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

基于通用属性知识引导的类增量学习

General Attribute Knowledge for Class-Incremental Learning

论文作者 翟江天 指导教师 程明明 教授
申请学位 工学硕士 培养单位 南开大学
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 评 阅 人 匿名评阅

南开大学研究生院

二〇二五年三月

摘要

类增量学习（Class-Incremental Learning, CIL）旨在逐步学习新类别的同时，尽量减少对旧类别的遗忘。然而，现有方法往往缺乏对通用属性知识（General Attribute Knowledge）的有效利用，导致模型在适应新任务的过程中难以保持对旧任务的稳定性。为此，本文围绕通用知识的引导，从三个方面改进 CIL 方法，以增强模型的稳定性和泛化能力。首先，本文从通用视觉模式出发，采用掩码自编码器（Masked Autoencoder, MAE）框架，通过重建任务学习跨类别共享的通用视觉特征，使模型能够在自监督信号下提取稳定表征，提高新任务适应能力，并通过重构的重放数据进一步缓解遗忘。利用通用视觉模式指导表征学习，有助于构建跨任务一致的特征空间。其次，本文在通用语义属性层面提出了一种细粒度知识选择与恢复机制，用于优化知识蒸馏过程。传统知识蒸馏方法往往对特征或参数施加严格约束，限制了模型对新任务的可塑性。为此，本文引入类别间语义相似性等通用属性，指导旧任务知识的筛选和迁移，以更有效地继承历史信息，减少关键特征的遗忘，同时增强模型的可塑性，使其更适应新类别的学习。最后，本文关注通用注意力模式，针对显著性漂移问题提出任务自适应显著性监督机制。由于任务间的显著性模式可能发生偏移，导致关键特征无法被有效保持，本文引入边界引导的显著性机制与底层显著性蒸馏任务，确保模型在新任务学习时能够持续关注关键区域。同时设计了一种显著性噪声注入与恢复模块，通过扰动显著性信息并进行自适应恢复，使模型在不断适应新任务的同时，仍能稳定地维护对旧任务的重要特征记忆。本文在 CIFAR-100、TinyImageNet 和 ImageNet-Subset 数据集上进行了广泛实验，结果表明，引入通用属性知识后，本文的方法在有示例与无示例类增量学习任务上均取得了更优的稳定性-可塑性平衡，并在多个基准测试中达到了最先进的性能，证明了方法的有效性和优越性。

关键词： 类增量学习；无示例学习；灾难性遗忘；任务自适应显著性；知识蒸馏

Abstract

Class-Incremental Learning (CIL) aims to sequentially learn new categories while minimizing the forgetting of previously learned ones. However, existing methods often fail to effectively leverage General Attribute Knowledge, making it challenging for models to maintain stability when adapting to new tasks. To address this issue, we enhance CIL methods by introducing general knowledge guidance from three perspectives, thereby improving model stability and generalization capability. First, we incorporate general visual patterns by employing a Masked Autoencoder (MAE) framework, which enables the model to learn cross-category shared visual features through a reconstruction task. This allows the model to extract stable representations from self-supervised signals, enhances its adaptability to new tasks, and alleviates forgetting through the replay of reconstructed data. Leveraging general visual patterns for representation learning facilitates the construction of a consistent feature space across tasks. Second, we introduce a fine-grained knowledge selection and recovery mechanism at the level of general semantic attributes to optimize the knowledge distillation process. Conventional knowledge distillation methods often impose strict constraints on features or parameters, limiting the model’s plasticity in adapting to new tasks. To mitigate this issue, we incorporate category-wise semantic similarity as a general attribute to guide the selection and transfer of past knowledge, enabling more effective inheritance of historical information, reducing the forgetting of critical features, and enhancing the model’s plasticity for learning new categories. Finally, we focus on general attention patterns and propose a task-adaptive saliency supervision mechanism to address saliency drift. Since saliency patterns may drift across tasks, leading to the loss of crucial features, we introduce a boundary-guided saliency mechanism and a low-level saliency distillation task to ensure that the model continuously attends to key regions when learning new tasks. Additionally, we design a saliency noise injection and recovery module, which perturbs saliency information and adaptively restores it, allowing the model to maintain robust memory of critical features from previous tasks while

adapting to new ones. We conduct extensive experiments on the CIFAR-100, Tiny-ImageNet, and ImageNet-Subset datasets. The results demonstrate that incorporating General Attribute Knowledge significantly improves the stability-plasticity trade-off in both exemplar-based and exemplar-free class-incremental learning tasks. Our approach achieves state-of-the-art performance across multiple benchmarks, further validating its effectiveness and superiority.

Key Words: Class Incremental Learning; Exemplar-Free Learning; Catastrophic Forgetting; Task-Adaptive Saliency; Knowledge Distillation

目录

第一章 绪论

目前，深度学习在计算机视觉、自然语言处理等领域取得了突破性进展，使得人工智能技术得到了广泛应用。在图像分类、目标检测等任务中，模型的性能通常依赖于大规模数据和强大的计算资源。然而，在实际应用中，模型往往需要在不断变化的环境中持续学习新知识，而不是仅在固定数据集上进行训练。因此，如何在不存储大部分旧任务数据的情况下，使模型能够增量学习新任务，同时保持对旧任务的认知能力，具有重要的研究意义。

另一方面，无示例类增量学习作为增量学习的重要研究方向，通过不保存任何旧类别样本，为解决数据隐私、存储成本等问题提供了一种可行方案。然而，由于模型在学习新任务时容易遗忘旧任务的知识，如何有效缓解灾难性遗忘，使模型在稳定性和可塑性之间取得平衡，成为任务的核心挑战。针对这一问题，本文从通用属性知识的角度出发，研究如何在任务中优化模型的知识保留与迁移能力，以提升增量学习的实际应用价值。

第一节 研究背景和意义

深度神经网络在许多计算机视觉任务上取得了最先进的性能，其强大的特征学习能力使其在图像分类、目标检测和语义分割等任务中表现优异。传统的深度学习模型通常是在静态环境下训练的，假设所有类别和数据在训练时已经可用，并且任务定义明确、数据分布稳定。然而，在现实世界中，数据分布是动态变化的，新类别、新任务不断出现，而直接使用标准训练方式更新模型往往会导致灾难性遗忘——即模型在学习新任务的同时，无法保持对旧任务的记忆。这一问题在计算资源受限或数据存储受限的场景下尤为严重，尤其是在隐私保护、医疗诊断和在线学习等应用中，长期存储和访问旧数据可能会受到严格限制。因此，如何在动态环境下高效地进行增量学习，同时避免灾难性遗忘，成为当前研究的一个核心问题。

类增量学习是一种增量学习场景，允许模型在训练后续任务时逐步扩展其识别的类别集合。大多数现有的类增量学习方法依赖于内存缓冲区，其中存储了一部分来自过去任务的样本，用于辅助模型在学习新任务时维持对旧任务的

知识。然而，在许多现实场景中，由于存储限制或数据隐私的考量，保留旧任务的样本并不现实。因此，无示例类增量学习应运而生，它要求模型在不保留任何旧任务数据的情况下进行增量学习，这无疑是一个更具挑战性的研究问题。

无示例类增量学习在需要严格隐私保护、数据存储受限或法规要求明确禁止历史数据保留的应用场景中展现出独特价值。在医疗诊断、金融风控以及用户行为建模等领域，原始数据往往因涉及敏感隐私信息而不能长期存储或复用。此类情况下，传统基于样本重放的方法将面临严重的现实约束，而无示例方法通过不依赖于旧数据的方式，仍能够实现对新类别的学习与旧类别知识的保留，为合规性要求高的行业提供了可行的解决方案。

此外，在边缘计算与嵌入式设备等资源受限环境中，无示例类增量学习亦具有明显优势。这类设备通常无法负担大规模样本缓存和频繁的数据访问操作，而这些方法通过精简模型设计和优化训练流程，有效地降低了对存储与计算资源的依赖。它能够在动态环境中持续适应新任务，具备良好的迁移性与鲁棒性，因此被广泛视为实现真实世界中持续学习系统的关键路径之一。随着对绿色 AI 和可持续计算的重视，无示例增量学习的研究有望在实际部署中获得更大应用价值。

由于缺乏旧数据的直接监督，模型在学习新类别时很容易发生显著性漂移，导致其注意力从旧类别的判别性特征转移到新类别特征上，从而加速了灾难性遗忘的发生。为了解决无示例类增量学习中的灾难性遗忘问题，研究者们提出了多种方法，主要包括基于知识蒸馏、参数正则化和生成模型的方法。

