解决连续学习过程中存在的问题。在连续类别语义分割的场景下，语义分割的标签中只能包含当前要学习的类别的标签，之前学习的旧类别都会被标记为背景类别。所以，当前学习步骤中的训练图像中可能会包含一些旧类别的区域被标记为背景，MiB[16] 为了解决这个优化目标的问题，提出了一个无偏知识蒸馏的优化目标，来挖掘背景区域中的旧类别。为了学习新类别，MiB[16] 认为目前训练的模型在旧类别上的概率应该与背景类合并为当前图像中标注的背景类别，从而计算交叉熵分类损失。同时，为了避免模型学习到的旧知识的模式崩塌，MiB[16] 将目前模型的新类别的预测概率与背景的预测概率合并，之后和教师模型的输出之间计算交叉熵分类损失。PLOP[15] 提出在模型的隐藏层之间使用知识蒸馏策略，从而优化模型对旧知识的记忆。在连续学习场景下，不能直接按照普通的知识蒸馏策略对模型进行蒸馏，如果对模型采用像素级知识蒸馏，能够维持模型的稳定性，但会削弱模型学习新知识的能力，降低模型的可塑性。所以，PLOP[15] 设计了一种基于条带池化的蒸馏方式，通过条带池化操作将像素之间的约束减小，从而能够给模型学习新知识的空间，从而增强模型的可塑性。SDR[125] 在特征空间，利用特征原型匹配的方式进行一致性约束，减少对旧知识的遗忘。其他方法会利用高维特征、自训练和模型自适应等技术[79, 126, 127] 来帮助连续学习的训练。更多地，目前除了连续类别语义分割，连续领域语义分割也是一个重要的设置。与连续类别语义分割关注的点不同，连续领域语义分割会假设所有领域中的类别都是相同的，只具有领域的差异。举例来说，目前常用的设置是将来源于不同城市的数据看作不同的领域，但是数据中的类别都是相同的，在连续学习过程中，每次会让模型学习新的领域的知识，但需要同时保证对旧领域的处理能力。PLOP[15] 使用相同的知识蒸馏策略来处理连续领域语义分割的问题。

与之前的方法不同，本文提出的基于表征补偿机制的方法来动态地拓展网络的学习能力，将对旧类和新类别的学习进行解耦，从而能在网络的稳定性和可塑性上取得较好的平衡。

第四节 知识蒸馏

知识蒸馏[8, 128, 129]，是一种流行的模型压缩方法，可以显著提高轻量级网络的性能。知识蒸馏在许多任务[21, 130–132] 中被广泛应用。Hinton 等人[8] 表明，知识蒸馏的成功是由于隐藏在模型预测中的类别之间的关系。这表明深度神经网络可以发现隐藏在预测中的不同类别[8, 129] 之间的相似性。受知识蒸馏的启发，一些工作[12, 13, 129] 利用自蒸馏策略来提高分类准确性。BYOT[12] 设计了一种基于自知识蒸馏的网络结构，将知识从深层蒸馏到浅层。Xu 等人[13] 采用了一种基于数据的自知识蒸馏方法，并约束增强样本的输出与原始样本保持一致。Furlanello 等人[129] 提出将教师模型的知识蒸馏到具有相同架构的学生模型。同时，Tommaso 等人[129] 也证明了软标签中类别间相似性的重要性。本文的工作受到知识蒸馏的启发，目的在于找到合理的类别间的相似性。知识蒸馏和本文的方法都使用网络的输出概率向量作为软标签，并受益于模型预测概率中隐藏的相似性[8, 129]。但是本文提出的在线标签平滑方法和知识蒸馏有很多不同之处。在没有任何教师模型的情况下，与知识蒸馏相比，该方法可以节省训练开销，即不需要额外的前向传播。此外，该方法可以适用于任何的网络结构，具有高度灵活性和极低训练代价。

第三章　基于标签平滑的在线知识迁移

数据驱动的卷积神经网络在图像分类中发挥了重要的作用。实际应用中，在边缘设备上部署的模型对模型大小和效率有很强的限制条件。因此，如何训练出来一个具有高精度的小模型得到了研究人员的广泛关注。包括像剪枝、网络架构搜索、量化和知识迁移在内的训练小模型的相关技术得到了广泛的研究。其中，以知识蒸馏为代表的知识迁移，对边缘设备的硬件条件没有限制要求，也不需要对小模型结构进行改变。通过知识蒸馏，能够有效地将具有高表现的大模型的知识迁移给当前的小模型，从而提高小模型的表现。

目前基于“教师-学生”架构的知识蒸馏方法会要求具备一个预训练好的具有较高精度的模型作为“教师模型”，而训练较大规模的“教师模型”也存在挑战。相比而言，自蒸馏具有较高的灵活性，不需要“教师模型”，得到了许多研究人员的青睐。然而，目前自蒸馏方法主要利用模型结构来构建从深层到浅层的蒸馏，或是利用样本之间的一致性来约束模型的学习效果。这样的两种方式仍然给模型的训练带来了一些挑战，比如必须要依赖于特定的模型结构设计，并不能对所有的模型结构都能起到作用。而标签平滑则具有最高的灵活性，不会带来任何的额外代价。标签平滑可以被理解成使用一个输出为均匀分布的固定的教师模型来对学生模型进行蒸馏。尽管该方法可以有效地缓解模型的过拟合现象，但是由于教师模型是固定的均匀分布，所以该方法没有充分利用到教师模型中的有效信息，没有进行有效的知识迁移。

该章节从设计标签平滑中包含更多可迁移知识的教师模型的角度出发，提出了一种使用在线累积方法得到教师模型的方法。这一方法考虑到了不同类别标签之间的关系，为每个样本施加了类内的约束，使得同类样本之间更加紧密。通过这一方法，能够将之前训练好的模型的知识迁移给当前模型，从而提升当前模型的表现。之后，为了验证所提出的在线标签平滑方法的有效性，该章节在公开的图像分类数据集上进行了广泛的实验，相比于传统的标签平滑方法，均取得了显著的提升。

本章节的内容安排如下：首先阐述了研究动机以及贡献，之后详细阐述了在线标签平滑方法，最后在一些公开图像分类数据集上进行实验，并对实验结

果进行分析和讨论。

第一节 研究动机以及贡献

深度神经网络[1, 2, 133–137] 在图像识别方面取得了显著的成绩[138, 139]。然而，基于全监督的深度神经网络很容易陷入过拟合，对样本的预测往往过于自信（即以很高的置信度对样本给出预测结果），极大地影响了其对测试样本的泛化能力。最近，研究人员提出了许多正则化方法来克服模型对训练集分布的过拟合问题，包括 Label Smoothing[29]，Bootstrap[25]，CutOut[140]，MixUp[141]，DropBlock[142] 和 ShakeDrop[143] 等等。这些方法试图从数据增强[140, 141]，模型设计[142, 143]，或标签转换[25, 29, 144] 的角度来缓解模型的过拟合现象。

标签平滑是 Christian 等人首次提出的[29]，目的在于为可学习的分类模型提供正则化。该方法不再利用硬标签进行训练（图 3.1(a)所示），而是通过在硬标签和一个均匀分布之间取平均值来作为用于监督模型训练的软标签（如图 3.1(b)所示）。虽然这种标签平滑方式可以提供很强的正则化并防止模型对预测结果过于置信，但它通过给非目标类别分配固定的均匀分布，忽略掉了不同类别之间的关系。例如，如图 3.1(c)所示，“cat”类别具有与“dog”类别更高的相似性。因此，本文认为应该充分考虑非目标类别与所给图像类别的相似性，来确定平滑标签中的非目标类别的概率分布。而对每个非目标类别不加以区分地进行处理会削弱标签平滑的能力，限制模型的性能。Hinton 等人[8] 表明，模型的预测为揭示不同类别之间的隐含关系提供了一种很有前景的方法，而知识蒸馏的有效性则得益于模型的预测中隐含的类别关系。受该观点的启发，本文提出了一个简单而有效的方法以代替标签平滑，考虑了不同类别之间的关系来生成更可靠的软标签。具体来说，我们为每个类别保存一个动态标签分布，可以在训练过程中更新。保存的标签分布在每次训练迭代中都是不断变化的，用来监督模型直到模型收敛。该方法利用了中间模型预测的统计数据，较好地建模了目标类别和非目标类别之间的关系。从图 3.1(c)可以看出，当标签为“猫”时，本文提出的标签平滑方法使动物类别比那些非动物类别更加置信。

目前知识蒸馏方法得到了广泛的应用，基于“教师-学生”架构的蒸馏方式的灵活性较低，需要提前预训练教师模型。所以不需要预训练教师模型的自蒸馏方法得到了广泛的关注。目前广泛使用的两种自蒸馏方法一是利用特定的模型结构来设计从深层到浅层的蒸馏[12]，二是利用同一训练样本的不同数据增强

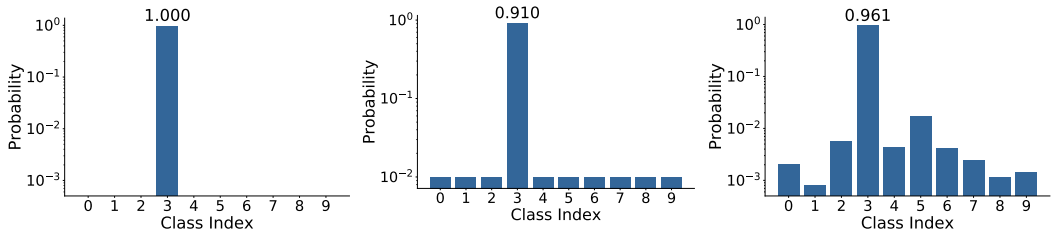


图 3.1 CIFAR-10 数据集上不同类型的标签的分布。目标类别为“cat”，图中使用 *log* 函数缩放 y 轴以进行可视化。（a）硬标签。（b）标签平滑[29] 生成的软标签，这种软标签是硬标签和均匀分布的混合。（c）本文提出的在线标签平滑方法生成的软标签。

的视角的一致性来对模型进行正则化[13]。然而这两种自蒸馏方式会对模型的结构有特殊限制或者会引入额外的训练负担。相比于自蒸馏方法，本文提出的在线标签平滑方法不仅考虑了类别之间的关系，而且具有更高的灵活性，能够以一种在线过程将之前训练好的模型的知识迁移到当前的训练过程。通过这种在线过程，该方法只需要带来额外的极小的存储代价来存储不同的类别的软标签。

为了验证在线标签平滑（OLS）的有效性，本文在 CIFAR-100、ImageNet[139] 和四个细粒度分类数据集[145–148] 上进行了大量的实验。实验结果表明，本文所提出的在线标签平滑方法 OLS 在基线方法上得到了显著的提升。具体而言，将 OLS 直接应用于 ResNet56 和 ResNeXt29-2x64d，可使这两个网络在 CIFAR-100 上的 top-1 准确率分别提高 1.57% 和 2.11%。对于 ImageNet 数据集，在线标签平滑方法可以分别为 ResNet-50 和 ResNet-101[2] 分别带来 1.4% 和 1.02% 的性能提升。在四个细粒度分类数据集上，OLS 在四个不同主干网络上的性能平均比标签平滑 LS[29] 提高了 1.0%，在本文的实验中使用的四个主干网络分别为 ResNet-50[2]、MobileNetv2[136]、EfficientNet-b7[149] 和 SAN-15[150]。另一方面，本文讨论了这种数据相关的在线标签平滑方式可以为训练样本引入类内约束，从而使得类内样本之间距离更加紧密，从而能够为模型起到正则化的作用。本文也在噪声标签的图像分类任务上进行了实验，实验结果表明本文所提出的在线标签平滑方法能够通过对模型施加正则化，从而减少其对噪声样本的拟合，在噪声标签上获得更好的表现。另外，本文进一步在对抗攻击上进行了实验，由于在线标签平滑能够使得同类样本之间更加紧密，所以会使得处于决策边界的样本远离决策边界，从而使得模型具有对对抗攻击更高的鲁棒性。由于本文提

出的在线标签平滑方式灵活易用，所以本文希望该方法能作为一种有效的正则化工具来增强分类模型的训练。

第二节 在线标签平滑方法

3.2.1 在线知识迁移

给定一个有 K 个类别的数据集 $\mathcal{D}_{\texttt{train}}=\{(\boldsymbol{x_i},y_i)\}$，其中 $\boldsymbol{x_i}$ 表示输入的图像，y_i 表示相应的真值标签。对于每个样本 $(\boldsymbol{x_i},y_i)$，神经网络模型会使用 softmax 函数预测一个 k 类别的概率 $p(k|\boldsymbol{x_i})$。硬标签 y_i 的分布 q 可以表述为 $q(k=y_i|\boldsymbol{x_i})=1$ 与 $q(k\neq y_i|\boldsymbol{x_i})=0$。然后，$(\boldsymbol{x_i},y_i)$ 图像分类的标准交叉熵损失可以写成

$$\begin{aligned} L_{hard} &= -\sum_{k=1}^{K} q(k|\boldsymbol{x_i}) \log p(k|\boldsymbol{x_i}) \\ &= -\log p(k=y_i|\boldsymbol{x_i}). \end{aligned} \tag{3.1}$$

标签平滑 LS[29] 没有使用硬标签进行模型训练，而是使用通过均匀分布平滑硬标签分布而生成的软标签。具体而言，$\boldsymbol{x_i}$ 为软标签中的类别 k 的概率可表示为

$$q'(k|\boldsymbol{x_i}) = (1-\varepsilon)q(k|\boldsymbol{x_i}) + \frac{\varepsilon}{K}, \tag{3.2}$$

其中，ε 表示实际中通常设置为 0.1 的平滑参数。如图 3.1(b)所示，标签平滑 LS 所具有的一个假设是非目标类别的置信度可以被同等程度来对待。虽然将均匀分布与原始硬标签相结合对于正则化是有用的，但是标签平滑本身并不考虑不同类别之间的相似性关系[151]。考虑到这一点，本文提出了利用不同类别之间相似性关系的在线标签平滑方法。知识蒸馏在知识迁移以及提升全监督模型精度上起到了重要的作用。许多研究工作表明[8, 129]，利用知识蒸馏，可以从模型预测中有效地发现类别之间的相似性，而知识蒸馏也得益于模型预测中的隐藏知识。基于这一事实，与标签平滑 LS 使用静态软标签不同，本文提出在训练阶段利用模型的预测不断更新用于监督模型训练的软标签。具体来说，在训练过程中，该方法为每个类别保存一个类别级的软标签。给定输入图像 $\boldsymbol{x_i}$，如果分类正确，则使用预测概率 $p(\boldsymbol{x_i})$ 更新对应于目标类 y_i 的软标签。之后，更新后的软标签将用于模型监督。本文提出的方法如图 3.2所示。形式上，让 T 表示训练的迭代次数。然后定义 $\mathcal{S}=\{S^0,S^1,\cdots,S^t,\cdots,S^{T-1}\}$ 作为不同训练时期类别级软标签的集合。其中，S^t 是一个 K 行和 K 列的矩阵，S^t 中的每一列对应于一个

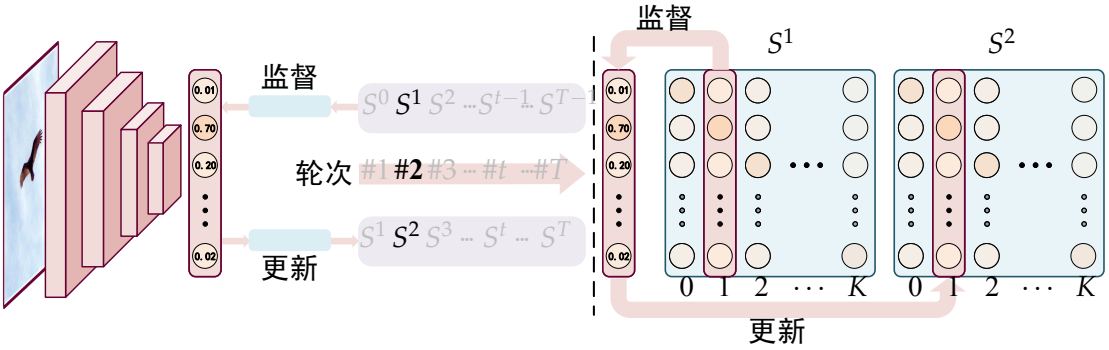


图 3.2 使用本文的在线标签平滑方法训练深度神经网络模型的示例。图的左侧显示了整个训练过程。本文简单地根据训练轮次将训练过程分为 T 个阶段。K 表示数据集中的类别数。本文定义了 S^t 的每一列来表示目标类别的软标签。在每次训练中，本文使用前一训练轮次中生成的软标签来监督模型，同时，本文为下一训练轮次生成软标签。图中展示了第 #2 训练轮次的一个详细示例。