知识蒸馏是一种常见的方法，通过在新任务训练过程中引入旧模型的软目标来引导新模型的学习，从而减少对旧类别的遗忘。模型可以使用旧任务的预测分布作为额外的监督信号，以维持旧任务的特征表征。然而，在无示例类增量学习场景下，由于旧数据无法访问，模型难以生成准确的蒸馏目标，这限制了该方法的有效性。

参数正则化方法通过对神经网络的关键参数施加约束，以保持旧任务的知识。例如，可以计算神经网络权重对旧任务损失的敏感度，并对关键权重施加较大的惩罚，以防止其在新任务训练中发生剧烈更新。此外，其他方法也采用类似的思想来稳定模型参数，减少灾难性遗忘。

生成模型方法尝试通过生成旧任务的数据来辅助增量学习，例如利用自编码器或对抗生成方法来合成旧任务的数据，从而提供类似于真实样本的监督信

息。然而，这类方法面临生成质量和计算成本的问题，生成的数据往往难以完全匹配真实数据的分布，从而影响模型性能。

尽管上述方法在一定程度上缓解了无示例类增量学习中的灾难性遗忘问题，但它们仍然存在局限性。例如，知识蒸馏方法依赖于教师模型的高质量输出，而在无示例学习场景下，教师模型往往无法提供稳定的知识迁移。参数正则化方法虽然能保持部分旧知识，但无法有效适应新任务的多样性。生成模型方法虽然理论上可以恢复旧数据，但生成的数据质量仍然是一个挑战。

因此，如何在无示例类增量学习中进一步提高模型的稳定性和泛化能力，仍然是一个亟待解决的问题。研究者们正在探索更加高效的知识迁移策略、参数调整方法以及更稳定的特征保持机制，以应对这一挑战。本文从通用属性知识的引导与维护层面，从多个角度改进这一任务固有的知识遗忘问题。

第二节 研究现状与挑战

增量学习和无示例类增量学习在应对灾难性遗忘方面取得了一定进展，但仍面临诸多挑战。现有增量学习方法主要包括基于重放、正则化和参数隔离的策略，然而，它们分别受限于存储需求、跨任务关系建模不足和知识共享能力受限等挑战。无示例类增量学习由于无法存储旧类别样本，使得遗忘问题更加严重，现有方法如正则化、知识蒸馏和生成模型均存在各自的局限性，难以根本性地解决长期记忆保持的问题。未来的研究可以围绕通用属性知识的补充与引导对这些挑战加以改善。

一、 增量学习

增量学习旨在使模型能够适应不断变化的任务序列，同时避免灾难性遗忘。近年来，随着深度学习的进步，增量学习在计算机视觉、自然语言处理等领域得到了广泛关注，并催生了多种缓解遗忘的方法，包括基于重放、正则化和参数隔离的策略。尽管这些方法在一定程度上缓解了遗忘问题，但仍然存在对任务特定信息的依赖较强、通用知识建模能力不足等局限性，导致在长期任务学习过程中，模型仍可能丢失旧任务的关键信息。

基于重放的方法通过存储旧任务的一部分样本，在训练新任务时重新输入这些样本，以保持旧任务的知识。例如，GEM[1] 和 AGEM[2] 通过调整梯度方向，以减少对旧任务的遗忘。然而，这类方法依赖于样本存储，难以适用于隐

私受限或存储受限的场景，同时重放策略主要缓解短期遗忘，难以有效提升跨任务的泛化能力。相比之下，基于正则化的方法尝试通过约束模型参数的更新，使其在学习新任务时保留对旧任务的记忆。例如，EWC[3] 通过计算参数的重要性，减少关键参数的更新幅度，以降低遗忘风险。然而，这类方法通常采用全局参数约束，忽略了任务间的特征分布差异，可能难以适应具有较大变化的任务。此外，参数隔离方法通过为不同任务分配独立的模型参数，以减少任务间干扰，例如 L2P[4] 和 DualPrompt[5] 采用动态提示机制优化增量学习能力。然而，随着任务数量的增加，这些方法的存储和计算成本会显著上升，同时不同任务间的知识共享能力受限，可能影响模型的泛化性。

总体来看，现有增量学习方法在应对灾难性遗忘方面虽然取得了一定进展[6, 7]，但仍面临诸多挑战。例如，重放方法依赖具体样本，正则化方法难以建模跨任务关系，参数隔离方法限制了知识共享，使得模型难以真正从本质上克服遗忘问题。未来的研究可以进一步探索如何提取任务不变的特征，使模型能够基于通用属性知识进行知识迁移，提高其在无示例增量学习场景下的适应能力。此外，结合自监督学习、知识蒸馏和对比学习等策略，有望在不增加额外存储开销的前提下，使模型在学习新任务的同时保持对旧任务的长期记忆。

二、 无示例类增量学习

无示例类增量学习是增量学习中的一个特殊研究方向，要求模型在逐步接收新类别的同时，不存储旧类别的任何样本，从而进一步加剧了灾难性遗忘的问题。传统的基于重放的方法无法直接应用于该场景，因此研究者们提出了其他策略来维持模型对旧类别的记忆。例如，EWC[3] 等正则化方法通过约束关键参数的更新幅度来减少遗忘，但由于无法提供具体的旧任务信息，模型仍然难以在新任务到来时保持旧类别的判别能力。此外，部分研究探索了知识蒸馏的方式，即利用旧模型的输出作为软标签，引导新模型学习[8, 9]。然而，蒸馏方法通常受到类别间语义偏移的影响，难以保证模型在长期增量过程中维持稳定的类间关系。

另一类方法是通过生成模型合成重放样本来减少遗忘。这类方法通常利用生成对抗网络或变分自编码器在连续学习过程中生成先前任务的数据，从而实现对旧任务知识的重现和巩固。例如，DGR[10] 方法使用 GAN 来同时训练生成器和任务模型，生成器负责合成历史任务的样本，而任务模型在当前任务数据和生成数据上联合训练，以缓解灾难性遗忘。后续的研究如 MeRGAN[11] 进一步

改进了生成样本的质量与多样性，增强了对旧任务的保留能力。与传统的样本缓存策略相比，生成重放方法不依赖于真实数据的存储，因此更符合隐私保护和资源受限场景下的应用需求。然而，生成模型的训练本身面临稳定性和模式崩溃等挑战，其在复杂任务下的泛化能力仍是一个亟待解决的问题。Data-Free Class-Incremental Learning 是近年来无示例类增量学习中的重要方向，旨在基于不访问原始训练数据的前提下，通过生成模型或其他替代机制，实现持续学习能力。典型工作如 ABD[12] 提出利用模型反演生成的伪样本进行知识回放，并诊断了传统蒸馏策略在合成图像上的失效问题。为此，ABD 引入了局部交叉熵损失、重要性加权的特征蒸馏以及线性分类头的微调策略，有效缓解了由语义偏移与分布偏移共同引起的灾难性遗忘现象。另一代表性工作 R-DFCIL[13] 则从特征表示层面切入，提出关系引导的表示学习框架以缓解合成数据与真实数据之间的域间差异所带来的灾难性遗忘问题。该方法通过引入硬知识蒸馏与关系知识蒸馏相结合的方式，在保持旧类知识稳定性的同时提升对新类的可塑性表达，并通过局部分类损失与后续的全局类别平衡精调策略，进一步优化旧新类之间的判别边界。

当前无示例类增量学习的主要挑战在于，现有方法大多依赖模型权重的调整，而缺乏对通用属性知识的有效建模，使得模型难以形成稳定的长期记忆[14]。未来研究可以从两个方向入手：一方面，探索更高效的知识蒸馏和自监督学习方法，使模型能够在无示例场景下进行稳健的知识迁移[15]；另一方面，结合视觉 Transformer 的全局特性，提出基于通用属性学习的增量学习框架，以减少对具体任务数据的依赖，从而提高增量学习的长期稳定性[5]。

第三节　本文研究内容

本文围绕类增量学习任务中如何利用通用属性知识进行有效知识迁移和保持展开研究，旨在解决现有方法中由于重放数据不足、缺乏通用知识约束、知识蒸馏粗糙、缺少适应性知识以及对遗忘机制理解不足等挑战。