类别的软标签。在第 t 次训练迭代中，给定一个样本 $(\boldsymbol{x_i}, y_i)$，该方法使用软标签 $S_{y_i}^{t-1}$ 形成临时标签分布，以监督模型，其中 $S_{y_i}^{t-1}$ 表示目标类别 y_i 的软标签。由 $(\boldsymbol{x_i}, y_i)$ 的 $S_{y_i}^{t-1}$ 监督模型的训练损失可以表示为

$$L_{soft} = -\sum_{k=1}^{K} S_{y_i,k}^{t-1} \cdot \log p(k|\boldsymbol{x_i}). \tag{3.3}$$

在本文中直接使用上述软标签来监督模型训练，由于在开始训练时，模型参数初始化是随机的，并且缺少硬标签，所以模型在训练初期收敛速度比较慢。因此本文同时使用软标签和硬标签来监督模型的训练。现在，训练总损失可以表示为

$$L = \alpha L_{hard} + (1-\alpha) L_{soft}, \tag{3.4}$$

其中 α 是超参数，用于平衡 L_{hard} 和 L_{soft}。

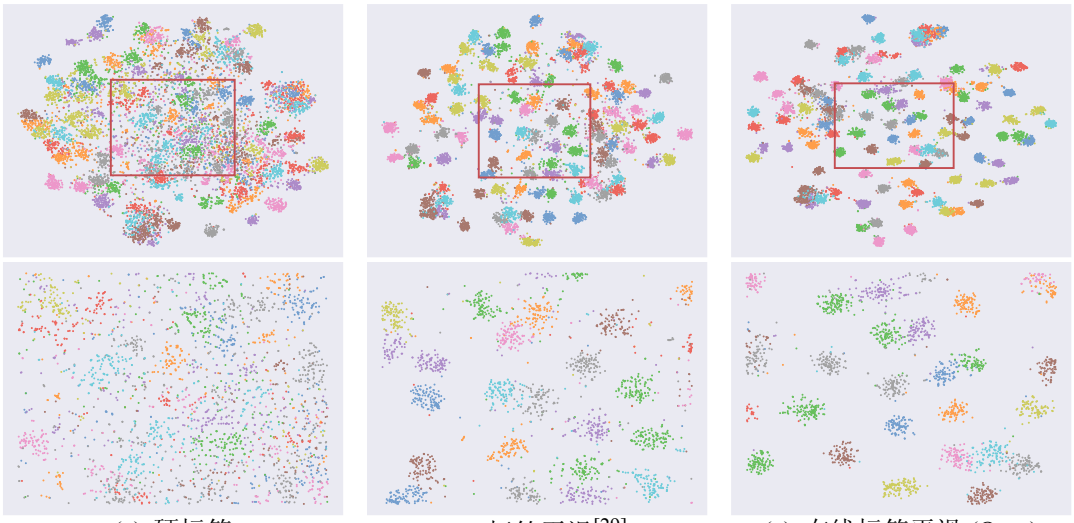
3.2.2 在线生成标签

在第 t 个训练迭代中，该方法还使用输入样本的预测概率来更新 $S_{y_i}^t$，这将用于监督 $t+1$ 次迭代的模型训练。在第 t 个训练迭代开始时，该方法将软标签 S^t 初始化为零矩阵。当模型对输入样本 $(\boldsymbol{x_i}, y_i)$ 进行正确分类时，该方法利用其预测得分 $p(\boldsymbol{x_i})$ 更新 S_t 中的 y_i 列，其公式如下

$$S_{y_i,k}^t = S_{y_i,k}^t + p(k|\boldsymbol{x_i}), \tag{3.5}$$

Algorithm 1 在线标签平滑方法

输入： 数据集 $\mathcal{D}_{\texttt{train}} = \{(\boldsymbol{x_i}, y_i)\}$, 模型 f_θ, 训练轮次 T
初始化： 软标签矩阵 $\boldsymbol{S^0} = \frac{1}{K}\mathbf{1}$, K 表示类别数
for 训练轮次 $t = 1$ **to** T **do**
  初始化： $\boldsymbol{S^t} = 0$
  for *iter* $= 1$ **to** *iterations* **do**
    采样一个批次 $\mathcal{B} \subset \mathcal{D}_{\texttt{train}}$, 输入 f_θ
    获得预测概率 $\{f(\theta, \boldsymbol{x_i}), \boldsymbol{x_i} \in \mathcal{B}\}$
    通过 式 (3.4) 计算损失, 反向传播更新模型参数 θ
    for $i = 1$ **to** $|\mathcal{B}|$ **do**
      更新 $\boldsymbol{S}^t_{y_i} \leftarrow \boldsymbol{S}^t_{y_i} + f(\theta, \boldsymbol{x_i})$
    end for
  end for
  归一化 $\boldsymbol{S^t}$ 的每一列
end for



(a) 硬标签　(b) 标签平滑[29]　(c) 在线标签平滑 (Ours)

图 3.3　使用 t-SNE[152] 对 CIFAR-100 训练集上 ResNet-56 特征表示进行可视化。图中每 10 个类使用相同的颜色。本文可视化了所有 100 个类的表示（第一行）。本文在红色框中进行放大，以便更好地可视化（第二行）。

其中 $k \in \{1, \cdots, K\}$ 表示软标签 $S^t_{y_i}$ 的类别索引。在第 t 次迭代结束时，该方法将按列对累积的软标签 S^t 进行归一化，表示如下

$$S^t_{y_i,k} \leftarrow \frac{S^t_{y_i,k}}{\sum_{l=1}^{K} S^t_{y_i,l}}. \tag{3.6}$$

在归一化之后可以获得所有 K 个类别的软标签 S^t，用于在下一个训练轮次监督模型。特殊的是，在第一个训练轮次时，该方法无法获得累积出的类别级的软

标签，所以使用均匀分布来初始化 S^0 中每一个类别对应的软标签。在这种特殊情况下，该方法等价于普通的标签平滑方式。本文在算法 1中描述了所提出的在线标签平滑方法的整体流程。

3.2.3 引入类内约束

本文分析这样的数据相关的标签平滑方法能够引入类内约束，从而使得类内样本之间距离更加紧密。具体来说，由第 $t-1$ 个轮次生成的软标签 $S^{t-1}_{y_i,k}$ 可表示为

$$S^{t-1}_{y_i,k} = \frac{1}{N}\sum_{j=1}^{N} p^{t-1}(k|\boldsymbol{x_j}), \tag{3.7}$$

其中 N 表示标签为 y_i 的正确预测的样本数。$p^{t-1}(k|\boldsymbol{x_j})$ 是 x_j 在第 $t-1$ 个轮次中输入网络时为类别 k 的概率。那么式 (3.3) 可以重写为

$$\begin{aligned} L_{soft} &= -\sum_{k=1}^{K}\frac{1}{N}\sum_{j=1}^{N} p^{t-1}(k|\boldsymbol{x_j})\cdot\log p(k|\boldsymbol{x_i}) \\ &= -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} p^{t-1}(k|\boldsymbol{x_j})\cdot\log p(k|\boldsymbol{x_i}). \end{aligned} \tag{3.8}$$

该式表明所有正确分类的样本 $\boldsymbol{x_j}$ 将对当前样本 $\boldsymbol{x_i}$ 施加一个一致性约束，会使得属于同一类别的样本之间距离更加紧密。为了给出更直观的解释，本文利用 t-SNE[152] 分别将利用硬标签、标签平滑（LS）和在线标签平滑（OLS）训练的 ResNet-56 模型在 CIFAR-100 上的倒数第二层的表示可视化。如图 3.3所示，本文提出的在线标签平滑的方法能够使得不同类的表示之前更具有区别性，并且同类样本之间会距离更加紧密。此外，由于在整个训练过程中，在线标签平滑只会累积正确的预测，所以这保证了生成的软标签的正确性。

第三节 实验结果对比和分析

3.3.1 图像分类

CIFAR 数据集分类。首先，本文在 CIFAR-100 数据集上进行实验，将提出的在线标签平滑与其他相关方法进行比较，包括标签正则化方法 (Bootstrap[25]，Disturb Label[26]，Symmetric Cross Entropy[44]，Label Smoothing[29] 和 Pairwise Confusion[27]) 以及自知识蒸馏方法（Xu 等人[13] 和 BYOT[12]）。为了与它们进行公平的比较，本文对所有方法保持相同的实验配置。具体来说，本文对所有模型

表 3.1 本文提出的在线标签平滑方法和其他方法之间的比较。本文在 CIFAR-100 上进行三次实验，并计算 Top-1 错误率（%）的平均值和标准差。最佳结果用粗体突出显示。

方法	ResNet-34	ResNet-50	ResNet-101	ResNeXt29-2x64d
Hard Label	20.6 ± 0.2	21.2 ± 0.3	20.3 ± 0.4	20.9 ± 0.5
Bootsoft[25]	21.7 ± 0.1	21.3 ± 0.7	20.4 ± 0.1	21.2 ± 0.1
Boothard[25]	22.6 ± 0.1	20.8 ± 0.1	21.5 ± 0.2	21.0 ± 0.1
Disturb Label[26]	20.9 ± 0.3	22.1 ± 0.5	21.0 ± 0.1	21.6 ± 0.2
SCE[44]	22.9 ± 0.1	22.1 ± 0.1	22.6 ± 0.6	23.1 ± 0.3
LS[29]	20.9 ± 0.1	21.2 ± 0.3	20.1 ± 0.1	20.3 ± 0.2
Pairwise Confusion[27]	22.9 ± 0.1	23.1 ± 0.5	22.7 ± 0.4	21.6 ± 0.1
Xu[13]	22.7 ± 0.1	22.1 ± 0.4	21.7 ± 0.8	22.8 ± 0.1
OLS	**20.0 ± 0.1**	**20.7 ± 0.1**	**19.7 ± 0.2**	**18.8 ± 0.5**
BYOT[12]	20.4 ± 0.1	19.2 ± 0.3	18.5 ± 0.5	19.7 ± 0.1
BYOT[12] + OLS	**19.4 ± 0.1**	**18.2 ± 0.2**	**18.1 ± 0.1**	**18.3 ± 0.2**

进行了 300 个轮次的训练，批大小为 128。学习率最初设置为 0.1，在第 150 个 epoch 和第 225 个 epoch 时衰减为 0.1 倍。对于不同方法中的其他超参数，本文保留其原始设置。此外，为了与 BYOT[12] 和 Xu 等人[13] 进行公平比较，本文删除了其中的特征级监督，只使用类标签来监督模型。

本文在表 3.1中展示了基于不同网络架构的每种方法的分类结果。可以看出，本文提出的在线标签平滑方法显著提高了轻量级和复杂模型的分类性能，这证明了它对不同网络结构的鲁棒性。由于 BYOT[12] 是在深度监督下学习的，因此它在更深的模型（如 ResNet-50 和 ResNet101）上的性能比在线标签平滑方法要好。然而，在线标签平滑可以很容易地与 BYOT[12] 进行联合使用，并且在更深的模型上获得比 BYOT 更好的结果。此外，与标签平滑 LS[29] 相比，OLS 在不同的模型上实现了稳定的性能提升。特别是，OLS 在 ResNeXt29-2x64d 上的性能比 LS 高出约 1.5%。本文分析性能的提高归功于提出的在线标签平滑机制能够利用到类别之间的关系。

ImageNet 数据集分类。为了进一步验证提出的在线标签平滑的有效性，本文在大型数据集 ImageNet 上评估了该方法。ImageNet 数据集包含 1 千个类别，共 120 万张训练图像和 5 万张验证图像。具体来说，本文使用 SGD 优化器对所有模型进行训练，训练轮次为 250，数据批大小为 256。学习率最初设置为 0.1，分别在第 75、150 和 225 个 epoch 时衰减。本文在图 3.2中展示了不同方法在 ImageNet 数据集上的分类表现。在使用在线标签平滑的情况下，ResNet-50 可以

表 3.2 ImageNet 数据集上的分类结果。

Model	Top-1 Error(%)	Top-5 Error(%)
ResNet-50	23.68	7.05
ResNet-50 + Bootsoft[25]	23.49	6.85
ResNet-50 + Boothard[25]	23.85	7.07
ResNet-50 + LS[29]	22.82	6.66
ResNet-50 + CutOut[140]	22.93	6.66
ResNet-50 + Disturb Label[26]	23.59	6.90
ResNet-50 + Tf-KD[153]	23.58	-
ResNet-50 + BYOT[12]	23.04	6.51
ResNet-50 + OLS	22.28	6.39
ResNet-50 + CutOut[140] + OLS	21.98	**6.18**
ResNet-50 + BYOT[12] + OLS	**21.88**	6.27
ResNet-101	21.87	6.29
ResNet-101 + LS[29]	21.27	5.85
ResNet-101 + CutOut[140]	20.72	5.51
ResNet-101 + OLS	20.85	5.50
ResNet-101 + CutOut[139] + LS[29]	20.47	5.51
ResNet-101 + CutOut[139] + OLS	**20.25**	**5.42**

实现 22.28% 的 Top-1 错误率，能够实现比使用标签平滑 LS[29] 提升 0.54%。并且，ResNet-101 可以实现 20.85% 的错误率，能够实现相对基线方法 1% 的提升，比使用标签平滑 LS 提高了 0.42%。这表明，本文提出的在线标签平滑策略在大规模数据集上仍然具有良好的表现。除此之外，本文也探索了本文提出的方法与其他策略的结合，以数据增强（CutOut[140]）和自知识蒸馏（BYOT[12]）为例，本文在表 3.2中展示了将本文提出的在线知识蒸馏策略与其他策略结合的表现。实验结果表明，本文提出的在线知识蒸馏策略能够在数据增强以及自蒸馏方法的基础上带来额外的性能提升。这表明，本文提出的在线标签平滑 OLS 可以用作即插即用的正则化模块，与其他的方法可以简便的进行结合。

细粒度数据集分类。细粒度图像分类任务[154–158] 侧重于从粗粒度类别中判别细粒度的类别[27, 159–161]。本文分别在四个细粒度图像识别数据集上进行了实验，包括 CUB-200-2011[146]、Flowers-102[145]、Cars[147] 和 Aircrafts[148]。为了公平比较，对于所有实验，本文保持相同的实验设置。具体地说，本文使用 SGD 作为优化器，将所有模型训练 100 个轮次。初始学习率设置为 0.01，分别在第 45

表 3.3 细粒度分类数据集上不同网络结构的 Top-1 和 Top-5 错误率 (%)。所有结果在三次运行中取平均值。平均提升表示在所有数据集和主干网络上相对于硬标签性能的平均提升幅度。

数据集	主干网络	Hard Label		LS[29]		Tf-KD[153]		OLS	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CUB-200-2011[146]	ResNet-50[2]	19.19	5.00	18.11	4.88	19.04	4.92	17.53	4.01
Flowers-102[145]		9.31	2.43	7.58	1.93	8.70	2.46	7.14	1.55
Cars[147]		9.58	1.79	8.32	1.57	8.65	1.46	7.46	0.92
Aircrafts[148]		11.88	3.86	9.92	3.73	10.55	3.34	9.19	2.60
CUB-200-2011[146]	MobileNetv2[136]	22.24	6.61	21.33	7.05	22.36	6.41	20.05	5.08
Flowers-102[145]		8.97	2.51	8.06	2.46	8.05	2.23	7.27	1.77
Cars[147]		11.71	2.29	10.17	2.33	10.57	2.14	9.25	1.33
Aircrafts[148]		13.16	4.15	12.05	4.08	11.95	4.04	10.53	2.96
CUB-200-2011[146]	EfficientNet-b7[149]	18.44	5.07	17.40	5.02	20.24	6.33	16.21	3.34
Flowers-102[145]		9.50	2.04	9.42	2.34	8.58	2.07	8.16	1.63
Cars[147]		9.24	1.84	8.42	1.76	9.52	1.64	7.53	0.97
Aircrafts[148]		11.61	3.72	9.60	3.62	9.45	2.01	8.83	2.71
CUB-200-2011[146]	SAN-15[150]	19.05	5.37	17.54	5.43	19.88	5.81	17.28	4.08
Flowers-102[145]		7.85	1.78	8.08	1.95	7.87	1.91	7.09	1.56
Cars[147]		9.23	1.78	8.55	1.87	8.98	1.76	7.55	1.08
Aircrafts[148]		11.31	3.79	9.96	3.45	10.77	4.18	9.43	2.95
平均提升		0.00	0.00	1.11 ↑	0.02 ↑	0.44 ↑	0.19 ↑	**2.00** ↑	**0.96** ↑

和 80 个轮次时衰减。在表 3.3中本文报告了三次运行的平均 Top-1 错误率（%）和 Top-5 错误率（%）。实验结果表明，在线标签平滑 OLS 还可以提高模型在细粒度数据集上的分类性能，这表明利用类别间的关系进行知识迁移仍然有利于细粒度类别的分类。