为了更好地维护类增量学习过程中的通用属性知识，本文提出了一种双边 MAE 框架，该方法能够在知识迁移过程中高效保持通用属性信息，从而缓解因重放数据不足带来的灾难性遗忘问题。其次，针对现有方法中知识蒸馏过于粗糙、无法有效区分任务无关知识的问题，本文提出了一种细粒度知识选择策略，通过学习任务无关知识，使得模型在增量学习过程中能够更好地保持通用知识。

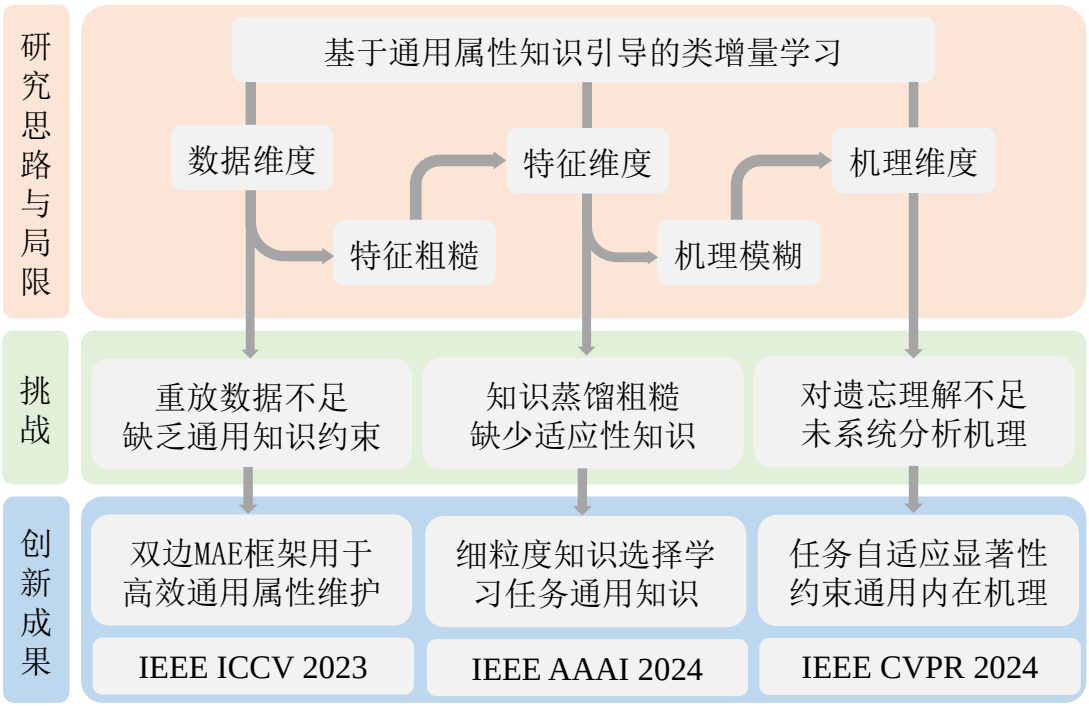

图 1.1 本文内容简述

此外，为了提升模型的自适应能力，本文进一步引入任务自适应显著性约束，通过建模通用知识的内在机理，引导模型在不同任务之间进行更有效的知识迁移，从而提高模型的泛化能力和稳定性。整体研究内容与成果见图 1.1。

一、 基于掩码自编码器的类增量学习

在类增量学习过程中，模型需要在不断接收新类别的同时，尽可能保留对旧类别的知识。然而，由于存储资源受限，通常只允许使用固定大小的内存，这使得基于示例重放的增量学习方法面临较大的挑战。现有研究提出了一些缓解遗忘的策略，例如利用生成式网络[10, 11, 16] 生成旧任务样本，以支持重放训练。然而，尽管生成式方法可以在一定程度上缓解遗忘，但它们仍存在一些固有问题，例如生成图像的质量难以保证，同时生成器自身也容易受到灾难性遗忘的影响，使得长期知识保持变得更加困难。

在本研究中，引入了掩码自编码器[17] 作为一种高效的重放机制，以缓解类增量学习中的遗忘问题。MAE 通过对输入图像进行大比例的随机掩码，并利用 Transformer 结构重建被遮挡部分，仅需部分可见的图像块即可恢复完整的视觉信息。相比于传统的示例存储方法，这一特性使得 MAE 具备更高效的存储能力，在相同的内存预算下能够存储更多的视觉线索。此外，与基于 GAN 的方法相

比，MAE 的重放策略更加稳定，因为它利用部分已知信息推理全局结构，能够减少因任务转换而导致的模型不稳定性，同时降低跨任务干扰带来的遗忘效应。

相比于已有的 CIL 方法，基于 MAE 的重放方法具备以下优势：（1）存储效率更高。传统示例存储方法依赖于完整样本，而 MAE 仅需存储部分关键图像块，即可在训练过程中恢复原始信息，从而提高存储效率。（2）稳定性更强：GAN 及其他生成式模型在不同任务间的图像合成质量可能发生波动，而 MAE 通过部分线索推理全局信息的方式减少了任务间的语义偏移，使得重放数据在不同任务中保持较高的稳定性。（3）通用属性知识引导：相比于传统的重放方法，MAE 依赖于全局视觉结构的恢复能力，而这种全局信息通常是任务无关的。因此，MAE 在不同任务中的遗忘较少，使得模型能够更好地保留通用属性知识，并提升跨任务知识迁移的能力。同时，本文进一步提出了一种双边 Transformer 架构，以高效利用 MAE 进行重放。本文在模型中引入了通用属性引导机制，鼓励模型在不同任务间共享稳定的特征表示，减少因任务漂移导致的遗忘。

整体而言，基于 MAE 的类增量学习方法在兼顾存储效率、模型稳定性和知识迁移能力方面展现出了良好的潜力。结合双边 Transformer 结构和通用属性引导策略，本文的框架能够在固定内存约束下，实现更有效的跨任务知识保持和迁移，为未来的类增量学习研究提供了新的思路。

二、 基于细粒度知识选择的无示例类增量学习

在无示例类增量学习场景下，模型需要在不存储旧类别样本的情况下，持续适应新任务，同时尽可能保持对旧任务的知识。然而，由于缺乏旧类别样本，模型容易受到灾难性遗忘的影响，导致对先前任务的知识丧失。为此，本文提出了一种基于细粒度知识选择的策略，以提高模型在增量学习过程中的稳定性和可塑性。该方法主要包括分块级知识选择和改进的原型恢复策略，引入通用属性知识引导机制，以更有效地跨任务共享稳定的特征表示，从而缓解遗忘问题并提升知识迁移能力。

分块级知识选择基于视觉 Transformer 结构，该结构通过对输入图像进行分块级表示，使得模型能够在细粒度级别上区分任务相关的重要特征。在此基础上，本文利用 [CLS] 标记嵌入与各分块的相似性来度量任务相关性，从而对不同类别的知识施加不同的蒸馏约束，以最大程度地保持旧任务的知识，同时提升模型对新类别的适应能力。具体而言，前景分块通常携带与当前任务高度相关的信息，因此在训练过程中对这些分块施加较低强度的正则化，使其能够灵

活调整，从而更好地适应新任务的学习。而背景分块通常对具体任务的依赖较小，因此可以在不同历史模型版本之间保持稳定。本文通过对这些分块施加更强的正则化约束，使得模型能够在不同增量阶段保持一致的背景表示，提高模型的稳定性，减少灾难性遗忘。

除了分块级知识选择，本文还提出了一种改进的原型恢复方法，以进一步减少因类别漂移导致的分类器偏差。传统的类别原型恢复方法通常假设类别原型服从高斯分布，但这一假设在实际应用中可能并不成立，从而导致恢复的原型与真实分布存在较大偏差，进而影响分类器的决策性能。为了解决这一问题，本文提出了一种更为稳健的原型恢复策略，主要包括两个步骤。首先，计算当前任务样本与其类别中心之间的偏移距离，并进行正则化，以保证类别间的数据分布尽可能稳定，从而减少因任务变化带来的分布漂移。其次，基于当前任务的原型信息与历史任务的原型信息，共同生成更精确的旧任务原型，并将其用于分类器重放。这种方式能够有效缓解分类器的偏差，同时降低因缺乏旧任务样本而导致的遗忘现象，使得模型在不存储旧类别样本的情况下，依然能够较好地保持对旧任务的分类能力。这一任务无关的通用原型维护也提升了模型在不同任务间的稳定性。