3.3.2 对噪声标签的鲁棒性分析

如 SL[44, 162] 所述，数据集中存在噪声（不正确）标签，尤其是从网络中获取的标签。由于深度学习强大的拟合能力，模型仍然可以很容易地拟合有噪声的标签[163]，但这对神经网络的泛化性是有害的。本文发现，在线标签平滑方法可以通过减少对噪声样本的拟合来提高深度神经网络对噪声标签的鲁棒性。

本文在 CIFAR-100 上进行实验，以验证本文的方法对噪声数据的正则化能力。为了公平比较，本文与 Arazo 等人[46] 使用了相同的实验设置。在这些实验中，本文根据噪声率随机选取一定数量的样本，并在训练前将这些样本的标签均匀地翻转到错误的标签上（对称噪声）。由于 Ren 等人[41] 和 MetaWeightNet[42]

表 3.4 不同方法在不同噪声率下的分类性能。本文在不同的噪声率下运行三次每种方法，并计算 Top-1 错误率 (%) 的平均值。最好的两个结果以粗体显示。

方法 / 噪声率	0%	20%	40%	60%	80%
Hard Label	26.81	37.75	47.07	62.06	81.56
Bootsoft[25]	27.28	37.99	46.96	63.76	80.32
Boothard[25]	**26.02**	36.21	42.73	54.95	81.20
SCE[44]	28.97	38.40	46.97	62.13	82.66
Ren 等人[41]	38.38	43.74	49.83	57.65	**73.04**
MetaWeightNet[42]	29.51	35.06	43.58	56.15	87.25
Arazo 等人[46]	33.80	**33.91**	**40.87**	**52.91**	83.92
PENCIL[49]	29.36	36.33	43.55	57.49	79.24
Han 等人[40]	32.07	35.08	44.39	62.50	80.39
标签平滑 LS[29]	26.37	35.48	43.99	59.51 0	80.36
在线标签平滑 OLS（Ours）	**25.24**	**32.67**	**38.86**	**50.04**	**78.22**

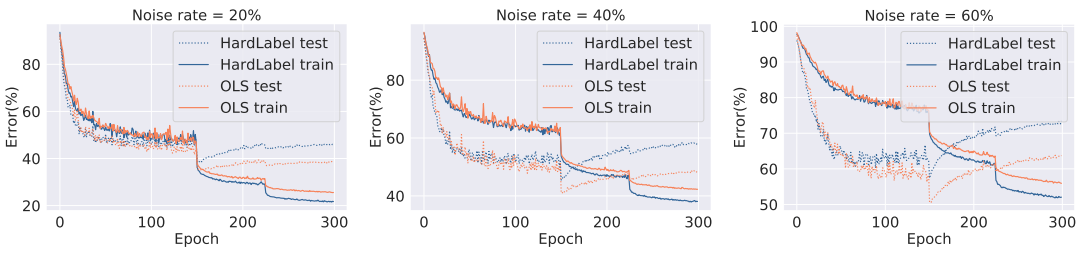


图 3.4 在不同的噪声率 (20%，40%，60%) 下的训练和测试错误率。

都需要从训练集分割出一部分作为正确标签的验证集，因此本文在验证集中保留了他们的默认最佳样本数。在表 3.4中，本文分别报告了当噪声率设置为 0%，20%，40%，60%，80% 时，基于 ResNet-56 模型的分类结果。可以看出，本文提出的在线标签平滑方法与那些专门为噪声标签设计的方法[41, 42, 44, 46] 取得了相当的结果。与标签平滑 LS 相比，OLS 在不同的噪声率下取得了稳定的提升。为了进一步讨论在线标签平滑的正则化作用，本文在图 3.4中可视化了整个训练过程中训练和测试的错误率。从图中可以看出，在训练过程中，在训练集上，在线标签平滑会具有更高的错误率。而在测试集上，在线标签平滑会具有更低的错误误差。这表明，在线标签平滑方法可以有效地减少对噪声样本的过拟合，从而实现更好的泛化性能。

表 3.5 在 CIFAR-10 上对攻击的鲁棒性。本文分别使用 FGSM 和 PGD 算法攻击在 CIFAR-10 上训练的 ResNet-29。本文将 PGD 攻击算法的迭代次数设置为 20 次。

方法	ResNet-29 Top-1 Err(%)	+ FGSM Top-1 Err(%)	+ PGD Top-1 Err(%)
硬标签	7.18	82.46	93.18
Bootsoft[25]	6.91	79.83	92.57
Boothard[25]	7.73	82.68	90.01
Symmetric Cross Entropy[44]	8.66	77.68	93.96
标签平滑[29]	6.81	79.48	87.32
在线标签平滑（Ours）	**6.46**	**60.39**	**76.29**

3.3.3 对对抗攻击的鲁棒性分析

本节首先解释了在线标签平滑对对抗攻击具有鲁棒性的原因。为了获得样本 x 的对抗性样本，FGSM[164] 在样本 x 的邻域 ϵ-ball 中寻找穿过决策边界的点，该点令 x 被错误分类。对抗性样本 x_{adv} 可以表示为：

$$x_{adv} = x + \gamma sign(\nabla_x L(\theta, x, y)), \tag{3.9}$$

其中 L 表示损失函数，γ 是表示优化步长的系数。函数 $sign()$ 是

$$sign(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0. \end{cases} \tag{3.10}$$

如 图 3.5(a) 所示，FGSM[164] 的目的是为每个样本在邻域（ϵ-ball）中找到一个错误分类的扰动点。因此，很容易找到决策边界附近样本的对抗样本。在本文的方法中，对于每个类别 k，软标签由同一类别中所有样本的预测累积而成。式 (3.8) 表明所有正确分类的样本 x_j 都会对当前训练样本 x_i 施加类内约束，使得同一类别内的样本距离更加紧密。如 图 3.5(b) 所示，在一次训练迭代中，OLS 引入的类内约束将促使当前训练样本与同一类中的其他样本更加接近。因此，如图 3.5(c) 所示，OLS 会使得同类样本之间距离更加紧密，这会使得决策边界附近的样本数量减少，所以模型在面对对抗攻击时会更加鲁棒。

本文分别在 CIFAR-10 和 ImageNet 上评估了不同方法训练的模型对于现有对抗攻击算法的鲁棒性。本文分别使用 FGSM[164] 和 PGD[165] 来生成对抗样本。为了实验的公平性，本文采用了 FGSM 的默认参数设置，ℓ_∞ 设置为 8。对于

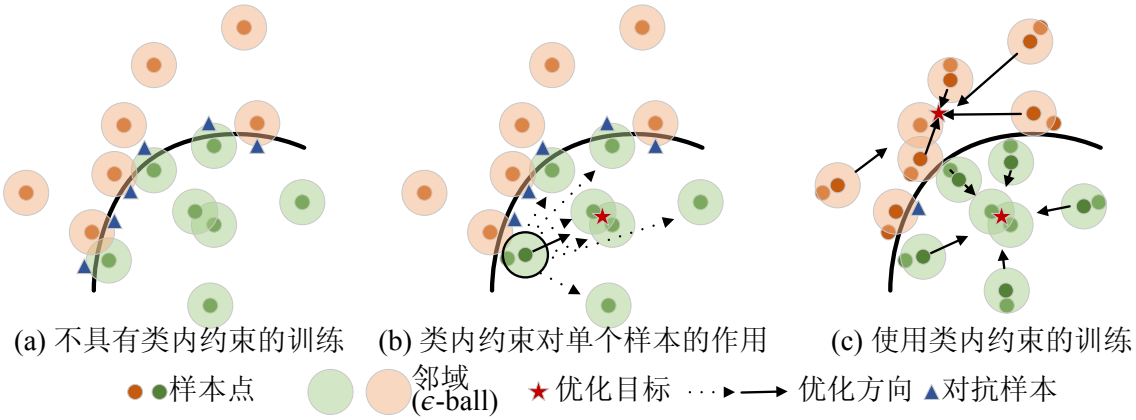


图 3.5 类内约束对模型的正则化作用。(a) 使用硬标签训练的模型很容易得到其对抗样本，因为有许多样本的 ϵ-ball 越过了决策边界。(b) 在一次训练迭代中，OLS 对当前训练样本和同一类的所有其他样本之间施加一致性约束，使得当前训练样本远离决策边界。虚线表示同一类中的每个样本对当前训练样本具有一致性约束。实线表示此训练样本的优化方向。(c) 类内约束使同一类中的样本更近，距离决策边界更远，这会使得找到对抗性样本变得更加困难。

表 3.6 在对抗攻击后，ResNet-50 在 ImageNet 上的 Top-1 和 Top-5 错误率 (%)。对于两种对抗性攻击算法，FGSM 和 PGD，本文保留其默认设置。本文将 PGD 攻击算法的迭代次数设置为 20 次。

ResNet-50	+ FGSM		+ PGD	
	Top-1 Err(%)	Top-5 Err(%)	Top-1 Err(%)	Top-5 Err(%)
Hard Label	91.07	66.21	94.93	31.82
Bootsoft[25]	91.29	67.29	94.56	31.07
LS[29]	**74.44**	50.63	80.31	24.46
OLS	75.79	**48.13**	**74.43**	**22.14**

PGD，本文采用与 Peterson 等人[166] 相同的实验设置，但是我们将对抗攻击的迭代次数增加到 20 次，以获得更好的攻击效果。在表 3.5中，本文在 CIFAR-10 数据集上报告了对各种方法使用 FGSM 和 PGD 算法的对抗性攻击后的 Top-1 错误率。从该表中可以看出，经过 FGSM 和 PGD 攻击后，使用 OLS 训练的模型具有最低的 Top-1 错误率。使用本文提出的的 OLS 算法训练的模型比使用其他方法训练的模型对对抗性攻击更具鲁棒性。此外，如表 3.6所示，本文在 ImageNet 上进行了相同的实验。与硬标签相比，OLS 在 Top-1 错误率上的平均增益为 17.9%，在 Top-5 错误率上的平均增益为 13.9%。本文所提出的方法在 Top-1 错误率和 Top-5 错误率方面分别比标签平滑 LS[29] 提升 2.3% 和 2.4%。这些提升得益于本

表 3.7 集成模型的 Top-1 错误率。本文分别集成了 6 个、10 个、15 个和 20 个在不同时期训练的模型。模型从所有训练轮次中均匀选择。

方法	1 Model	6 Models	10 Models	15 Models	20 Models
硬标签	26.41	26.07	25.93	25.87	25.88
标签平滑 LS[29]	26.37	25.30	25.11	24.97	24.96
OLS (ours)	25.27	24.52	24.22	24.10	23.91

文所提出的数据相关的软标签所引入的类内约束能够让同类样本之间距离更加紧密，从而减少决策边界附近的样本数量。这些实验表明，本文提出的在线标签平滑 OLS 能够有效地提高模型对对抗攻击的鲁棒性。

3.3.4 与模型集成的关系

集成不同迭代次数的训练好的模型是一种有效且节省成本的集成方法。将在不同迭代中训练的模型的输出进行集成的方法如下所述：

$$z_i = \frac{1}{||T||}\sum_{t\in T} softmax(W(x_i|\theta_t)), \tag{3.11}$$

其中，z_i 表示集成预测，T 表示不同迭代中所选定模型集合，W 表示网络模型，θ_t 表示第 t 个轮次的网络参数，x_i 表示输入样本。

本文所提出的在线标签平滑方法和模型集成都利用了不同训练迭代的模型的知识。模型集成是对不同训练时期的模型的输出进行平均，从而给出最终的预测结果。然而，与集成方法不同，本文的方法利用前一个迭代轮次的知识来帮助当前轮次的学习。具体来说，本文所提出的方法在前一个训练轮次中生成软标签，并使用其来监督当前训练轮次的模型的训练。这样的一个操作实现了将前一个训练轮次模型的知识迁移给当前训练的模型。但是，这样的一个迁移策略与模型集成并不冲突，为了验证这一点，本文使用 ResNet-56 在 CIFAR-100 数据集上进行了模型集成的实验。具体来说，本文采用了 节 3.3.1中描述的相同的实验设置，实验结果如 表 3.7所示。对于所有方法，本文采用了相同的集成策略，我们从整个训练过程（300 个轮次）中均匀地选择模型，分别选择 6 个、10 个、15 个和 20 个模型用于集成模型。在表 3.7中，本文所提出的方法的 Top-1 错误率为 25.27%。当我们提出的方法与模型集成相结合时，性能将进一步大幅度提高（“20 个模型”：23.91%）。所以，实验结果表明，虽然本文提出的在线标签平滑策略能够将之前训练的模型的知识迁移给当前模型，但是这与模型集成并

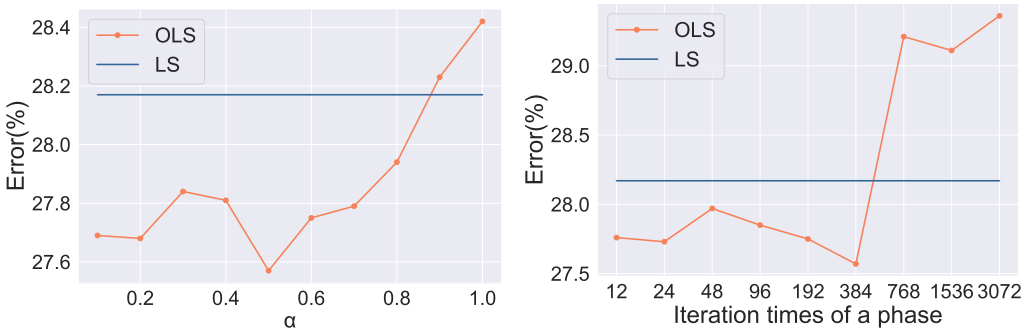
不冲突，配合模型集成使用能进一步提升模型的表现。

3.3.5 消融实验

在本小节中，我们首先进行实验来研究提出的方法中的超参数对于模型最终表现的影响。之后，本文分析了软标签所表示的类别之间的关系的重要性。最后，本文分析了提出的方法对模型的正则化效果。在本小节中，所有的实验都是在 CIFAR 数据集上进行的。

超参数的影响。基于 ResNet-29，本文分析了式 (3.4) 中的超参数 α。与之前直接将 α 设置为 0.5 的实验不同，在本小节中枚举了一些可能的值 $\alpha \in \{0.1, 0.2, \cdots, 1.0\}$。之后在图 3.6(a) 中绘制了实验结果。从图中可以看出，当 α 设置为 0.5 时，模型可以实现最低的 Top-1 错误率。当 α 设置为 0 时，模型此时缺少正确类别的信息的监督，所以很难收敛，这一实验现象表明模型仍然需要原始的硬标签提供的正确类别的信息。当 α 从 0.1 开始逐渐增加到 0.5 时，模型的性能在逐渐提升，然而当 α 处于区间 $[0.5, 1.0]$ 时，模型的性能在逐渐下降。此外，本文进一步通过实验研究了训练过程中软标签矩阵 S 的更新周期对模型表现的影响。在本文所提出的方法中，将更新周期设置为一个训练轮次。如图 3.6(b) 所示，本文使用不同的更新周期 (迭代次数 $\in \{12, 24, 48, \cdots, 1536, 3072\}$) 来评估在线标签平滑方法。实验表明，当更新周期设置为一个训练轮次时，模型可获得最佳性能。从该图中可以观察到当更新周期小于一个训练轮次时（1 个训练轮次大约是 384 次迭代），模型的分类表现非常接近。然而，当更新周期大于一个训练轮次时，模型的性能会急剧下降。这是因为随着网络的训练，模型的预测结果会变得越来越好，然而当使用更多的迭代来更新软标签时，目前监督模型使用的软标签很容易过时，早期模型预测所包含的知识已经与目前预测有很大不同。所以，更新软标签的周期并不是越长越好。

类别之间关系的重要性。本文认为包含不同类别之间的关系的软标签有益于提升分类模型的表现，所以本文在本小节中探索类别之间的关系的重要性。具体而言，本文利用人类标注的不确定性数据集 CIFAR10H[166] 来验证类别之间关系的重要性。CIFAR-10H 通过收集 CIFAR-10 测试集中每个样本的许多人的分类判断，从而确定人类对每个样本的不确定性。人类不确定性标签可以看作是一种软标签，它可以反映出不同类别之间的相似性。Peterson 等人[166] 发现，在人类不确定性标签上训练的模型比在硬标签上训练的模型具有更好的准确性和泛化性。这表明，合理的类别之间的关系对于提升模型性能是非常有利



(a) 超参数 α 对于模型表现的影响　　(b) 更新软标签的周期对于模型表现的影响

图 3.6　超参数对模型表现的影响。

表 3.8　不同方法在 CIFAR-10 和 CIFAR-10H 上的评价结果。本文在 CIFAR-10 上用不同的方法训练 ResNet-29。本文使用平均 KL 散度来测量模型的预测分布和 CIFAR-10H 测试集上的人类不确定性判断之间的差异。