整体而言，本文的方法通过分块级知识选择提升了模型的任务相关性建模能力，使得模型能够更有效地区分任务相关和任务无关的特征信息，从而减少灾难性遗忘的发生。同时，改进的原型恢复策略减少了分类器的偏差，使得分类器在不同任务间的切换更加平稳。更重要的是，本文通过通用知识引导机制在不同任务间建立了一致的知识表征，使得模型能够更好地在长期增量学习过程中保持知识稳定性，并增强跨任务的适应能力。实验结果表明，该方法在无示例 CIL 任务中显著提升了模型的性能。

三、 基于任务自适应显著性监督的无示例类增量学习

相比于基于 ViT 结构的细粒度知识选择框架，本文同时提出任务自适应显著性监督，专注于缓解跨任务显著性漂移问题，以确保网络在增量学习过程中能够稳定关注与任务相关的显著区域，同时兼顾模型的可塑性与稳定性。由于无示例类增量学习需要在不同任务间迁移知识，显著性区域的漂移可能导致模型关注错误的特征，从而影响分类性能。为了解决这一问题，TASS 主要由膨胀边界监督、低层辅助监督任务以及显著性噪声去噪机制三部分组成，共同作用以保持任务间的显著性一致性，提高模型的跨任务适应能力。

在跨任务的学习过程中，显著性漂移通常发生在中间特征层，并导致注意力从前景区域扩散至无关的背景区域。因此，本文引入了一种膨胀边界监督机制，以在特征提取过程中提供额外的约束。具体而言，该机制通过控制模型的注意力边界，防止其在跨任务时发生偏移。当模型学习新类别时，其注意力可能会在缺乏旧类别数据的情况下发生漂移，使得某些关键特征区域被忽略或错误地扩展至背景区域。通过膨胀边界监督，本文在中间层对注意力进行约束，使其聚焦于前景区域，减少对背景区域的扩散。这样可以有效提升模型的任务适应能力，同时避免因注意力分布的异常变化而导致的灾难性遗忘。

本文进一步引入了一种任务无关的低层监督任务，以增强模型在不同任务间的注意力稳定性。低层特征通常包含丰富的边缘和轮廓信息，这些信息对于显著性区域的稳定性至关重要。因此，本文设计了一种辅助监督机制，使模型在低层特征空间中执行额外的显著性预测任务，以提供稳定的显著性引导。该监督任务能够帮助模型更好地识别前景区域，并在跨任务时保持一致的注意力模式。此外，由于显著性预测任务是任务无关的，因此它不会受到类别漂移的影响，使得模型能够在不依赖旧类别数据的情况下，保持长期的显著性稳定性。

为了进一步减少显著性漂移，本文提出了一种显著性噪声去噪机制，以增强模型的鲁棒性。该机制采用噪声注入与去噪的策略，使得模型能够在不同任务间保持稳定的注意力模式。具体而言，方法在特定的特征通道中注入显著性噪声，并训练模型学习如何去除这些噪声，使得最终的注意力分布更加稳定。由于模型在不同任务间的注意力分布可能发生较大的变化，显著性噪声去噪机制能够有效抑制由于任务漂移导致的注意力异常，确保模型始终关注最具判别性的特征区域。这种去噪机制不仅提升了模型的任务适应能力，还增强了其对跨任务知识迁移的稳健性。

实验结果表明，方法在无示例类增量学习任务中取得了显著的性能提升。这些关于显著性信息都属于通用属性知识，能够在任务间稳定维护模型的内在注意力，从而减少灾难性遗忘。

第四节　论文章节安排

本文围绕类增量学习任务，研究如何有效缓解灾难性遗忘，并从数据，模型蒸馏和机理的角度围绕通用属性知识提出优化方案，以提高模型在增量学习场景中的稳定性和适应性。

本文提出使用掩码自编码器作为 CIL 的高效学习器。它们可以很容易地与监督损失集成以进行分类。此外，MAE 还能从随机选取的块中可靠地重建原始输入图像，在 CIL 中，用它来更有效地存储来自过去任务的示例。本文还提出了一种双边 MAE 框架，用于从图像级和嵌入级的融合中学习，从而产生更高质量的重建图像和更稳定的表征。

现有的知识蒸馏方法在无示例类增量学习中面临较大的挑战，主要体现在模型难以有效保持旧任务知识，并在新任务学习过程中发生灾难性遗忘。针对这一问题，本文提出了一种细粒度知识蒸馏策略，利用任务相关区域信息来增强模型对关键特征的保持能力。

另一方面，增量学习过程中，模型需要在多个任务之间进行知识迁移，但现有方法难以在保持旧知识和学习新知识之间取得平衡。为此，本文设计了一种显著性增强的训练策略，通过在不同任务阶段对显著性区域进行动态调整，引导模型在学习新知识的同时，最大程度地保留已有知识。

为了验证提出方法的有效性，本文在多个公开数据集上进行了实验，并与当前主流增量学习方法进行对比。实验内容包括分类精度、灾难性遗忘度量、显著性漂移分析以及消融实验，以全面评估方法的有效性和适用性。本文共分为六章，其主要内容如下：

第一章为绪论，介绍了类增量学习的研究背景和意义，分析了在增量学习过程中模型面临的主要挑战，包括灾难性遗忘和显著性漂移等问题。随后，回顾了国内外在增量学习、知识蒸馏以及任务自适应方法方面的研究进展，并对本文的主要研究内容和创新点进行了概括。

第二章为相关工作，综述了增量学习、掩码自编码器、知识蒸馏、显著性建模的相关研究工作。回顾了当前在图像分类任务上的增量学习方法，介绍了这些方法的核心思想，并分析了其在增量学习中的应用。

第三章介绍了一种图像级和嵌入级融合的双边 MAE 框架，该框架旨在结合高质量的原始图像与重建图像，以增强重放数据的有效性，缓解灾难性遗忘。在类增量学习任务中，如何在缺乏旧任务样本的情况下保持模型对旧类别的认知，是影响模型性能的关键挑战之一。为此，本文提出了一种双边 MAE 结构，使模型在自监督学习过程中充分利用重建图像所包含的丰富细节信息。通过融合原始图像与重建图像，并结合通用属性知识进行特征提取，模型能够更有效地学习跨类别的稳定模式，从而提升对旧类别的记忆能力。此外，在嵌入层面，

设计了两个互补分支的信息融合机制，使不同层次的特征信息在学习过程中得以充分整合，从而提高特征表示的稳定性和多样性。该机制不仅能在可塑性与稳定性之间取得更优的平衡，还能借助通用属性知识约束，使特征表达更加具有泛化性，确保跨任务的知识迁移损失更少。此外，MAE 框架本身具备高效的自监督约束特性，能够促进模型在不同任务之间保持一致的通用特征，从而进一步提升类增量学习的性能。

第四章提出了一种基于分块级知识选择的增量学习方法，该方法基于视觉 Transformer 结构，通过分块级知识选择增强模型的稳定性与可塑性。ViT 采用分块级表示建模输入图像，并利用 [CLS] 标记嵌入与各分块的相似性衡量任务相关性。据此，本文对不同类别的知识施加差异化蒸馏约束。前景分块包含关键信息，因此施加较低强度正则化，以便灵活调整，学习新类别特征，并结合通用属性知识保持旧任务的核心信息。此外，为缓解原型恢复偏差问题，本文提出了一种改进策略，计算样本与类别中心的偏移并进行正则化，结合通用属性知识生成更稳定的旧任务原型，从而降低遗忘现象，提升无示例类增量学习任务的性能。

第五章介绍了一种任务自适应显著性监督策略，用于缓解跨任务显著性漂移问题，使模型在增量学习过程中稳定关注任务相关区域，同时兼顾可塑性与稳定性。由于不同任务的显著区域分布可能存在较大差异，TASS 主要包括三个关键部分。首先，膨胀边界监督机制约束模型关注前景区域，并结合通用属性知识引导显著性特征，防止注意力向背景扩散。其次，引入低层辅助监督任务，如显著性预测，以增强模型在不同任务间的注意力稳定性，并通过通用属性知识确保显著性特征的一致性，减少漂移影响。此外，本文设计了一种显著性噪声去噪机制，在特定特征通道中注入噪声并训练模型去噪，并借助通用属性知识维持跨任务的一致特征去噪能力。

第六章为总结与展望，本章总结了提出的三种关键方法：双边 MAE 框架通过图像级与嵌入级融合增强知识保持能力，分块级知识选择利用 ViT 结构优化任务相关性建模，任务自适应显著性监督通过显著性引导提升跨任务稳定性。三者均结合通用属性知识，以增强模型的泛化能力。最后，展望了未来可能的研究方向。