Method	CIFAR-10 Top-1 Err(%)	CIFAR-10H KL Divergence
Hard Label	7.18	0.2974
Bootsoft[25]	6.91	0.3247
Boothard[25]	7.73	0.3188
Symmetric Cross Entropy[44]	8.66	0.5563
LS[29]	6.81	0.1866
OLS	**6.46**	**0.1399**

的。为了探索本文提出的方法发现的类别之间关系的合理性，本文使用 KL 散度来衡量模型的预测概率分布与 CIFAR-10H 上的人类不确定性分布之间的差异。为了公平的比较，在计算 CIFAR-10H 的 KL 散度时，本文只考虑每个模型的正确预测的样本。表 3.8中列出了不同方法在 CIFAR-10H[166] 上的平均 KL 散度和在 CIFAR-10 上的 Top-1 错误率 (%)。实验结果表明，使用在线标签平滑 OLS 训练的模型的预测分布更加接近人类不确定性，这证明本文提出的在线平滑方法训练出的模型能够建模类别之间更加合理和正确的关系。

样本级软标签。在线标签平滑中对同类的不同样本的预测进行累积是非常重要的操作。为了验证累积模型预测的统计特征的有效性，本文使用单个样本的预测分布来作为监督信息（表示为 OLS-Single）进行了实验。具体来说，对于每个训练样本，本文随机选择另一个具有相同类别的训练样本，之后获取该随机选择的训练样本的预测分布，并利用该分布作为软标签，对当前训练样本进行监督。基于 ResNet-56 模型，OLS(25.24 ± 0.18) 比 OLS-single(26.18 ± 0.30) 的

表 3.9 CIFAR-100 上的 Top-1 Error(%) 和预期校准误差（ECE）。

Method	ResNet-56		ResNet-74		ResNet-110	
	Top-1 Error(%)	ECE	Top-1 Error(%)	ECE	Top-1 Error(%)	ECE
Hard Label	26.81	11.37	25.86	12.70	25.54	13.14
LS[29]	26.37	3.35	25.90	2.37	25.14	2.32
OLS	**25.24**	**2.85**	**24.89**	**1.81**	**23.86**	**2.05**

精度高出约 1%。这一实验结果表明，对不同样本的预测进行累积得到的软标签可以具有更好的稳定性来反映不同类别之间的关系。

校准效果。 Guo 等人[10] 提出了置信度校准的评价指标，用于衡量模型对训练集的过拟合程度。期望校准误差（ECE）[10] 用于衡量分类模型的过拟合程度，在本小节中也使用这一指标来衡量在线标签平滑 OLS 的校准模型的能力。表 3.9报告了几个模型的 Top-1 错误率 (%) 和 ECE，实验结果表明，在线标签平滑 OLS 的错误率比标签平滑 LS 平均低 1.14%。并且，本文提出的方法在三种不同深度的模型上也获得了更低的期望校准误差。这表明在线标签平滑 OLS 可以更有效地防止过度置信的预测，并显示出更好的校准能力，表现出了良好的正则化的能力。

第四节 本章小结

在单次学习场景下，知识蒸馏在提升模型表现以及进行知识迁移上发挥了重要的作用。然而目前现有知识蒸馏方法具有灵活性低的问题，而标签平滑则提供了一种具有高灵活性的知识蒸馏方式。由于目前的知识蒸馏得益于模型预测中隐藏的类别之间的关系，所以本文提出了一种引入类别之间关系的在线标签平滑方式。该方法通过统计模型的预测来得到类别之间的关系，能够有效地将之前训练的模型的知识迁移到当前训练的模型中。本章首先阐述了所提出的方法的动机和贡献，之后详细介绍了提出的在线标签平滑的方法。最后，为了验证方法的有效性，本章分别在 CIFAR、ImageNet 和四个细粒度数据集上评估了提出的在线标签平滑方法的表现，表明了该方法的有效性。

第四章　基于表征补偿机制的在线知识迁移

基于单次学习的深度学习模型在许多领域已经实现了很多里程碑式的进展。然而，这些模型只能处理预先设定好的固定类别或固定领域的数据，难以满足在实际应用中的可拓展性的要求。因此，连续学习这种研究如何有效地对模型进行拓展的机器学习算法得到了研究人员的广泛关注。目前，现有连续学习方法存在两方面的局限性：一是基于知识蒸馏的方法仍然是优化同一套参数，难以在保留旧知识和学习新知识之间进行权衡；二是基于拓展模型的方法随着连续学习步骤的增多，会使得模型越来越庞大。为了解决现有连续学习方法存在的这两点问题，在本章中，本文提出了一种基于表征补偿机制的自适应知识迁移策略。具体来说，该策略通过在参数空间中解耦对旧知识的记忆和对新知识的学习，其中一个分支的参数是固定的，用于保持旧知识，另外一个分支的参数是可训练的，用于学习新知识。通过这种解耦机制，可以尽可能减少对模型中的旧知识的破坏。并且，在最终优化目标的作用下，可训练的分支不仅要负责学习新的知识，并且要自适应地适配另外分支的旧知识。于是，这种解耦机制能够使得模型在这两者之间实现更好的权衡。在本章节中，为了验证所提出解耦机制的有效性，本文在三个公开的语义分割数据集上进行了实验。

本章节的内容安排如下：首先阐述了研究动机以及贡献，之后详细阐述了基于表征补偿机制的在线知识迁移机制以及提出的知识蒸馏机制，并且在三个公开的语义分割数据集上对连续类别语义分割和连续领域语义分割两种设置进行了广泛的实验，最后对实验结果进行分析和讨论。

第一节　研究动机以及贡献

目前数据驱动的神经网络[167–170] 在语义分割领域已经取得了巨大的进展。然而，这些依赖于训练数据的全监督的模型[171–173] 只能用来处理固定数量的类别。真实世界中的应用场景期望模型能够动态地拓展，从而能够识别新的类别。一个非常简单的解决方案是在新数据集上微调模型，但是这样会导致模型很快偏差到新的数据类别上，失去对旧类别的判别能力，这被称之为灾难性遗忘现象[74]. 另一个非常直接的解决方案是重新构建训练集，使其包含所有目前需要

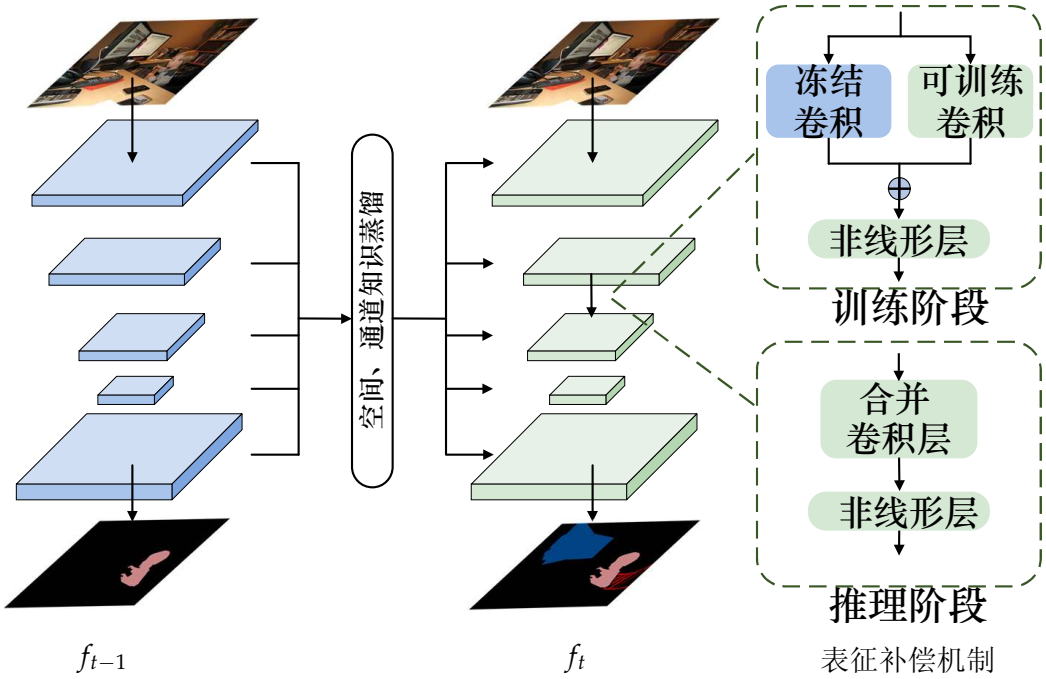


图 4.1 本文提出的连续学习的训练框架。该方法中设计了两种机制，分别是表征补偿机制和池化立体蒸馏方法。

模型处理的类别的数据。之后使用这个重新构建的训练集重新训练模型，这种方法称之为联合训练。

然而，这样做存在很多问题，首先是重新训练模型存在的成本问题。对于大量数据训练出的模型，在频繁拓展新类别的情况下，就需要频繁地构建新训练集和训练模型，这会带来高昂的成本。其次，数据隐私问题也是联合训练方法的一个限制条件，很多情况下，因为隐私政策等等问题，原来用于训练模型的数据无法一直保留。这会导致无法使用原来的数据和现在新增加的数据重新构建数据集。所以，只利用现在新增加的数据来拓展模型是非常必要的，从而实现能够同时识别旧类别和新类别的目的。于是，本文研究使用连续学习算法来缓解学习过程中的灾难性遗忘问题，使得模型能够同时识别旧类别和新类别。

本章以连续语义分割[15, 16, 125, 174] 作为研究的任务，在给定之前训练好的模型和新类别的训练数据的情况下，模型被期望能够鉴别所有见过的类别，包括之前见过的类别（旧类别）和新增加数据中的类别（新类别）。并且，为了节省语义分割标签的标注成本，新增加的数据只需要标注新类别的像素级标签，旧类别的区域会被当成背景，不进行标注。如果直接使用新数据进行模型的训练，这样的一个设置是非常具有挑战性的，会非常容易导致灾难性遗忘[74] 的问题。

如[24, 66, 74] 表明，在新数据上进行微调模型很可能导致灾难性遗忘，模型会快速拟合新类别的数据分布，失去对旧类别的判别能力。一些方法[74–80] 在模型参数上添加正则化来提升模型的稳定性，缓解灾难遗忘的问题。然而，所有的参数仍然会在新类别的训练数据上被更新。这给缓解灾难遗忘带来了很大的挑战，因为新知识和旧知识在模型参数中耦合在一起，这会让模型在维持学习新知识和保持旧知识的脆弱的平衡极其困难。一些其他方法[19, 81–85] 增加了模型的容量来在模型的稳定性和可塑性上获得一个更好的权衡，但是会导致增加模型的计算代价和存储代价。模型的稳定性是指模型保持旧知识的能力，可塑性是指模型学习新知识的能力。

本章提出了一个即插即用的表征补偿机制，能够在记住旧知识的同时，具有额外的模型容量来学习新知识。受到结构重参数化方法的启发[50, 51]，在训练过程中，本文将卷积层替换成了两个并行的分支，本文将这个结构称为表征补偿机制。如 图 4.1所示，在训练过程中，两个并行的卷积层的输出在模型中的非线性激活层前进行融合。在每个连续学习步骤的开始，本文等价地合并两个并行卷积层的参数为一个，并且这个合并后的卷积层参数会被冻结来保留旧的知识。另外一个分支仍然是可训练的，并且它会从之前步骤中相对应的分支来继承参数。这样的一个表征补偿机制希望利用冻结的分支来记忆住旧知识，同时利用可训练的分支作为额外的模型容量来学习新知识。更重要的是，这样的的一个机制在模型的推理阶段，不会带来任何额外的参数和计算代价。

为了进一步缓解灾难遗忘的问题[74]，本文在隐藏层之间引入了一个显式的知识蒸馏机制[175]。如图 4.1所示，本文将这个知识蒸馏机制命名为池化立体蒸馏。这个蒸馏机制可以抑制局部特征图中的误差和噪声的负面影响。本节的主要贡献总结如下：

- 本节提出了一个表征补偿机制，这个机制在训练时具有两个并行分支。其中一个用于保留旧知识，另一个用于自适应到新数据上。在连续学习的步骤中，该机制在推理阶段不会带来额外的的计算和存储代价。
- 本文在连续类别分割和连续域分割两种设置上进行了实验。实验结果表明，本文提出的方法在三个数据集上取得了最高的精度。

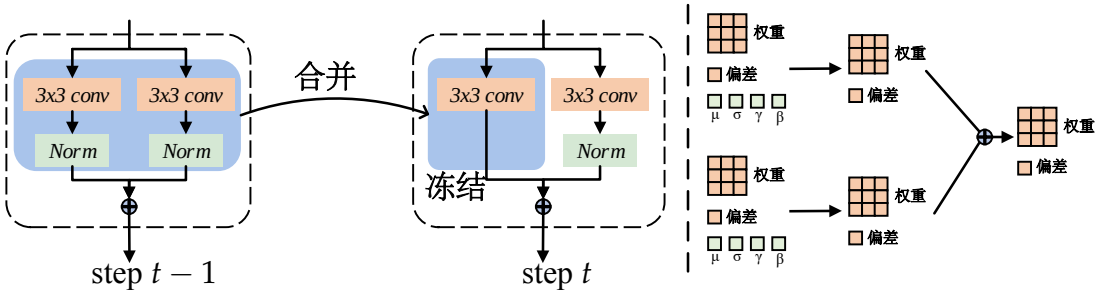


图 4.2 本文提出的基于重参数的表征补偿机制的示意图。该方法将 3×3 卷积修改为两个并行的卷积。来自于这两个分支的特征在激活层之前进行融合。于是，在第 t 个步骤的训练开始阶段，在第 $t-1$ 个步骤训练好的两个并行的分支可以被合并为一个等价的卷积层，而这个卷积层会被冻结，并且会被看作第 t 个步骤的一个分支。在第 t 个步骤中模型的另外一个分支会使用第 $t-1$ 步骤中的模型对应的分支进行初始化。图中右半部分展示了合并操作的过程。

第二节 表征补偿机制

4.2.1 基于表征补偿机制的知识迁移机制

如 图 4.2所示，为了解耦保留旧知识和学习新知识，本文提出了表征补偿机制。在目前大多数的深度神经网络中，一个常见的组件是一个 3×3 卷积层，后面紧跟着归一化层和非线性层。本文修改了这一结构，将结构修改为一个并行的 3×3 卷积层，之后跟着归一化层。两个并行的卷积-归一化层的输出融合在一起之后，特征图会经过一个非线性层进行处理。形式化地，这样一个结构包含两个并行的卷积层，本文分别使用 $\{W^0,W^1\}$ 和 $\{b^0,b^1\}$ 来分别表示两个卷积层的权重和偏差。假定 $Norm^0=\{\mu^0,\sigma^0,\gamma^0,\beta^0\}$ 和 $Norm^1=\{\mu^1,\sigma^1,\gamma^1,\beta^1\}$ 代表两个归一化层的均值、方差、权重和偏差。于是，在非线性激活层之前的输入 x 的计算可以表示为

$$\begin{aligned}\hat{x}&=\sum_{i=0}^{1}Norm_i(W_ix+b_i)\\&=\sum_{i=0}^{1}(\gamma_i\frac{W_ix+b_i-\mu_i}{\sigma_i}+\beta_i)\\&=(\sum_{i=0}^{1}\frac{\gamma_iW_i}{\sigma_i})x+\sum_{i=0}^{1}(\frac{\gamma_ib_i-\gamma_i\mu_i}{\sigma_i}+\beta_i)\\&=\hat{W}x+\hat{b}.\end{aligned}\tag{4.1}$$

这个公式表明两个并行的分支可以被等价地表示为一个卷积层，其权重和偏差分别使用 $\hat{W}$ 和 $\hat{b}$ 来表示。本文也在 图 4.2的右半部分展示了合并卷积层参数的

过程。于是，对于这个编辑后的结构，本文可以等价地合并两个分支的参数为一个卷积。

更加具体地，在第 0 个学习步骤，所有的参数都是可以训练的，从而训练出一个能够判别 $\mathcal{C}_0$ 个类别的初始模型。对于后续的学习步骤，模型被期望能够分割新增加的类别。在这些连续学习的步骤中，网络将会使用之前步骤中训练好的参数进行初始化，这对知识迁移是非常有利的[16]。在第 t 个学习步骤的开始，因为模型被期望用于避免旧知识，所以本文合并了在第 $t-1$ 个学习步骤中训练好的并行的分支，将其合并为一个卷积层。如 图 4.2所示，这个合并后的卷积层中的参数是冻结的，用于记忆旧知识。另外一个分支仍然是可以训练的，用于学习新的知识，这个分支使用之前步骤中训练好的相关的分支的参数进行初始化。除此之外，本文设计了一个随机路径丢弃策略，这个策略被用在聚合两个分支的输出 x_1 和 x_2。在训练过程中，在非线性激活层之前的输出可以被表示为

$$\hat{x} = \eta \cdot x_1 + (1-\eta) \cdot x_2, \tag{4.2}$$

其中 η 是一个随机通道维度上的加权向量，其中的每个值都是从集合 $\{0, 0.5, 1\}$ 中均匀选取的。在推理阶段，向量 η 的每个内容都被设置为 0.5。实验结果表明这样一个策略带来了一些微弱的额外提升。

4.2.2 表征补偿机制的有效性分析

如 图 4.3所示，这种并行卷积结构可以被看作是一种许多子网络的隐式集成[2, 176]。在这些子网络中，一些层的参数是合并之后的卷积层（在上一个学习步骤中训练好的），并且是冻结的。在训练过程中，类似于交互式知识蒸馏[177]，这些冻结的层将对可训练的参数引入正则化，使得可训练的层表现得更像在上一个训练步骤中训练好的模型中的对应层。在一种特别的情况下，如 图 4.3(a) 所示，在子网络中只有一层是可训练的，在训练过程中，这一层将会同时考虑适应冻结层的表示，和学习新知识。于是，这种机制可以缓解可训练层的灾难性遗忘的问题。本文进一步推广这个形式到普通的子网络，例如 图 4.3(b) 所示，这也能够鼓励可训练层去适应冻结层的表示。之后，所有的子网络可以被集成在一起，从不同的子网络中集成知识到一个网络中，例如 图 4.3(c) 所示。

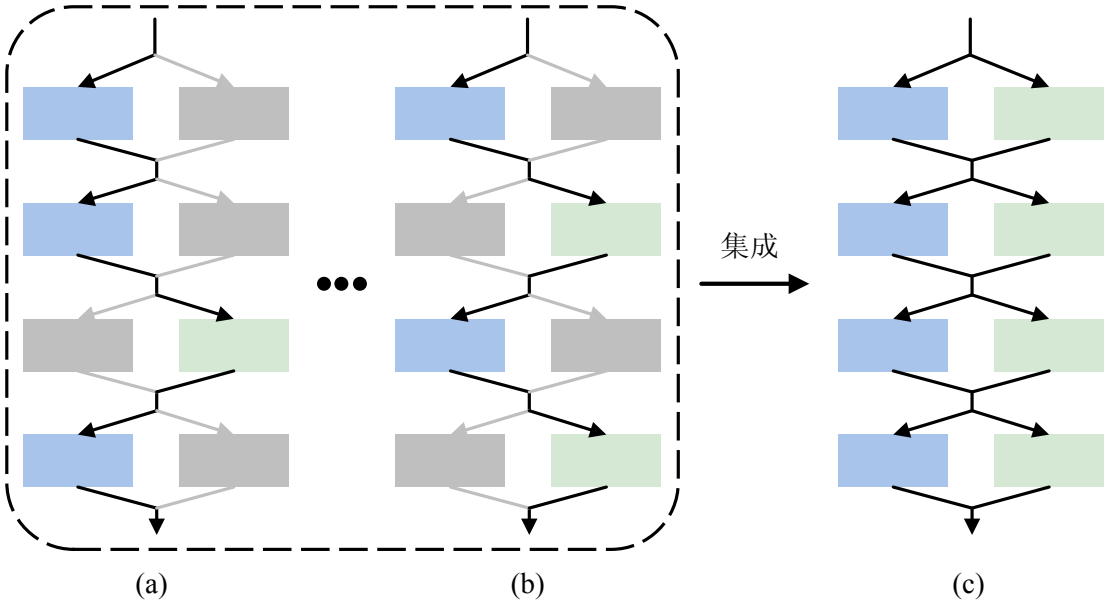


图 4.3 表示补偿网络的有效性解释。本文提出的网络架构（c）可以看成是许多子网络如（a）和（b）的隐式集成。图中使用蓝色来表示从合并的教师模型中继承的冻结的参数层。图中使用绿色来表示可训练的层。图中使用灰色来表示在子网络中被忽略的层。

4.2.3 知识蒸馏机制

为了进一步缓解对旧知识的遗忘问题，与 PLOP[15] 类似，本文也探索在隐藏层之间进行知识蒸馏。如 图 4.4(a) 中所示，PLOP[15] 引入了条带池化[111] 来分别从教师模型和现在要训练的模型的特征图中聚合信息。这样一个池化操作在保持对旧类别的判别能力和允许学习新类别上起到了关键的作用。在本文的方法中，本文设计了在空间维度上的基于平均池化的知识蒸馏方法。另外，为了保持每个像素位置的特征强度的分辨性，本文也在通道维度上使用了平均池化来进行知识蒸馏。总之，如 图 4.4(b) 所示，本文设计的知识蒸馏方法会在空间和通道维度上都使用平均池化操作。

形式上，本文选择使用所有的 L 个网络阶段中的最后一层非线性激活层之前的特征图 $\{X^1, X^2, ..., X^L\}$，包括解码器和主干网络中的所有阶段。对于教师模型和学生模型中的特征，本文首先计算每个像素值的平方来保留特征中的负信息。之后，本文分别在空间和通道两个维度上进行多尺度的平均池化。来自于

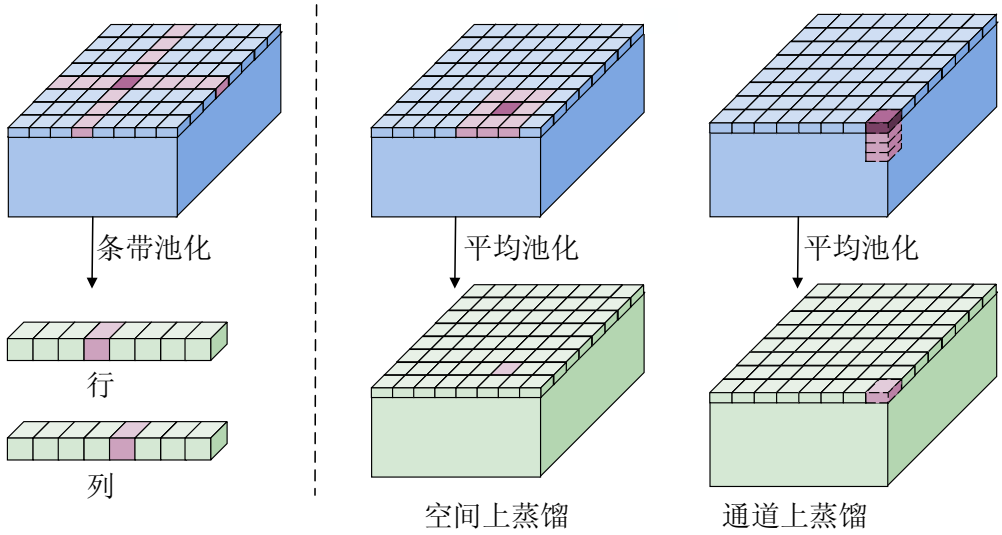


图 4.4 PLOP[15] 和本文提出的蒸馏机制之间的比较。

教师模型和学生模型的特征 $\hat{X}_T^l, \hat{X}_S^l$ 通过平均池化操作 $\odot$ 计算得到，表示如下：

$$\begin{aligned} \hat{X}_T^{l,m} &= M \odot [(X_{T,ij}^l)^2] \\ \hat{X}_S^{l,m} &= M \odot [(X_{S,ij}^l)^2], \end{aligned} \tag{4.3}$$

其中 M 表示第 m 个平均池化核，l 表示第 l 个网络阶段。对于在空间维度上的平均池化操作，本文使用多尺度窗口来建模局部区域内的像素之间的关系。池化核 M 的尺寸会从 $\mathcal{M} = \{4,8,12,16,20,24\}$ 中进行选择，并且池化操作的步长会设置成 1。并且对于在通道维度上的平均池化，本文简单地设置窗口的尺寸大小为 3。之后，在隐藏层上的空间知识蒸馏的损失函数 L_{skd} 可以被表示为

$$L_{skd} = \frac{1}{L}\frac{1}{|\mathcal{M}|}\sum_{l=1}^{L}\sum_{m=1}^{|\mathcal{M}|}\sqrt{\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{d=1}^{D}[(\hat{X}_{T,ijd}^{l,m} - \hat{X}_{S,ijd}^{l,m})^2]}, \tag{4.4}$$

其中，H,W,D 分别表示高度，宽度和通道数。在通道维度上进行蒸馏，其损失函数也使用相同的计算方式来计算 L_{ckd}，不同的是池化核的大小会设置成 $\mathcal{M} = \{3\}$。总之，蒸馏的优化目标可以被表示为

$$L = L_{skd} + L_{ckd}. \tag{4.5}$$

(a) 输入图像 (b) 第 2 步骤 (c) 第 3 步骤 (d) 分割标注

图 4.5 在 PLOP[15] 中使用的条带池化（第一行）和本文方法中使用的平均池化（第二行）的对比。

4.2.4 平均池化与条带池化的比较

得益于条带池化强大的建模长范围依赖的能力，条带池化在许多全监督语义分割模型中起到了重要的作用[111, 178]。目前，连续语义分割的表现仍然要比全监督分割模型的精度差很多。在连续语义分割的场景下，相比于全监督分割模型，模型的预测结果中经常包含有更多的噪声或者错误。于是，在蒸馏过程中，当使用条带池化来聚合特征时，这样的一个长范围依赖会对交叉点引入一些无关的噪声，会导致噪声的扩大。这将导致学生的预测结果进一步恶化。本文所提出的方法使用在局部区域内的平均池化来抑制噪声的负面影响。具体来说，因为局部区域的语义往往是相似的，当前的关键点可以从更多的局部区域内的邻域像素聚合特征。于是，当前的关键点受到噪声和误差的负面影响会更小。

本文在 图 4.5(b) 中展示了一个例子，由于条带池化引入的长范围依赖，其可能会对关键点引入远距离的噪声和错误。在蒸馏阶段，噪声被进一步传播到了学生模型，使得连续学习的训练过程发生恶化。为了解决这一问题，本文发现使用平均池化 图 4.5能使得关键点考虑到邻域内的许多相关点，从而使得聚合之后的特征对噪声更加鲁棒。

第三节 实验结果对比和分析

4.3.1 实验设置

4.3.1.1 数据集

PASCAL VOC 2012[179] 是一个常用的公开数据集，包含 10582 张训练集图像和 1449 张图像，共分为 20 个物体类别和 1 个背景类别。**ADE20K**[180] 是一个涵盖日常生活的用于语义分割的数据集。它包含有 20210 张训练集图像和 2000 张验证集图像，共分为 150 个类别。**Cityscapes**[181] 包含有 2975 张训练集图像，500 张验证集图像和 1525 张测试集图像。该数据集包含从 21 个城市中采集的 19 个类别。

4.3.1.2 实验协议

连续类别语义分割在多个序列化的连续学习步骤中，训练模型用于识别不同的类别。在每一步骤中，模型需要增加对一些类别的识别能力。保持和 MiB[16]、PLOP[15]、SDR[125] 相同的设置，本文假定之前训练步骤中的训练数据是不可得到的，换句话说，模型只能获得现在学习步骤中的数据。除此之外，对于现在步骤中的训练数据，只有当前学习步骤要学习的新类别才会被标注。所有其他的类别都会忽略，并标注为背景。这样一种设置是出于代价最小的原则，既要考虑到原来的训练数据的不可得到的风险，也要考虑到目前新数据的标注成本。目前，MiB[16] 提出了两种公共使用的连续类别语义分割的实验设置，分别称为非重叠设置和重叠设置。在非重叠设置中，假定知道在将来会出现的所有类别，现在学习步骤中的训练图像不会包含将来出现的任何类别。相对而言，重叠设置是更加符合真实应用场景的设置。这一设置允许在将来可能出现的类别出现在现在的训练图像中。

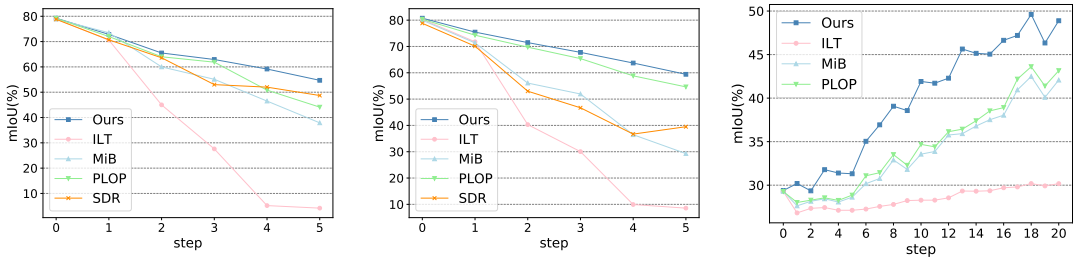
本章在 PASCAL VOC 2012[179] 和 ADE20K[180] 两个数据集上进行了连续类别语义分割实验。本文保持与之前方法相同的设置[15, 16, 125]，本文将每个在新增加的数据集上的训练阶段称之为一个连续学习步骤。X-Y 表示本文实验中的连续学习的不同设置，其中 X 代表模型需要在第一个连续学习步骤中学习的类别数。在之后的每个连续学习步骤中，新增加的数据集包含有 Y 个类别。在 PASCAL VOC 2012[179] 上，本文在三个不同的设置上进行了实验，分别是 15-5 (2 steps), 15-1 (6 steps) 和 10-1 (11 steps). 举例来说，15-1 表示在第一个学习步骤

中，本文在初始的 15 个目标类别和 1 个背景类别上训练模型。在后续的 5 个步骤中，模型需要在新数据集上进行训练，每个步骤中的数据集包含一个新增加的类别。于是，模型在最后一个学习步骤的时候就可以判别 20 个目标类别和 1 个背景类别。在 ADE20K[180] 上，本文应用了四个不同的设置，分别是 100-50（共包含 2 个学习步骤），50-50（共包含 3 个学习步骤），100-10（共包含 6 个学习步骤）以及 100-5（共包含 11 个学习步骤）。

连续领域语义分割这一实验设置是 PLOP[15] 提出的。与连续类别语义分割不同，这一设置目的在于处理深度学习中的领域漂移现象，而不是聚合新类别。在真实应用中，领域漂移可能会经常发生。在这一设置中，本文假设所有领域中的类别都是相同的，本文不考虑不同领域中的类别之间的不同。在当前学习步骤中，模型在新领域数据上进行训练，而之前学习步骤中的旧领域数据是不可见的。本文在 Cityscapes[181] 数据集上进行了连续域语义分割实验。与 PLOP[15] 保持相同的设置，本文将来自于每个城市的训练数据作为一个域数据。本文也采用了三个不同的设置，11-5（共分为 3 个步骤），11-1（共分为 11 个步骤），以及 1-1（共分为 21 个步骤）。在这些实验设置中，本文使用了和连续类别语义分割相同的记号表示。不同的是每个学习步骤添加的是新的领域数据（城市），而不是类别。

4.3.1.3 实现细节

与之前的方法保持一致的设置[15, 16, 125]，本文使用 ResNet-101[2] 作为主干网络的 Deeplab-v3[182] 作为语义分割模型。该模型的降采样倍数被设置为 16 倍。与上述方法保持一致的设置，本文也在主干网络中采用了 IABN[183]，并且将主干网络在 ImageNet[139] 数据集上进行了预训练。本文利用 MiB[16] 提出的损失函数来辅助模型的训练过程。并且，本文应用了和其他方法[15, 16, 125] 相同的训练策略。具体来说，本文采用了相同的数据增强方式，例如水平翻转和随机裁剪。在所有的实验中，本文将批次大小设置为 24。本文将第一个学习步骤中的初始学习率设置为 0.02，之后的连续学习步骤的学习率设置为 0.001。在每个训练过程中，模型采用 poly 学习率衰减策略。在每个步骤中，本文使用 SGD 优化器对模型分别训练 30 个轮次（PASCAL VOC 2012[179] 数据集），50 个轮次（Cityscapes[181] 数据集）和 60 个轮次（ADE20K[180]）。与许多方法相同[15, 16, 125]，本文也使用了训练集中的 20% 作为验证集。本文在数据集的原始的验证集上报



(a) 15-1 非重叠 (b) 15-1 重叠 (c) 1-1 连续领域分割

图 4.6 在三种不同的实验设置下的连续学习过程中的不同方法的表现。(a)(b) 是连续类别语义分割的两种不同设置。(c) 是连续领域分割的一种设置。

告平均交并比（mIoU）作为评价指标。

表 4.1 不同方法在不同连续学习设置下的 mIoU(%)。所有的实验都是在 Pascal VOC 2012 数据集上进行的。表格中使用加粗字体来表示最高的精度。