第二章　相关工作

本章介绍了增量学习的相关技术，重点讨论了正则化方法、重放策略和参数隔离等经典方法，以及增量学习在不同应用场景中的实践。

本章探讨了自监督学习在增量学习中的应用，介绍了自监督学习的基本原理，并重点分析了掩码自编码器如何与增量学习相结合，以提高模型的适应性和泛化能力。在知识蒸馏部分，回顾了传统知识蒸馏方法，并讨论了知识蒸馏在增量学习中的应用，包括如何利用知识蒸馏缓解灾难性遗忘问题。此外，本章还介绍了知识蒸馏与其他增量学习技术（如正则化方法和参数隔离）的结合方式，以进一步提升模型的性能。最后，本章总结了显著性监督技术及其在增量学习中的应用。显著性监督方法能够引导模型关注关键信息，提高模型的稳定性和适应能力。同时探讨了显著性监督与其他增量学习方法的结合方式，并分析了其在不同任务中的应用效果。

第一节　增量学习

增量学习是一种允许模型在新数据到来时进行持续学习，而无需遗忘旧知识的学习范式。这一能力在深度学习应用中尤为重要，特别是在数据分布随时间变化的场景下，如在线学习[18]、长期知识积累[19] 以及动态环境适应[20] 等。增量学习主要面临灾难性遗忘、模型稳定性与可塑性的权衡、以及知识迁移等挑战[21]。针对这些问题，研究者提出了一系列有效的方法，主要包括正则化方法、重放策略和参数隔离方法[22]。

一、　正则化方法

正则化方法通过引入额外的损失项，使模型在学习新知识的同时保留旧知识。这类方法的核心思想是约束模型参数的更新，从而减少灾难性遗忘。弹性权重保持[3] 是一种经典方法，它通过估计旧任务参数的重要性，在损失函数中添加一个正则项，以限制关键参数的更新幅度。在线 EWC[23] 是对原始 EWC 的改进版本，它在训练过程中持续更新参数重要性估计，使得方法适用于更大规模的增量学习场景。此外，突触智能[24] 通过跟踪训练过程中的参数变化，计算

参数的重要性，并在优化过程中对重要参数进行约束，从而减少遗忘。一些较新的方法，如 PSSR[25] 通过原型相似性重放和相似性调整正则化的类增量学习。OVOR[26] 提出了一种基于虚拟异常值的正则化方法来收紧分类器的决策边界，从而减轻不同任务之间的类别混淆。Hu 等人[27] 提出了一个简单但有效的协方差约束损失，以强制模型学习具有相同协方差矩阵的每个类分布。

这些方法在一定程度上缓解了灾难性遗忘问题，但仍然存在局限性。例如，EWC 及其变种通常需要存储重要性参数矩阵，并在优化过程中计算额外的正则项，增加了计算成本[28]。此外，当任务之间的相似性较低时，这些方法的有效性可能会降低[29]。

二、 重放策略

重放策略通过存储部分旧数据或生成类似旧数据的样本来缓解灾难性遗忘。经验重放方法直接存储旧任务数据，并在训练新任务时混合旧数据进行训练[30]。这种方法的优点在于，它能够保持旧任务数据的完整性，使得模型在训练新任务时仍然可以访问过去的知识。然而，直接存储旧数据可能会带来存储成本过高的问题，并且在某些隐私敏感的场景下可能不可行[20]。

为了解决存储数据的问题，生成式重放方法被提出，该方法通过生成对抗网络[31] 或变分自编码器[32] 生成旧任务的数据，使得模型可以在不存储实际数据的情况下回忆过去任务[33]。生成式重放方法避免了直接存储数据的问题，但生成的样本质量可能不足，从而影响模型的增量学习能力。此外，训练高质量的生成模型本身是一个具有挑战性的问题[34]。一些最近的方法[35] 提出了一种称为动态重放训练的新方法来解决模型对以前学习过的任务的动态遗忘问题。GenFCIL[36] 提出了一个称为生成联邦类增量学习的 FCIL 框架，引入了一个轻量级生成器，以促进客户端之间的知识共享并保留来自所有客户端的累积知识。此外，MixER[37] 提出了一种缓解表示偏移的解决方案，即将非对称混合训练纳入重放方法中。

三、 参数隔离方法

参数隔离方法通过为不同任务分配不同的模型参数，从而防止新任务对旧任务的干扰。这类方法的一个代表是渐进式神经网络[38]，它为每个新任务引入新的网络模块，并保持旧模块的权重不变，从而实现任务间的独立性。然而，PNN 的主要问题是模型规模会随着任务的增加而不断增长，限制了其在长期任

务中的应用[39]。

PathNet[40] 通过进化算法选择特定的网络路径，使得不同任务之间的共享程度可以动态调整，从而在一定程度上减少模型膨胀问题。此外，PackNet[41] 采用网络剪枝技术，为不同任务分配不同的子网络，并在训练新任务时冻结已使用的参数。这种方法能够有效利用固定的模型容量，但当任务数量较多时，可能会面临模型可用参数资源逐渐减少的问题。Wen 等人[42] 提出了方法来为 CIL 寻找多个不同的老师，采用权重置换特征扰动和多样性正则化技术来确保老师的机制多样化，实现老师模型参数的隔离化。Hu 等人[43] 通过在任务专家网络的中间层之间引入密集连接来实现，这些连接使知识能够通过特征共享和重用从旧任务转移到新任务。Liang 等人[44] 在每个新任务到来时，首先评估模型的可塑性，然后根据评估结果，自适应地扩展每一层的参数。

四、 增量学习的应用

增量学习在多个领域有着广泛应用。例如，在图像分类任务中，增量学习可以帮助模型适应新类别，而无需重新训练整个模型[30]。在语音识别领域，语音模型可以通过增量学习适应新的发音、语言或环境噪声[45]。在机器人学习中，增量学习允许机器人在与环境交互时逐步学习新的技能，而不遗忘已有的技能[20]。此外，在自然语言处理领域，增量学习可用于动态扩展词汇表、适应新的语言风格或领域知识[46]。

增量学习作为深度学习中的关键技术，为持续学习和动态适应提供了可能。未来的研究方向可能包括更高效的知识整合策略[39]、更强的任务适应能力[22]，以及更低计算成本的增量学习方法[28]。

第二节 自监督学习在增量学习中的应用

随着深度学习在各类任务中的广泛应用，增量学习成为解决数据动态变化场景下模型训练的重要研究方向。然而，传统的增量学习方法往往依赖监督信号，难以有效适应数据的不断更新，同时容易产生灾难性遗忘问题。自监督学习作为一种无需人工标注的学习范式，能够通过构造人工监督信号，使模型在无标签数据上学习稳定且具有泛化能力的特征表示，为增量学习提供了一种新的解决思路。

一、 自监督学习的基本原理

自监督学习是一种无监督学习范式[47]，它通过构造人工监督信号，使模型在无标签数据上学习有意义的特征表示[48, 49]. 其核心思想是利用数据本身的内在结构设计学习任务，使得模型能够提取稳健的特征，而不依赖人工标注。在图像领域，自监督学习通常通过前置任务或对比学习来实现[50]. 早期的自监督学习方法主要依赖于设计前置任务，例如块排列任务[51]，该方法将图像随机划分为多个块并打乱顺序，让模型学习如何恢复正确顺序，从而学习局部与全局结构之间的关系。此外，旋转预测任务[52] 要求模型预测图像的旋转角度，这一方法促使模型关注物体形状和方向等关键特征。颜色恢复任务[53] 让模型从灰度图像恢复彩色图像，以鼓励模型学习不同物体类别的颜色分布。此外，图像修复任务[54] 让模型填补被遮挡的部分，从而促进模型学习上下文信息。这些前置任务在一定程度上能够帮助模型学习有效的特征表示，但其泛化能力仍存在局限性[55]。

近年来，对比学习成为自监督学习的主流方法，该方法的基本思想是通过构造正样本对和负样本对，让模型学习如何将相似数据投影到相近的表示空间，并将不同数据区分开来[56]. SimCLR 方法[57] 通过数据增强构造相同图像的两个视图，并要求模型学习将这两个视图的嵌入尽可能接近。MoCo 方法[58] 则引入动量编码器，构建更稳定的动态负样本队列，使得模型能够在更大范围的样本空间上进行对比学习。此外，BYOL 方法[59] 采用自蒸馏策略，通过教师-学生模型架构学习特征，而不依赖负样本，从而有效提升了模型的稳定性和可迁移性。SwAV[60] 提出了基于聚类的对比学习框架，使得模型能够在无标签数据上学习可泛化的视觉特征。对比学习在多个任务上取得了优异的表现，并为增量学习提供了一种潜在的解决方案[61, 62]。