	15-5 非重叠			**15-1 非重叠**			**10-1 非重叠**		
方法	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*
微调	5.7	33.6	12.3	4.6	1.8	3.8	6.3	1.1	3.8
联合训练	79.8	72.6	78.2	79.8	72.6	78.2	78.2	78.0	78.2
LwF[66]	60.4	37.4	54.9	5.8	3.6	5.3	7.2	1.2	4.3
ILT[174]	64.9	39.5	58.9	8.6	5.7	7.9	7.3	3.2	5.4
MiB[16]	73.0	43.3	65.9	48.4	12.9	39.9	9.5	4.1	6.9
SDR[125]	74.6	44.1	**67.3**	59.4	14.3	48.7	17.3	11.0	14.3
PLOP[15]	71.0	42.8	64.3	57.9	13.7	46.5	9.7	7.0	8.4
Ours	75.0	42.8	**67.3**	66.1	18.2	**54.7**	30.6	4.7	**18.2**

表 4.2 不同方法在不同连续学习设置下的 mIoU(%)。所有的实验都是在 Pascal VOC 2012 数据集上进行的。表格中使用加粗字体来表示最高的精度。

	15-5 重叠			**15-1 重叠**			**10-1 重叠**		
方法	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*	*0-15*	*16-20*	*all*
微调	6.6	33.1	12.9	4.6	1.8	3.9	6.4	1.2	3.9
联合训练	79.8	72.6	78.2	79.8	72.6	78.2	78.2	78.0	78.2
LwF[66]	60.8	36.6	55.0	6.0	3.9	5.5	8.0	2.0	4.8
ILT[174]	67.8	40.6	61.3	9.6	7.8	9.2	7.2	3.7	5.5
MiB[16]	76.4	49.4	70.0	38.0	13.5	32.2	20.0	20.1	20.1
SDR[125]	76.3	50.2	70.1	47.3	14.7	39.5	32.4	17.1	25.1
PLOP[15]	75.7	51.7	70.1	65.1	21.1	54.6	44.0	15.5	30.5
Ours	78.8	52.0	**72.4**	70.6	23.7	**59.4**	55.4	15.1	**34.3**

表 4.3 不同方法在 ADE20K 数据集上的在不同设置下的实验结果。表格中使用加粗来表示最高的表现。

	100-50 重叠			**100-10 重叠**			**50-50 重叠**			
方法	*1-100*	*101-150*	*all*	*1-100*	*101-150*	*all*	*1-50*	*51-100*	*101-150*	*all*
ILT[174]	18.3	14.8	17.0	0.1	2.9	1.1	13.6	12.3	0.0	9.7
MiB[16]	40.7	17.7	32.8	38.3	11.3	29.2	45.3	26.1	17.1	29.3
PLOP[15]	41.9	14.9	32.9	40.6	14.1	31.6	48.6	30.0	13.1	30.4
Ours	42.3	18.8	**34.5**	39.3	17.6	**32.1**	48.3	31.3	18.7	**32.5**
联合训练	44.3	28.2	38.9	44.3	28.2	38.9	51.1	38.3	28.2	38.9

表 4.4 不同方法在 ADE20K 数据集上 100-5 重叠设置下的实验结果。

方法	*1-100*	*101-150*	*all*
ILT[174]	0.1	1.3	0.5
MiB[16]	36.0	5.6	25.9
PLOP[15]	39.1	7.8	28.7
Ours	38.5	11.5	**29.6**

4.3.2 连续类别语义分割

PASCAL VOC 2012. 与之前的方法[15, 16, 125]采用相同的实验设置，本文在不同的连续学习设置下进行了实验，包括 15-5，15-1 和 10-1 三种不同的实验设置。如 表 4.1与 表 4.2所示，本文报告了在连续学习的最后一个步骤的精度。简单的微调方法会导致发生灾难性遗忘，模型会很快忘记旧的知识，并且无法很好地学习新知识。实验结果表明本文的方法在重叠和非重叠两种设置上都具有很大的分割表现的提升。尤其是在富有挑战性的 15-1 设置下，本文的方法在 mIoU 指标上分别优于现有方法最高的精度 6.0% 和 4.8%。如 图 4.6(a)和 图 4.6(b)中所示，本文也展示了不同方法在整个 15-1 的学习过程中每个步骤的精度。这表明本文的方法在连续学习过程中可以减少对旧知识的遗忘。在 表 4.1与 表 4.2中，本文也分别报告了在旧类别和新类别上的模型表现。从该表中可以看出，在所有不同的设置下，模型的旧类别上的表现得到了很大的提升。这得益于本文提出的表征补偿机制和知识蒸馏方法，这两个模块都可以有效地缓解模型对于旧知识的遗忘。在另外一方面，本文提出的表征补偿机制和蒸馏机制也允许模型具有学习新知识的空间。在 节 4.3.4中，本文将会进一步分析这样两个机制的有效性。本文进一步在 图 4.7中分析了定性结果，展示了不同方法在 15-1 重叠设置下的分割结果的表现。

表 4.5 不同方法在连续领域语义分割的不同设置下的实验结果。

方法	11-5	11-1	1-1
微调	61.7	60.4	42.9
LwF[66]	59.7	57.3	33.0
LwF-MC[67]	58.7	57.0	31.4
ILT[174]	59.1	57.8	30.1
MiB[16]	61.5	60.0	42.2
PLOP[15]	63.5	62.1	45.2
Ours	**64.3**	**63.0**	**48.9**

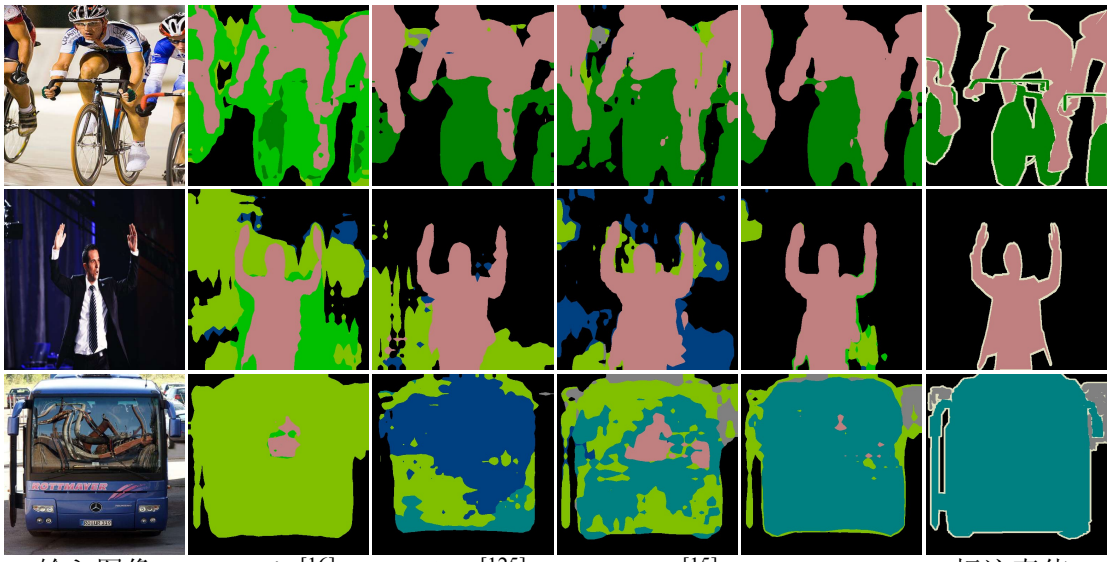
ADE20K. 为了进一步验证本文方法的有效性，本文在一个具有挑战性的语义分割数据集，ADE20K[180] 上进行了实验。实验结果展示在 表 4.3和 表 4.4中。在三个不同的连续学习任务设置中，100-50，100-10 和 50-50，本文的方法实现了相对于目前现有方法最高精度的平均 1.4% 的提升。本文也在一个更具有挑战性的设置，100-5 重叠设置下进行了实验。如 表 4.4所示，在这样一个包含有 11 个学习步骤的设置下，本文的方法也实现了最先进的表现，在 mIoU 指标上，比之前的方法提升了 0.9%。这样的提升得益于本文提出的表征补偿机制和蒸馏方式。

4.3.3 连续领域语义分割

在连续语义分割的情况下，模型除了需要能够处理新的类别，增加对新的领域数据的处理能力也是非常重要的。与 PLOP[15] 保持相同的设置，本文在 Cityscapes[181] 数据集上进行了连续领域语义分割的实验。本文将 Cityscapes[181] 中的一个城市中的数据看作一个领域，这样的设置在领域自适应语义分割[184] 中得到了广泛的使用。在这种连续领域语义分割的场景下，本文不再考虑不同领域中类别之间的不同。如 表 4.5所示，实验结果表明本文的方法在所有的三个设置下，实现了比之前的方法[15, 16, 174] 更高的 mIoU。在具有 21 个连续学习步骤的 1-1 设置下，本文的方法比之前最高的精度提升了 3.7%。本文也在 图 4.6(c)中展示了这一设置的每一步的模型的精度。由于 MiB[16] 目的在于解决连续学习过程中的语义漂移的问题，这一问题在连续领域语义分割中并不存在，所以 MiB[16] 会具有比微调稍微差点的表现。这些实验表明本文的方法对于连续领域语义分割也是非常有效的，这得益于本文方法的保留旧知识和允许学习新知识的特点。

表 4.6 关于表征补偿机制和在空间和通道上进行的池化立体蒸馏机制的消融实验。所有实验都在 15-1 重叠设置下在 Pascal VOC 2012 数据集上进行。† 表示基线方法通过 PLOP[15] 中提出的自适应因子进行提升的表现。

MiB†[16]	表示补偿机制	条带池化[111]	空间知识蒸馏	通道知识蒸馏	15-1
✓					36.1
✓	✓				43.0
✓	✓		✓		58.3
✓	✓			✓	58.4
✓			✓	✓	57.8
✓	✓	✓			57.9
✓	✓		✓	✓	**59.4**



输入图像 MiB[16] SDR[125] PLOP[15] Ours 标注真值

图 4.7 不同方法的预测结果的定性比较。所有的实验都在 15-1 重叠设置下进行。

4.3.4 消融实验

在这一小节, 本文首先分析了本文提出的表示补偿和池化立体蒸馏机制的有效性。之后，本文讨论了该方法在连续学习场景下对类别顺序的鲁棒性。

表征补偿机制。本文在 PASCAL VOC 2012[179] 数据集上进行了实验。如 表 4.6所示, 本文提出的表征补偿机制可以在 MiB[16] 基准方法上实现大约 7% 的提升。本文也研究了提出的知识蒸馏机制，实验结果表明在空间上和通道上同时使用蒸馏机制会具有更高的表现。之后，本文在使用了知识蒸馏的一个强的基准方法上消融研究了本文提出的表征补偿机制，最后使用这样一个模块，本文的方法实现了目前最高的精度。本文认为这样一个最高的精度得益

表 4.7 表征补偿机制中各个操作的消融实验。

并行卷积	合并操作	冻结操作	Drop-path	15-1
✓				40.1
✓	✓			42.0
✓	✓	✓		42.8
✓	✓	✓	✓	**43.0**

表 4.8 蒸馏机制中不同的池化方式的对比。所有的实验都在 15-1 重叠设置上进行。

不使用池化	全局平均池化	最大池化	条带池化[111]	平均池化
52.0	36.1	48.0	54.6	**56.1**

于本文提出的方法具有记忆旧知识和允许学习新知识的能力。在本文的方法中，合并卷积层和冻结参数的操作目的在于缓解模型对于旧知识的遗忘。于是，在 表 4.7中，本文进一步研究了这样两个操作的有效性。具体来说，基于一个简单的并行的两个卷积分支（Parallel-Conv），合并卷积层和冻结参数的两个操作可以取得 2.7% 的提升。实验结果表明模型可以从之前学习步骤中冻结的知识中受益。

知识蒸馏机制。在 表 4.6中，本文首先分别研究了在空间和通道两个维度上进行知识蒸馏的重要性。本文发现，在空间和通道两个维度上进行知识蒸馏具有相似的表现，能够在基准方法上提升大约 15.3% 的 mIoU。本文同时在空间和通道维度上使用知识蒸馏，再与本文提出的表征补偿机制联合使用，可以实现目前最高的精度。如 表 4.8所示，本文进一步比较了在知识蒸馏算法中使用的不同的池化方式的有效性。实验结果表明平均池化能够实现比条带池化的 1.5% 的提升。

其次，本文研究了提出的方法中不同的池化核大小的影响。在连续语义分割的场景下，知识蒸馏机制中的池化操作起到了非常关键的作用。在本文提出的知识知识蒸馏方法中，本文使用多尺度的平均池化操作，池化核的大小在集合 $\mathcal{M} = \{4,8,12,16,20,24\}$ 中进行选取。如 表 4.9所示，本文研究了不同的池化核大小的影响。实验结果表明如果只使用一个太小或者太大的池化窗口，就会导致最后的效果比较差。本文分析是因为当池化窗口尺寸比较小的时候，在为当前像素聚合信息的时候，无法考虑到附近的足够的信息，所以噪声带来的负面影响无法被有效抑制。当池化窗口尺寸比较大的时候，在为当前像素聚合信息的时候，会为当前像素带来不相关的噪声，这样也会导致模型的效果比较差。

表 4.9 本文提出的池化立体蒸馏机制中不同的平均池化核的尺寸的对比。

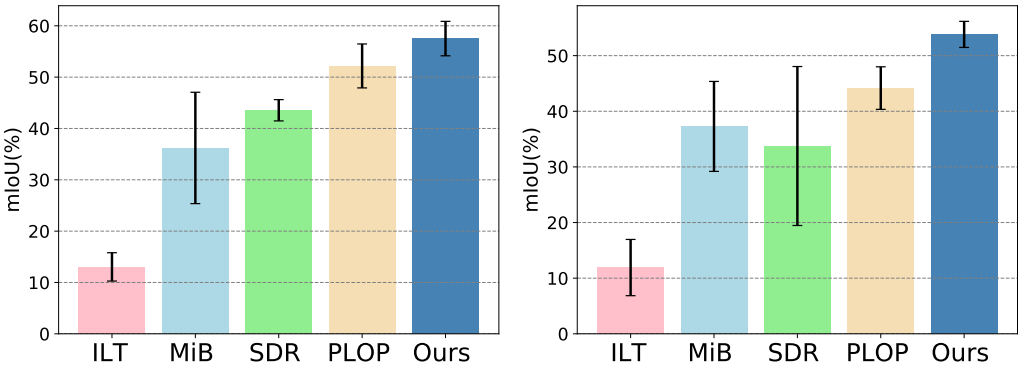
4	8	12	16	20	24	mIoU(%)
✓						55.1
	✓					**56.2**
		✓				**56.2**
			✓			55.4
				✓		54.7
					✓	53.7
✓	✓					55.8
✓	✓	✓				56.1
✓	✓	✓	✓			**56.2**
✓	✓	✓	✓	✓		56.1
✓	✓	✓	✓	✓	✓	56.1

表 4.10 在网络不同的阶段进行知识蒸馏的消融实验。

layer 1	*layer 2*	*layer 3*	*layer 4*	*decoder*	15-1
					36.1
✓					33.6
	✓				34.0
		✓			39.7
			✓		47.2
				✓	54.1
✓	✓				32.8
✓	✓	✓			34.0
✓	✓	✓	✓		46.6
			✓	✓	55.3
		✓	✓	✓	56.6
	✓	✓	✓	✓	57.4
✓	✓	✓	✓	✓	**57.8**

如 表 4.9所示，当本文使用多尺度窗口进行池化时，模型的表现可以稳定在一个比较高的精度，于是本文使用了所有的尺度来进行平均池化。

之后，本文进一步研究了在模型的不同的隐藏层进行知识蒸馏对模型表现的影响。实验结果展示在 表 4.10。从该表中可以发现，当知识蒸馏方法加在所有的层上时，模型能够比不使用知识蒸馏方法时在 mIoU 上获得 21.7% 的精度提升。一个很有意思的发现是当在解码器上加入知识蒸馏之后，会给模型带来最大的提升，这是解码器中的高级语义特征带来的提升。由于在深度监督的情况下，模型的梯度消失的问题能够得到有效的缓解，所以在所有层上进行知识蒸馏能够进一步提升模型的性能。于是本文在所有层上使用了知识蒸馏。



(a) 15-1 重叠　　(b) 15-1 非重叠

图 4.8　在不同的类别顺序下，不同方法的平均精度及其标准差。

表 4.11　不同超参数的影响。所有的实验都是在 15-1 重叠设置下进行的。本文最终选择超参数 λ 为 100，以及 γ 为 0.01。

λ \ γ	0.0001	0.001	0.005	0.01	0.05	0.1
1	35.4	39.8	46.3	49.3	46.5	42.8
10	44.3	49.3	52.1	51.0	46.5	44.7
20	49.0	56.9	57.6	56.1	50.0	47.8
50	48.5	57.4	**59.7**	59.1	53.6	50.6
100	42.9	55.0	**59.4**	**59.4**	55.5	50.8
150	52.6	52.6	58.2	58.9	55.4	50.7
200	50.0	50.0	57.8	58.3	55.1	51.0

对于类别顺序的鲁棒性研究。在连续语义分割的场景下，不同的类别顺序可能会对模型最终的精度带来比较大的影响。为了验证本文提出的算法对类别顺序的鲁棒性，本文在五种不同的类别顺序上进行了实验，包括四个随机的顺序和之前使用的升序顺序。在 图 4.8中，本文展示了对于不同方法[15, 16, 125] 的 mIoU 的平均表现和标准差。实验结果表明，本文的方法比之前的方法在不同顺序上更加鲁棒。

超参数的消融实验。在本文的方法中，采用了与 MiB[16] 相同的优化目标，共有两个优化目标中的超参数，分别是 λ 和 γ。在 表 4.11中，本文研究了这样两个超参数对模型的结果的影响。本文发现当 $\gamma = 0.005$ 和 $\gamma = 0.01$ 时，本文的方法取得了最好的表现。为了公平比较，在本文的实验中，考虑到 *lambda* 在 MiB[16] 中被设置为 100，于是本文采用了和 MiB[16] 相同的实验设置。最终，本文将 λ 设置为 100，并且将 γ 设置为 0.01。

第四节 本章小结

在连续学习场景下，为了保留住旧类别知识的同时，能够创造更多学习新类别的空间，本章提出了基于重参数化的表征补偿机制，能够动态地对模型进行拓展，而不增加任何的额外的推理代价。除此之外，为了进一步缓解对旧知识的遗忘，本章提出了一种同时在空间和通道上对模型进行知识蒸馏的方法。之后，为了进一步验证该方法的有效性，本文在三个常用的公开语义分割数据集上进行了连续语义分割的实验，均取得了比现有方法更高的精度。本文之后也进行了消融实验，来探讨表征补偿机制中对旧知识进行冻结的重要性。

第五章 总结展望

第一节 本文工作总结

如何高效、有效地进行知识迁移是深度学习中非常重要的研究问题。本文从单次学习和连续学习两种场景下探讨了知识迁移存在的问题，以及提出了解决思路，并进行了实验验证和分析。

在单次学习场景下，基于全监督的深度神经网络在许多任务上已经取得了重大的突破性进展，这依赖于数据的规模以及模型的拟合能力与泛化能力。由于实际应用中的边缘设备会对模型有众多的限制条件，比如参数量大小、功耗、算子类型等等，所以目前利用知识迁移技术将大模型的知识迁移到小模型上，受到了许多研究人员的关注。然而，现有知识迁移技术具有较低的灵活性，往往需要预训练好的教师模型或者依赖于特定的网络结构。另一方面，部署的模型在实际应用中面临着可拓展性的问题。由于单次学习场景下的深度模型在推理时只能处理固定的预设好的类别或者领域，所以在连续学习场景下，如何有效地将旧模型的知识迁移给新模型也具有重要的意义。然而，现有连续学习场景下的知识迁移方法无法将新旧知识解耦，难以在模型的稳定性和可塑性之间进行权衡。针对这两个场景下的知识迁移方法面临的问题，本文从在线标签平滑和表征补偿机制出发，设计了两种不同的在线知识迁移的策略，分别用于在单次学习场景和连续学习场景下的知识迁移。

具体来说，在单次学习场景下，为了提升模型的性能，可以在训练过程中将中间模型的知识迁移给当前正在训练的模型。该方法是一个在线过程，不需要存储中间模型，只需要极小的存储代价。在训练过程中，该方法会为每个类别维护一个类别级的软标签，用于监督模型训练。这个软标签是在训练过程中得到的，将所有预测正确的训练图像的预测概率累积到对应的类别。经过若干次迭代累积得到的软标签就可以用来监督模型的训练，同时，在新的训练阶段也会使用相同方法来累积新的软标签。相比于之前的标签平滑方法，本文提出的方法利用到了类别之间的关系，能够在时序上将之前训练好的模型的知识迁移给当前模型。并且，该方法具有比自蒸馏方法更高的灵活性和易用性。

在连续学习的场景下，本文提出了利用重参数化机制来解耦模型对于旧知识的记忆和对于新知识的学习，同时自适应地使得模型的旧知识适配到现有模型中，完成了隐式的自适应迁移。该方法能够使得基于知识蒸馏的方法在保持旧知识和学习新知识之间能够找到更好的平衡。相比于现有基于模型拓展的连续学习方法，能够避免因为连续学习步骤带来的模型规模越来越大的问题，不会带来额外的计算代价。本文在公开的图像分类、语义分割数据集上进行了广泛的实验，验证了本文所提出的算法的表现，并对实验结果进行了分析和讨论。

总之，本文针对目前单次学习和连续学习两种场景，从两个不同的角度设计了在线知识迁移的策略，能够为之后的相关研究工作以及实际应用提供启发。

第二节 未来工作展望

如何有效地进行知识迁移，无论在单次学习场景下，还是在连续学习场景下都具有重要的作用。本文在这两种不同场景下，分别设计了两种不同的在线的知识迁移策略，能够提升模型的表现。然而，这两种策略还有许多待改进的研究方向。对于基于在线标签平滑的在线知识迁移方式，在前期训练过程中面临着累积正确软标签的样本数太少的问题，所以用来约束训练的正确样本的数量较少，导致收敛比较慢。除此之外，利用数据去构建相关的类别级标签能够考虑到类别之间的相关性，从而引入了类内约束，但是本文所采用的累积预测的方式忽略掉了不同实例之间的差异性。在之后的研究中，考虑如何构建不同实例之间的区分性具有重要的研究意义。另一方面，基于表征补偿的在线知识迁移策略是一种自适应的方法，能够使得模型对于新知识的学习自适应地适配到旧的模式上。然而，这种方式本质上并不能提供更好的优化目标，仍然依赖于知识蒸馏方式来提升性能，所以将来如何能在整个框架中把显式的知识蒸馏去掉是一个非常有意义的研究方向。总之，希望本文工作可以带给领域内的研究人员一些启发，推动知识迁移算法的发展和进步。

参考文献

[1] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 4700–4708.

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 770–778.

[3] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39 (6): 1137–1149.

[5] LIU F, SHEN C, LIN G. Deep convolutional neural fields for depth estimation from a single image. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015.

[6] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. [C] // Eur. Conf. Comput. Vis.

[7] CHEN C.-F, FAN Q, PANDA R. Crossvit: Cross-attention multi-scale vision transformer for image classification. [C] // Int. Conf. Comput. Vis. 2021.

[8] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network. [C] // Adv. Neural Inform. Process. Syst. Worksh. 2015.

[9] 高钦泉, 赵岩, 李根, et al. 基于知识蒸馏的超分辨率卷积神经网络压缩方法. [J]. 计算机应用, 2019.

[10] GUO C, PLEISS G, SUN Y, et al. On calibration of modern neural networks. [C] // Int. Conf. Mech. Learn. 2017: 1321–1330.

[11] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018.

[12] ZHANG L, SONG J, GAO A, et al. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. [C] // Int. Conf. Comput. Vis. 2019: 3712–3721.

[13] XU T.-B, LIU C.-L. Data-Distortion Guided Self-Distillation for Deep Neural Networks. [C] // AAAI Conf. Artif. Intell. 2019: 5565–5572.

[14] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020.

[15] DOUILLARD A, CHEN Y, DAPOGNY A, et al. PLOP: Learning without Forgetting for Continual Semantic Segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[16] CERMELLI F, MANCINI M, BULO S R, et al. Modeling the background for incremental learning in semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 9233–9242.

[17] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay. [C] // Adv. Neural Inform. Process. Syst. 2017.

[18] VU T.-H, JAIN H, BUCHER M, et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019.

[19] LIU Y, SCHIELE B, SUN Q. Adaptive Aggregation Networks for Class-Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[20] ABATI D, TOMCZAK J, BLANKEVOORT T, et al. Conditional channel gated networks for task-aware continual learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 3931–3940.

[21] GE S, LUO Z, ZHANG C, et al. Distilling Channels for Efficient Deep Tracking. [J]. IEEE Trans. Image Process., 2020, 29: 2610–2621.

[22] WANG T, YUAN L, ZHANG X, et al. Distilling object detectors with fine-grained feature imitation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 4933–4942.

[23] MCCLOSKEY M, COHEN N J. Catastrophic interference in connectionist networks: The sequential learning problem. [G] // PsychologLearniny of learning and motivation. Vol. 24. Elsevier, 1989: 109–165.

[24] DOUILLARD A, CORD M, OLLION C, et al. Podnet: Pooled outputs distillation for small-tasks incremental learning. [C] // Eur. Conf. Comput. Vis. Vol. 12365. 2020: 86–102.

[25] REED S E, LEE H, ANGUELOV D, et al. Training Deep Neural Networks on Noisy Labels with Bootstrapping. [C] // Int. Conf. Learn. Represent. Worksh. 2015.

[26] XIE L, WANG J, WEI Z, et al. DisturbLabel: Regularizing CNN on the Loss Layer. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 4753–4762.

[27] DUBEY A, GUPTA O, GUO P, et al. Pairwise confusion for fine-grained visual classification. [C] // Eur. Conf. Comput. Vis. 2018: 70–86.

[28] LI C, LIU C, DUAN L, et al. Reconstruction Regularized Deep Metric Learning for Multi-Label Image Classification. [J]. IEEE Trans. Neural Netw. Learn Syst., 2020, 31 (7): 2294–2303.

[29] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 2818–2826.

[30] ZHANG L, QI G.-J, WANG L, et al. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 2547–2555.

[31] QI G.-J, ZHANG L, LIN F, et al. Learning generalized transformation equivariant representations via autoencoding transformations. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[32] QI G.-J, ZHANG L, CHEN C W, et al. Avt: Unsupervised learning of transformation equivariant representations by autoencoding variational transformations. [C] // Int. Conf. Comput. Vis. 2019: 8130–8139.

[33] WANG X, KIHARA D, LUO J, et al. EnAET: A Self-Trained framework for Semi-Supervised and Supervised Learning with Ensemble Transformations. [J]. IEEE Trans. Image Process., 2020.

[34] YAO J, WANG J, TSANG I W, et al. Deep Learning From Noisy Image Labels With Quality Embedding. [J]. IEEE Trans. Image Process., 2019, 28 (4): 1909–1922.

[35] DUNCAN J S, BIRKHOLZER T. Reinforcement of linear structure using parametrized relaxation labeling. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 1992, 14 (5): 502–515.

[36] WANG R, LIU T, TAO D. Multiclass Learning With Partially Corrupted Labels. [J]. IEEE Trans. Neural Netw. Learn Syst., 2018, 29 (6): 2568–2580.

[37] WEI Y, GONG C, CHEN S, et al. Harnessing Side Information for Classification Under Label Noise. [J]. IEEE Trans. Neural Netw. Learn Syst., 2020, 31 (9): 3178–3192.

[38] HAN B, TSANG I W, CHEN L, et al. Progressive Stochastic Learning for Noisy Labels. [J]. IEEE Trans. Neural Netw. Learn Syst., 2018, 29 (10): 5136–5148.

[39] TANAKA D, IKAMI D, YAMASAKI T, et al. Joint Optimization Framework for Learning with Noisy Labels. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 5552–5560.

[40] HAN J, LUO P, WANG X. Deep self-learning from noisy labels. [C] // Int. Conf. Comput. Vis. 2019: 5138–5147.

[41] REN M, ZENG W, YANG B, et al. Learning to Reweight Examples for Robust Deep Learning. [C] // Int. Conf. Mech. Learn. 2018: 4334–4343.

[42] SHU J, XIE Q, YI L, et al. Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting. [C] // Adv. Neural Inform. Process. Syst. 2019: 1919–1930.

[43] LIU T, TAO D. Classification with noisy labels by importance reweighting. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2015, 38 (3): 447–461.

[44] WANG Y, MA X, CHEN Z, et al. Symmetric cross entropy for robust learning with noisy labels. [C] // Int. Conf. Comput. Vis. 2019: 322–330.

[45] TANNO R, SAEEDI A, SANKARANARAYANAN S, et al. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 11236–11245.

[46] ARAZO E, ORTEGO D, ALBERT P, et al. Unsupervised Label Noise Modeling and Loss Correction. [C] // Int. Conf. Mech. Learn. 2019: 312–321.

[47] ZHANG J, SHENG V S, LI T, et al. Improving Crowdsourced Label Quality Using Noise Correction. [J]. IEEE Trans. Neural Netw. Learn Syst., 2018, 29 (5): 1675–1688.

[48] FANG M, ZHOU T, YIN J, et al. Data Subset Selection With Imperfect Multiple Labels. [J]. IEEE Trans. Neural Netw. Learn Syst., 2019, 30 (7): 2212–2221.

[49] YI K, WU J. Probabilistic end-to-end noise correction for learning with noisy labels. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 7017–7025.

[50] DING X, GUO Y, DING G, et al. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. [C] // Int. Conf. Comput. Vis. 2019.

[51] DING X, ZHANG X, MA N, et al. RepVGG: Making VGG-style ConvNets Great Again. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[52] KIM C D, JEONG J, MOON S, et al. Continual Learning on Noisy Data Streams via Self-Purified Replay. [C] // Int. Conf. Comput. Vis. 2021: 537–547.

[53] BANG J, KIM H, YOO Y, et al. Rainbow Memory: Continual Learning with a Memory of Diverse Samples. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[54] CHAUDHRY A, GORDO A, DOKANIA P K, et al. Using hindsight to anchor past knowledge in continual learning. [C] // AAAI Conf. Artif. Intell. 2021.

[55] SMITH J, HSU Y.-C, BALLOCH J, et al. Always Be Dreaming: A New Approach for Data-Free Class-Incremental Learning. [C] // Int. Conf. Comput. Vis. 2021.

[56] BELOUADAH E, POPESCU A. Il2m: Class incremental learning with dual memory. [C] // Int. Conf. Comput. Vis. 2019: 583–592.

[57] VERWIMP E, DE LANGE M, TUYTELAARS T. Rehearsal Revealed: The Limits and Merits of Revisiting Samples in Continual Learning. [C] // Int. Conf. Comput. Vis. 2021: 9385–9394.

[58] CHA H, LEE J, SHIN J. Co2l: Contrastive continual learning. [C] // Int. Conf. Comput. Vis. 2021: 9516–9525.

[59] SHIM D, MAI Z, JEONG J, et al. Online Class-Incremental Continual Learning with Adversarial Shapley Value. [C] // AAAI Conf. Artif. Intell. 2021.

[60] BUZZEGA P, BOSCHINI M, PORRELLO A, et al. Dark experience for general continual learning: a strong, simple baseline. [C] // Adv. Neural Inform. Process. Syst. 2020.

[61] ZHU F, ZHANG X.-Y, WANG C, et al. Prototype Augmentation and Self-Supervision for Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 5871–5880.

[62] ZHU K, CAO Y, ZHAI W, et al. Self-Promoted Prototype Refinement for Few-Shot Class-Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 6801–6810.

[63] HAYES T L, KAFLE K, SHRESTHA R, et al. Remind your neural network to prevent catastrophic forgetting. [C] // Eur. Conf. Comput. Vis. 2020: 466–483.

[64] MARACANI A, MICHIELI U, TOLDO M, et al. RECALL: Replay-based Continual Learning in Semantic Segmentation. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7026–7035.

[65] CHAUDHRY A, DOKANIA P K, AJANTHAN T, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence. [C] // Eur. Conf. Comput. Vis. 2018.

[66] LI Z, HOIEM D. Learning without forgetting. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 40 (12): 2935–2947.

[67] REBUFFI S.-A, KOLESNIKOV A, SPERL G, et al. Icarl: Incremental classifier and representation learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[68] SIMON C, KONIUSZ P, HARANDI M. On Learning the Geodesic Path for Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[69] CHERAGHIAN A, RAHMAN S, FANG P, et al. Semantic-aware Knowledge Distillation for Few-Shot Class-Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[70] DHAR P, SINGH R V, PENG K.-C, et al. Learning without memorizing. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019.

[71] EBRAHIMI S, MEIER F, CALANDRA R, et al. Adversarial Continual Learning. [C] // Adv. Neural Inform. Process. Syst. 2020.

[72] XIANG Y, FU Y, JI P, et al. Incremental learning using conditional adversarial networks. [C] // Int. Conf. Comput. Vis. 2019: 6619–6628.

[73] ZENKE F, POOLE B, GANGULI S. Continual learning through synaptic intelligence. [C] // Int. Conf. Mech. Learn. 2017.

[74] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks. [J]. Proceedings of the national academy of sciences, 2017, 114 (13): 3521–3526.

[75] PAN P, SWAROOP S, IMMER A, et al. Continual deep learning by functional regularisation of memorable past. [C] // Adv. Neural Inform. Process. Syst. 2020.

[76] ISCEN A, ZHANG J, LAZEBNIK S, et al. Memory-efficient incremental learning through feature adaptation. [C] // Eur. Conf. Comput. Vis. 2020: 699–715.