二、 掩码自编码器与增量学习的结合

掩码自编码器（Masked Autoencoder, MAE）[17] 是一种基于自监督学习的深度表示学习方法，它通过对输入图像进行随机遮蔽，并要求模型重建被遮蔽部分，从而学习更丰富的视觉特征。相比于传统的前置任务或对比学习方法，MAE 具有多方面的优势[63]。首先，由于大部分输入数据被遮挡，模型被迫关注全局上下文信息，而不仅仅是局部细节，这使得其能够学习到更具鲁棒性的特征。其次，MAE 仅使用部分数据进行训练，但仍能获得高质量的特征表示，使其在数

据有限的增量学习场景下具有较大的优势。此外，MAE 预训练的模型具有良好的迁移能力，能够较好地泛化到新任务，特别适用于增量学习中的知识转移[64]。基于这些特性，MAE 在增量学习场景下的应用具有较大的潜力[65]。

三、 自监督学习在增量学习中的应用

在增量学习场景下，模型需要在学习新任务时尽可能保留旧任务的知识，而不产生灾难性遗忘。已有研究尝试将自监督学习应用于增量学习，以增强模型的稳定性和可塑性。例如，PASS 方法[66] 结合旋转预测任务，让模型学习能够跨任务迁移的特征，以减少新任务对旧任务的干扰。此外，DualNet 方法[67] 采用 Barlow Twins[61] 机制，引入一个“慢”任务进行表征学习，从而对“快”增量学习进行正则化，使知识保持更加稳定。此外，Meta-CL[68] 通过自监督学习增强增量模型的泛化能力，使得其能够在无监督的情况下更好地适应新任务。CURL[69] 采用对比学习的方法，在强化学习的增量学习任务中提高了特征的可重用性，并有效减少了灾难性遗忘现象。SimSiam[70] 证明了即使在无负样本的情况下，自监督学习仍然能够有效提升增量学习模型的特征表示能力。这些方法表明，自监督学习可以有效缓解灾难性遗忘，并提升模型的适应性。然而，自监督学习在增量学习中的应用仍然面临挑战，例如如何设计更有效的预训练任务，使得特征能够更好地适应增量场景，以及如何降低自监督训练的计算开销等问题，仍需进一步研究[71]。

第三节 知识蒸馏

知识蒸馏最早由 Hinton 等人提出[72]，其核心思想是通过教师模型向学生模型传递知识，使得学生模型能够在较小的参数规模下达到与教师模型相近的性能。在增量学习场景中，知识蒸馏被广泛用于缓解灾难性遗忘[73]，并且在无存储旧数据或仅存储有限旧样本的情况下，有助于提高模型的学习能力。

一、 传统知识蒸馏方法

在标准的知识蒸馏框架中，蒸馏损失通常由教师模型的软目标和硬目标共同决定。Hinton 等人[72] 提出使用温度调节的软目标进行知识蒸馏，随后 Furlanello 等人[74] 提出了 Born-Again Networks 训练策略，采用多轮学生-教师迭代训练以提升模型的泛化能力。在增量学习场景下，Li 和 Hoiem 提出了 Learning without Forgetting 方法[9]，其基本思路是利用旧任务的模型作为教师，并在新任

务训练过程中引入蒸馏损失，以保持旧任务的知识。LwF 在分类任务中得到了广泛应用，但由于未对旧类别特征进行显式建模，其性能通常存在一定的局限性。

二、 知识蒸馏在增量学习中的应用

在增量学习中，知识蒸馏主要用于缓解灾难性遗忘，主要的方法可归纳如下：

逻辑回归空间蒸馏方法方面，Hou 等人提出了 Less-Forgetting Learning 方法[75]，在 LwF 的基础上进一步对中间层特征进行蒸馏，以保持旧任务的判别能力。Wu 等人[76] 提出了 Bias Correction 方法，通过蒸馏旧模型的输出分布，减少类别不平衡对模型的影响。此外，Ahn 等人[77] 提出的 Knowledge Retention Network 通过蒸馏教师模型的特征关系，以保持旧任务的类间关系。Park 等人[78] 提出了 Relational Knowledge Distillation，通过约束样本间的关系信息，提高增量学习过程中知识的保留能力。

近年来，研究者开始关注结构化蒸馏，以提升增量学习的表现。例如，Douillard 等人[15] 提出的 PODNet 采用层间特征图对齐的方式进行蒸馏，以提高旧任务特征的可分性。此外，Zhu 等人[66] 提出了 Prototype-based Knowledge Distillation 方法，该方法通过存储和蒸馏类别原型，提高模型的稳定性。

三、 知识蒸馏与其他增量学习技术的结合

为了进一步提高增量学习的效果，许多研究结合了知识蒸馏与其他方法，如样本重放[8]、参数正则化[79, 80] 以及元学习[81]。其中，iCaRL[8] 结合了蒸馏损失与最近邻分类器，实现了有效的类别增量学习。此外，AANets[82] 通过动态蒸馏策略，根据任务变化调整蒸馏目标，提高模型的适应性。

第四节 显著性监督

显著性监督是一种利用显著性信息来增强模型学习能力的方法，在计算机视觉、自然语言处理以及增量学习等领域均得到了广泛应用。在增量学习场景中，显著性监督常用于强化特征学习，提高模型的稳定性和适应性[83, 84]。其核心思想是利用显著性信息作为额外的监督信号，以引导模型学习更具判别性的特征，并减少增量训练过程中的知识遗忘。

一、 显著性监督的基本方法

显著性信息通常由显著性检测模型或注意力机制生成，并作为额外的监督信号约束模型学习[85–88]。常见的方法可以分为以下两类：

显式显著性监督方法直接使用显著性图或注意力权重来指导特征提取。例如，Grad-CAM[89] 通过梯度反向传播计算输入图像对模型决策的重要性区域，以此生成显著性图，从而帮助模型在增量学习任务中关注关键特征。类似地，Zagoruyko 和 Komodakis[90] 提出了基于注意力蒸馏的策略，通过显式监督引导学生模型学习教师模型的显著性分布。

隐式显著性监督方法通过自监督学习、对比学习或注意力蒸馏等方式，在特征空间中保持显著性结构的稳定性。例如，He 等人提出的 Momentum Contrast[58] 通过对比学习的方式学习稳定的特征表示，这在增量学习中可用于保持旧任务的特征结构。此外，Liu 等人[82] 提出的自适应知识蒸馏框架利用注意力机制进行特征增强，从而提高增量学习的鲁棒性。

二、 显著性监督在增量学习中的应用

在增量学习中，显著性监督的主要应用包括以下几方面：

减少灾难性遗忘方面，显著性信息可以作为知识保持的一种机制，在新任务训练过程中约束旧任务的特征空间，从而减少重要特征的漂移。例如，Hou 等人[91] 提出的 LUCIR 方法通过显著性增强保持旧类信息，减少新任务训练时对旧任务特征的破坏。此外，Douillard 等人[15] 采用层间显著性对齐策略，使得增量学习过程中特征分布更加稳定。

增强模型鲁棒性方面，显著性监督有助于提高模型在类别不平衡和任务转移时的适应能力。例如，Liu 等人[92] 提出的 Mnemonics Training 方法通过优化存储样本的显著性分布，提升了在不均衡类别分布下的增量学习性能。此外，Zhao 等人[93] 研究了注意力引导的特征保持策略，证明了在增量学习过程中利用注意力机制能够提升模型在新任务上的泛化能力。

结合知识蒸馏方面，显著性信息可用于指导知识蒸馏，提高学生模型对重要知识的保留能力。例如，Jin 等人[94] 提出了一种基于显著性指导的知识蒸馏方法，该方法在教师-学生模型的蒸馏过程中对关键区域施加更大的权重，以确保重要知识的传递。此外，Shi 等人[95] 结合显著性监督与自监督学习，提高了无监督增量学习的性能。

第三章 基于掩码自编码器的类增量学习

本章介绍了一种基于掩码自编码器的类增量学习方法——双边 MAE 框架。该方法的核心思想是利用 MAE 的特性高效存储示例，并通过双边 MAE 融合机制缓解灾难性遗忘。实验结果表明，该方法在多个增量学习基准任务上优于现有最先进方法。通过消融实验分析了各组件对整体性能的贡献，验证了双边 MAE 在类增量学习场景下的有效性。

第一节 研究动机与贡献

类增量学习旨在按顺序学习新的分类任务，同时避免灾难性遗忘。该方法大致可分为三类，即基于重放的方法、基于正则化的方法和基于架构的方法。其中，基于重放的方法通过存储过去任务中的示例或生成合成样本进行重放，从而达到最先进的性能。在这项工作中，引入了掩码自编码器作为重放的基础模型。它只需要一小部分图像块就能重建整个图像，从而实现高效的示例存储。因此，与其它基于示例的方法相比，可以用同样有限的内存存储更多的示例。与之前的生成式方法相比，基于 MAE 的重放方法更稳定，因为它使用部分线索来推断全局信息，而全局信息与任务无关，在不同任务中遗忘较少。该方法通过固定的图像块缓解了 GAN 在不同任务中的不稳定生成效应。

掩码自编码器最初为在自监督学习场景中学习更好的特征表征而提出。在这项工作中将其视为高效的类增量学习器，同时提出了一种新颖的双边 Transformer 架构，用于高效重放示例。主要想法很简单：通过随机遮蔽输入图像的图像块训练模型来重构遮蔽的像素，MAE 可以为类增量任务提供一种新的自监督表征学习范式，从而使模型学习到更多通用的表征，这对于该任务而言至关重要。此外，利用带有分类标签的监督目标还能提高无监督 MAE 的训练效率和模型稳健性。同时，遮蔽后的输入通过提供数据的一个随机子集可以在分类中起到很强的正则化作用。

此外，本章进一步引入了通用属性知识的引导，以增强模型在不同任务间的泛化能力。通用属性知识指的是类别间共享的高层语义特征，例如形状、纹理或功能属性等，这些特征通常超越具体的类别标签，在不同任务中具有稳定

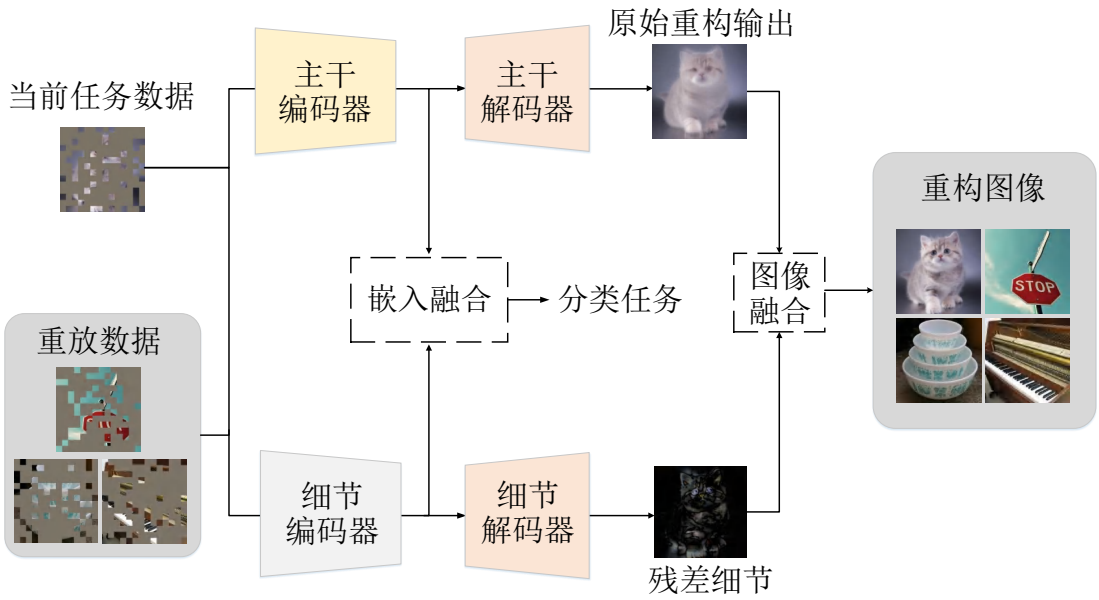

图 3.1 方法提出的用于高效 CIL 的双边 MAE。重放缓冲区包含从过去任务图像中选取的随机图像块，这比存储整幅图像更有效。将这些数据与当前任务中遮蔽的输入数据相结合，MAE 能够从遮蔽的输入中同时学习图像分类和重建。为了进一步改善重建的图像的质量并学习表征，方法使用了嵌入级和图像级的融合来学习更加稳定的表征以及包含更多细节的重建图像，从而用于 CIL。

的语义表达。

在学习新任务时，MAE 能够从示例的稀疏采样块中粗略地重构图像。这个过程能够使框架生成重建后的重放图像，但仍存在两个问题：生成的图像纹理往往不够精细和逼真，这减少了重放数据的多样性；在嵌入层面上，线性分类器缺乏来自低层次特征的信息。因此，方法为类增量任务引入了图像级和嵌入级融合的双边 MAE 框架。将互补的详细图像和重建图像融合在一起，可以用详细、高质量的数据分布来丰富不充分的重放数据，从而减轻灾难性遗忘。两个分支的嵌入层面融合还能保持嵌入的稳定性和多样性，因此框架能够在可塑性和稳定性之间取得更好的平衡。方法简要结构见图 3.1。

第二节 用于类增量任务的双边 MAE 框架

在本节中首先定义了类增量学习问题和基本的 MAE 模型，然后介绍了增量学习框架以及框架所基于的双边 MAE 架构。

一、 方法序言

类增量学习问题： CIL 旨在不断学习包含新类别的任务，同时避免或减轻对于旧任务的遗忘。在学习任务 $t \in \{1,2,...,T\}$ 的特定阶段，模型的训练仅能利用来自当前任务的数据 $\{(x_i^t, y_i^t)\}$，其中 x_i^t 表示任务 t 中的图像 i，y_i^t 表示对应的类别标签。一个 CIL 模型通常由一个特征提取器 F_θ 和一个常规的分类器 G_ϕ 构成，该分类器在遇到新任务时将进行扩张。在学习任务 $t+1$ 时，C_{t+1} 个新类将被添加到 G_ϕ。特征提取器 F_θ 首先将输入 x 映射到深度特征向量 $z = F_\theta(x) \in \mathbb{R}^d$，$d$ 为输出的特征表征的维度，之后统一的分类器 $G_\phi(z) \in \mathbb{R}^{C_{1:t}}$ 生成一个关于类别 $C_{1:t}$ 的概率分布，该分布将被用于预测输入图像 x。

在训练任务 t 时，模型的目标是尽量减少当前任务的损失，同时不降低之前任务的性能。减少遗忘的常用技术是保留一小部分先前任务的训练样本。令 ϵ 为先前任务样本的缓冲区。CIL 任务中一个关键的问题是重放数据的数量有限制。相比于当前任务 t 的全部数据而言，只有少量的旧任务类别样本是可用的（常用的设置是每一个类存储 20 个样本），这会带来新旧任务之间训练不平衡的问题。

一种用于分类的 MAE 框架： MAE 首先将输入图像 x 裁剪为不重叠的图像块，将一张完整图片 x 的图像块数目定义为 N_f。在完成分块后，MAE 随机将 N_f 个图像块中的一部分进行遮蔽，遮蔽的比例为 $r \in [0,1]$，剩余 $N = \lfloor N_f \times (1-r) \rfloor$ 个图像块。随后，这些采样后大小为 $K \times K$ 的像素块通过一个 MLP 被映射为 D 维的视觉嵌入。其与一个类别标记拼接后，得到大小为 $\mathbb{R}^{(N+1)\times D}$ 的张量。在对原始的块进行位置编码后，该输入将被送入 MAE transformer 编码器。这项操作保持了嵌入的形状不变。输出的类别标记嵌入能够通过交叉熵损失 $\mathcal{L}_t^{\text{cls}}$ 用于分类，如图3.2所示。

对于 MAE 解码器，可学习的掩码标记被插入到嵌入中以代替遮蔽的图像块，同时 MAE 编码器的输出形状从 $\mathbb{R}^{(N+1)\times D}$ 变为 $\mathbb{R}^{(N_f+1)\times D}$。虽然解码器不会被用于分类，但其有助于网络将图像级别的重建监督反向传播到嵌入级别。这可以稳定图像嵌入并有利于优化整个过程。此外，解码后重建的图像可提供更丰富、更高质量的重放数据。为了限制计算量，方法使用单层 Transformer 块进行解码。通过编码器得到的额外分类损失可以加快收敛速度，提高训练过程中的重构效率。输入图像 x 与重建图像 $\hat{x}$ 的均方误差被用作重建损失函数 $\mathcal{L}_t^{\text{rec}}(x, \hat{x})$。