[77] LIU Y, PARISOT S, SLABAUGH G, et al. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. [C] // Eur. Conf. Comput. Vis. 2020: 699–716.

[78] TAO X, CHANG X, HONG X, et al. Topology-preserving class-incremental learning. [C] // Eur. Conf. Comput. Vis. 2020: 254–270.

[79] YU L, TWARDOWSKI B, LIU X, et al. Semantic drift compensation for class-incremental learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 6982–6991.

[80] PARK D, HONG S, HAN B, et al. Continual learning by asymmetric loss approximation with single-side overestimation. [C] // Int. Conf. Comput. Vis. 2019: 3335–3344.

[81] VERMA V K, LIANG K J, MEHTA N, et al. Efficient Feature Transformations for Discriminative and Generative Continual Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[82] SINGH P, MAZUMDER P, RAI P, et al. Rectification-based Knowledge Retention for Continual Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 15282–15291.

[83] YAN S, XIE J, HE X. DER: Dynamically Expandable Representation for Class Incremental Learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[84] SINGH P, VERMA V K, MAZUMDER P, et al. Calibrating CNNs for Lifelong Learning. [C] // Adv. Neural Inform. Process. Syst. Vol. 33. 2020.

[85] KANAKIS M, BRUGGEMANN D, SAHA S, et al. Reparameterizing Convolutions for Incremental Multi-Task Learning without Task Interference. [C] // Eur. Conf. Comput. Vis. 2020: 689–707.

[86] JUNG S, AHN H, CHA S, et al. Continual Learning with Node-Importance based Adaptive Group Sparse Regularization. [C] // Adv. Neural Inform. Process. Syst. 2020.

[87] WU G, GONG S, LI P. Striking a Balance Between Stability and Plasticity for Class-Incremental Learning. [C] // Int. Conf. Comput. Vis. 2021: 1124–1133.

[88] ZHANG C, SONG N, LIN G, et al. Few-Shot Incremental Learning with Continually Evolved Classifiers. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[89] LIU Q, MAJUMDER O, ACHILLE A, et al. Incremental Meta-Learning via Indirect Discriminant Alignment. [C] // Eur. Conf. Comput. Vis. 2020.

[90] KIM C D, JEONG J, KIM G. Imbalanced continual learning with partitioning reservoir sampling. [C] // Eur. Conf. Comput. Vis. 2020: 411–428.

[91] ZHAO B, XIAO X, GAN G, et al. Maintaining discrimination and fairness in class incremental learning. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 13208–13217.

[92] HOU S, PAN X, LOY C C, et al. Learning a unified classifier incrementally via rebalancing. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 831–839.

[93] KOLTUN V, et al. Efficient inference in fully connected crfs with gaussian edge potentials. [C] // Adv. Neural Inform. Process. Syst. 2011.

[94] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional random fields as recurrent neural networks. [C] // Int. Conf. Comput. Vis. 2015.

[95] ARNAB A, JAYASUMANA S, ZHENG S, et al. Higher order conditional random fields in deep neural networks. [C] // Eur. Conf. Comput. Vis. 2016.

[96] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015.

[97] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Hypercolumns for object segmentation and fine-grained localization. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 447–456.

[98] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2015: 1520–1528.

[99] LIN D, JI Y, LISCHINSKI D, et al. Multi-scale context intertwining for semantic segmentation. [C] // Eur. Conf. Comput. Vis. 2018: 603–619.

[100] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[101] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39 (12): 2481–2495.

[102] PENG C, ZHANG X, YU G, et al. Large kernel matters–improve semantic segmentation by global convolutional network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 4353–4361.

[103] TIAN Z, HE T, SHEN C, et al. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 3126–3135.

[104] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 7794–7803.

[105] LIU S, DE MELLO S, GU J, et al. Learning affinity via spatial propagation networks. [C] // Adv. Neural Inform. Process. Syst. 2017.

[106] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 9167–9176.

[107] DING H, JIANG X, SHUAI B, et al. Context contrasted feature and gated multi-scale aggregation for scene segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 2393–2402.

[108] CHEN L.-C, YANG Y, WANG J, et al. Attention to scale: Scale-aware semantic image segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016.

[109] HONG S, OH J, LEE H, et al. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 3204–3212.

[110] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2019: 3146–3154.

[111] HOU Q, ZHANG L, CHENG M.-M, et al. Strip pooling: Rethinking spatial pooling for scene parsing. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 4003–4012.

[112] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers. [C] // Eur. Conf. Comput. Vis. 2020: 213–229.

[113] ZENG Y, FU J, CHAO H. Learning Joint Spatial-Temporal Transformations for Video Inpainting. [C] // Eur. Conf. Comput. Vis. 2020: 528–543.

[114] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. [C] // Int. Conf. Learn. Represent. 2021.

[115] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection. [C] // Int. Conf. Learn. Represent. 2021.

[116] WANG Y, XU Z, WANG X, et al. End-to-End Video Instance Segmentation with Transformers. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[117] ZHENG S, LU J, ZHAO H, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[118] XIE E, WANG W, YU Z, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. [C] // Int. Conf. Comput. Vis. 2021.

[119] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. [C] // Int. Conf. Comput. Vis. 2021.

[120] ZHANG D, ZHANG H, TANG J, et al. Feature pyramid transformer. [C] // Eur. Conf. Comput. Vis. 2020: 323–339.

[121] STRUDEL R, GARCIA R, LAPTEV I, et al. Segmenter: Transformer for Semantic Segmentation. [C] // Int. Conf. Comput. Vis. 2021.

[122] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. [C] // Int. Conf. Comput. Vis. 2021.

[123] YAN S, ZHOU J, XIE J, et al. An EM Framework for Online Incremental Learning of Semantic Segmentation. [C] // ACM Int. Conf. Multimedia. 2021.

[124] HUANG Z, HAO W, WANG X, et al. Half-Real Half-Fake Distillation for Class-Incremental Semantic Segmentation. [J]. ArXiv preprint arXiv:2104.00875, 2021.

[125] MICHIELI U, ZANUTTIGH P. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021.

[126] STAN S, ROSTAMI M. Unsupervised model adaptation for continual semantic segmentation. [C] // AAAI Conf. Artif. Intell. 2022.

[127] FREY J, BLUM H, MILANO F, et al. Continual Learning of Semantic Segmentation using Complementary 2D-3D Data Representations. [J]. ArXiv preprint arXiv:2111.02156, 2021.

[128] PASSALIS N, TEFAS A. Unsupervised Knowledge Transfer Using Similarity Embeddings. [J]. IEEE Trans. Neural Netw. Learn Syst., 2019, 30 (3): 946–950.

[129] FURLANELLO T, LIPTON Z C, TSCHANNEN M, et al. Born-Again Neural Networks. [C] // Int. Conf. Mech. Learn. 2018: 1602–1611.

[130] WANG N, ZHOU W, SONG Y, et al. Real-Time Correlation Tracking Via Joint Model Compression and Transfer. [J]. IEEE Trans. Image Process., 2020, 29: 6123–6135.

[131] GE S, ZHAO S, LI C, et al. Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation. [J]. IEEE Trans. Image Process., 2019, 28 (4): 2051–2062.

[132] PENG Z, LI Z, ZHANG J, et al. Few-shot image recognition with knowledge transfer. [C] // Int. Conf. Comput. Vis. 2019: 441–449.

[133] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition. [C] // Int. Conf. Learn. Represent. 2015.

[134] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017: 1492–1500.

[135] HU J, SHEN L, SUN G. Squeeze-and-excitation networks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 7132–7141.

[136] And ANDREW G. HOWARD M S, ZHU M, ZHMOGINOV A, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018: 4510–4520.

[137] GAO S.-H, CHENG M.-M, ZHAO K, et al. Res2Net: A New Multi-scale Backbone Architecture. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2020, 43 (2): 652–662.

[138] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images. [R]. 0. Toronto, Ontario: University of Toronto, 2009.

[139] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2009: 248–255.

[140] DEVRIES T, TAYLOR G W. Improved Regularization of Convolutional Neural Networks with Cutout. [J]. ArXiv preprint arXiv:1708.04552, 2017.

[141] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond Empirical Risk Minimization. [C] // Int. Conf. Learn. Represent. 2018.

[142] GHIASI G, LIN T.-Y, LE Q V. Dropblock: A regularization method for convolutional networks. [C] // Adv. Neural Inform. Process. Syst. 2018: 10727–10737.

[143] YAMADA Y, IWAMURA M, AKIBA T, et al. Shakedrop Regularization for Deep Residual Learning. [J]. IEEE Access, 2019: 186126–186136.

[144] QI G.-J. Loss-sensitive generative adversarial networks on lipschitz densities. [J]. Int. J. Comput. Vis., 2020, 128 (5): 1118–1140.

[145] NILSBACK M.-E, ZISSERMAN A. Automated flower classification over a large number of classes. [C] // 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. 2008: 722–729.

[146] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset. [R]. CNS-TR-2011-001. California Institute of Technology, 2011.

[147] KRAUSE J, STARK M, DENG J, et al. 3d object representations for fine-grained categorization. [C] // Int. Conf. Comput. Vis. Worksh. 2013: 554–561.

[148] MAJI S, KANNALA J, RAHTU E, et al. A Database for Fine-Grained Aircraft Recognition. [C] // IEEE Conf. Comput. Vis. Pattern Recog. Worksh. 2013.

[149] TAN M, LE Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. [C] // Int. Conf. Mech. Learn. Vol. 97. 2019: 6105–6114.

[150] ZHAO H, JIA J, KOLTUN V. Exploring self-attention for image recognition. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 10076–10085.

[151] MÜLLER R, KORNBLITH S, HINTON G E. When does label smoothing help? [C] // Adv. Neural Inform. Process. Syst. 2019: 4696–4705.

[152] MAATEN L V D, HINTON G. Visualizing data using t-SNE. [J]. Journal of machine learning research, 2008: 2579–2605.

[153] YUAN L, TAY F E, LI G, et al. Revisiting knowledge distillation via label smoothing regularization. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2020: 3903–3911.

[154] ISCEN A, TOLIAS G, GOSSELIN P, et al. A Comparison of Dense Region Detectors for Image Search and Fine-Grained Classification. [J]. IEEE Trans. Image Process., 2015, 24 (8): 2369–2381.

[155] ZHANG C, LIANG C, LI L, et al. Fine-Grained Image Classification via Low-Rank Sparse Coding With General and Class-Specific Codebooks. [J]. IEEE Trans. Neural Netw. Learn Syst., 2017, 28 (7): 1550–1559.

[156] ZHANG Y, WEI X, WU J, et al. Weakly Supervised Fine-Grained Categorization With Part-Based Image Representation. [J]. IEEE Trans. Image Process., 2016, 25 (4): 1713–1725.

[157] SHI W, GONG Y, TAO X, et al. Fine-Grained Image Classification Using Modified DCNNs Trained by Cascaded Softmax and Generalized Large-Margin Losses. [J]. IEEE Trans. Neural Netw. Learn Syst., 2019, 30 (3): 683–694.

[158] SHU X, TANG J, QI G.-J, et al. Image classification with tailored fine-grained dictionaries. [J]. IEEE Trans. Circuit Syst. Video Technol., 2016, 28 (2): 454–467.

[159] PENG Y, HE X, ZHAO J. Object-Part Attention Model for Fine-Grained Image Classification. [J]. IEEE Trans. Image Process., 2018, 27 (3): 1487–1500.

[160] ZHENG H, FU J, ZHA Z, et al. Learning Rich Part Hierarchies With Progressive Attention Networks for Fine-Grained Image Recognition. [J]. IEEE Trans. Image Process., 2020, 29: 476–488.

[161] LIN T, ROYCHOWDHURY A, MAJI S. Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2018, 40 (6): 1309–1322.

[162] XIAO T, XIA T, YANG Y, et al. Learning from massive noisy labeled data for image classification. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2015: 2691–2699.

[163] ZHANG C, BENGIO S, HARDT M, et al. Understanding deep learning requires rethinking generalization. [C] // Int. Conf. Learn. Represent. 2017.

[164] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples. [C] // Int. Conf. Learn. Represent. 2015.

[165] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Machine Learning at Scale. [C] // Int. Conf. Learn. Represent. 2017.

[166] PETERSON J C, BATTLEDAY R M, GRIFFITHS T L, et al. Human Uncertainty Makes Classification More Robust. [C] // Int. Conf. Comput. Vis. 2019: 9616–9625.

[167] SEYEDHOSSEINI M, TASDIZEN T. Semantic Image Segmentation with Contextual Hierarchical Models. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38 (5): 951–964.

[168] ZAND M, DORAISAMY S, ABDUL HALIN A, et al. Ontology-Based Semantic Image Segmentation Using Mixture Models and Multiple CRFs. [J]. IEEE Trans. Image Process., 2016, 25 (7): 3233–3248.

[169] NIRKIN Y, WOLF L, HASSNER T. Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 4061–4070.

[170] ZHU L, JI D, ZHU S, et al. Learning Statistical Texture for Semantic Segmentation. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2021: 12537–12546.

[171] DING H, JIANG X, SHUAI B, et al. Semantic Segmentation With Context Encoding and Multi-Path Decoding. [J]. IEEE Trans. Image Process., 2020, 29: 3520–3533.

[172] YANG K, HU X, STIEFELHAGEN R. Is Context-Aware CNN Ready for the Surroundings? Panoramic Semantic Segmentation in the Wild. [J]. IEEE Trans. Image Process., 2021, 30: 1866–1881.

[173] CHEN L.-Z, LIN Z, WANG Z, et al. Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation. [J]. IEEE Trans. Image Process., 2021, 30: 2313–2324.

[174] MICHIELI U, ZANUTTIGH P. Incremental learning techniques for semantic segmentation. [C] // Int. Conf. Comput. Vis. Worksh. 2019.

[175] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets. [C] // Int. Conf. Learn. Represent. 2015.

[176] HUANG G, SUN Y, LIU Z, et al. Deep networks with stochastic depth. [C] // Eur. Conf. Comput. Vis. 2016: 646–661.

[177] XU C, ZHOU W, GE T, et al. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. [C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 7859–7869.

[178] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation. [C] // Int. Conf. Comput. Vis. 2019: 603–612.

[179] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge. [J]. Int. J. Comput. Vis., 2010, 88 (2): 303–338.

[180] ZHOU B, ZHAO H, PUIG X, et al. Scene Parsing through ADE20K Dataset. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2017.

[181] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2016: 3213–3223.

[182] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. [J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 40 (4): 834–848.

[183] ROTA BULÒ S, PORZI L, KONTSCHIEDER P. In-place activated batchnorm for memory-optimized training of dnns. [C] // IEEE Conf. Comput. Vis. Pattern Recog. 2018.

[184] CHEN Y.-H, CHEN W.-Y, CHEN Y.-T, et al. No more discrimination: Cross city adaptation of road scene segmenters. [C] // Int. Conf. Comput. Vis. 2017: 1992–2001.

致谢

时光如白驹过隙，转眼间即将要结束我在南开大学的硕士阶段的求学之路。此时感慨万千，在此由衷感谢一直默默支持我的父母，感谢和我一起成长的老师、伙伴和挚友。

在过去的三年时光里，非常感谢我的导师程明明老师对我的谆谆教诲和指导。程明明老师以身作则，言传身教，在学术上给予了我很大的帮助，也非常感谢导师为我们提供的良好的计算资源。在过去的三年时光里，我和实验室的其他伙伴结下了深厚的友谊，他们不仅仅是我的同窗，更是我的挚友和伙伴。我们在学习和生活中互相帮助、互相陪伴。在无数个日夜里，当因科研问题没有进展而沮丧时，我们互相鼓励、共同进步。在学习之余，我们也会一起娱乐、放松身心。在过去的三年时光里，我和实验室的伙伴们以及老师们一起合作，完成了一些项目，并撰写了一些论文，这些项目锻炼了我的科研能力和解决问题的能力，使我受益匪浅。

在今后的学习和工作生涯中，我会继续努力，与伙伴们共同进步。

个人简历

张长彬，出生于 1997 年 11 月 10 日。在 2019 年毕业于中国矿业大学计算机科学与技术专业并获得学士学位。于 2019 年至今在南开大学就读专业学位硕士研究生。

研究生期间发表论文:

- **Chang-Bin Zhang**, et al. Delving deep into label smoothing[J]. IEEE Transactions on Image Processing, 2021.
- Yu Zhang, **Chang-Bin Zhang**, et al. Personalized image semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
- Peng-Tao Jiang, **Chang-Bin Zhang**, et al. Layercam: Exploring hierarchical class activation maps for localization[J]. IEEE Transactions on Image Processing, 2021.
- Kai Zhao, Qi Han, **Chang-Bin Zhang**, et al. Deep hough transform for semantic line detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
- **Chang-Bin Zhang**, et al. Representation Compensation Networks for Continual Semantic Segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.