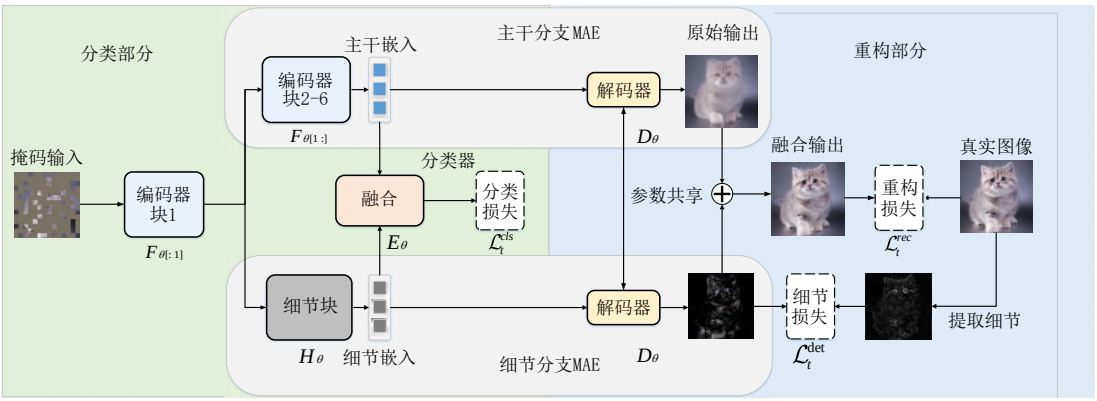

图 3.2 用于 CIL 的双边 MAE 总体框架。遮蔽后的输入经过两个分支，其中嵌入级的融合用于分类，图像级的融合用于重建。整张图像可以通过一小部分输入的图像块生成，同时重建的图像可以用于重放。

二、 利用 MAE 高效存储示例

每个任务训练完成后，保存一小部分样本图像并对其进行随机遮蔽。通过保持相同的存储容量，就能为每个类保存更多的重放数据，因为每个样本占用的空间更少。例如，相比于传统的基于重放的方法而言，以 0.75 作为遮蔽比例能够使保存四倍数量的（可重建的）样本。

令 S 和 P 表示图像和块的尺寸。编码器将输入图像切分为 $\frac{S}{P} \times \frac{S}{P}$ 个图像块。对于每个没有被遮蔽的图像块，保存其 2D 的索引 (i,j)。仅需一个字节来存储索引，因为索引的范围小于 255。两个用于存储 2D 索引的额外字节与存储的图像块相比可以忽略不计。遮蔽比例为 0.75 的大小为 224× 224 的图像仅仅占据 36.75KB 的存储空间。对于 $P = 16$ 的情况，保存的图像块的数量为 $(1 - 0.75) \times (\frac{224}{16})^2 = 49$，索引占据的存储空间仅为 98B。

三、 双边 MAE 融合

为进一步提高重建质量和嵌入的多样性，提出了一种双分支 MAE，以学习全局和细节的图像分类以及重建的知识。在图3.2中阐述了整体框架。嵌入层面的双边融合旨在提高表征的多样性。图像级的重构学习可为 CIL 提供高质量的重放数据和稳定的自监督。

嵌入融合： 在以下内容中，使用 $F_{\theta[:1]}$ 与 $F_{\theta[1:]}$ 表示 Transformer 编码器中第一个以及之后的块。令 H_θ 和 E_θ 表示图3.2中的细节块和嵌入融合模块，这些结构

为标准的 MLP 层以及注意力块。分类损失的计算公式为：

$$
\begin{aligned}
f &= F_{\theta[:1]}(\text{mask}(x,r)) && (3.1)\\
z &= E_\theta(F_{\theta[1:]}(f), H_\theta(f)) && (3.2)\\
\mathcal{L}_t^{\text{cls}}(x,y) &= \mathcal{L}_{\text{ce}}(G_\phi(z), y), && (3.3)
\end{aligned}
$$

其中 $\text{mask}(x,r)$ 表示将比例为 r 的随机遮蔽操作施加到图像 x，f 为第一个编码器块提取到的嵌入，这也是双边 MAE 两个分支的输入，$G_\phi(z)$ 为用于交叉熵损失的预估类别分布。

基于细节损失的图像融合： 对于细节头和相应的损失，可以发现在频域中工作更容易使网络关注高频细节，而这正是细节分支应该重建的。方法定义了一个频率掩蔽函数 $M(\cdot)$，它能将参数（一个图像块）转换到频域，然后使用一个围绕原点的圆形掩蔽器将低频分量掩蔽。如图3.2所示，MAE 解码器由模型的两个分支共享，因为它们具有相似的重建任务以及相同的输入和输出形状。令 D_θ 表示共享的解码器，之后两个分支图像级别的输出以及重建损失可以表示为：

$$
\begin{aligned}
f &= F_{\theta[:1]}(\text{mask}(x,r)) && (3.4)\\
x' &= D_\theta(F_{\theta[1:]}(f)) && (3.5)\\
x'' &= \text{ifft2}(M(D_\theta(H_\theta(f)))) && (3.6)\\
\hat{x} &= x' + x'' && (3.7)\\
\mathcal{L}_t^{\text{rec}} &= \mathcal{L}_{\text{mse}}(x,\hat{x}), && (3.8)
\end{aligned}
$$

其中 x' 和 x'' 分别为主要的和残差的细节输出，ifft2 为逆快速傅立叶变换。

细节损失 $\mathcal{L}_t^{\text{det}}$ 还利用了频率掩蔽函数 M 来比较细节分支的输出与输入图像的两个 MAE 重建结果之间的差值：

$$
\begin{aligned}
\hat{x}_1 &= D_\theta(F_\theta(\text{mask}(x,r_1))) && (3.9)\\
\hat{x}_2 &= D_\theta(F_\theta(\text{mask}(x,r_2))) && (3.10)\\
\mathcal{L}_t^{\text{det}} &= ||M(D_\theta(H_\theta(f))) - M(\hat{x}_2 - \hat{x}_1)||_1, && (3.11)
\end{aligned}
$$

其中 $\hat{x}_1$ 和 $\hat{x}_2$ 是两张使用了不同遮蔽比例 r_1 和 r_2 的重建图像（从计算图剥离）。残差 $\hat{x}_2 - \hat{x}_1$ 被用作在频域中损失 $\mathcal{L}_t^{\text{det}}$ 内的细节分枝。

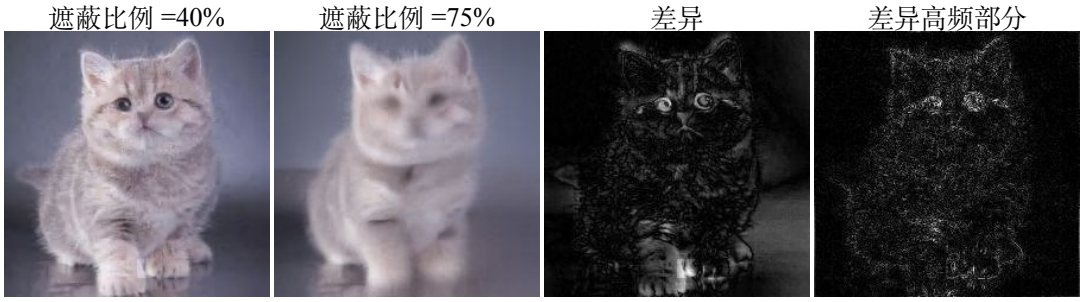

图 3.3 一个关于从不同遮蔽比例 r_1 和 r_2 的重建结果中提取细节图像的示例。第三幅图像来自前两幅图像的差值，最后一幅图像是从第三幅图像中提取的高频分量。

Algorithm 1 双边 MAE 的伪代码

输入: 任务总数 T，任务 t 的训练样本 $D_t = \{(x_i, y_i)\}_t$，初始模型 Θ^0，重放缓冲区 ϵ，掩码比例 r、r_1、r_2。
输出: 模型 Θ^T

for $t \in \{1, 2, ..., T\}$ **do**
  $\Theta^t \leftarrow \Theta^{t-1}$
  $R_t \leftarrow$ 重构旧样本 (ϵ_t, r)
  while 未收敛 **do**
    $(x, y) \leftarrow$ 采样 (R_t, D_t)
    $(\mathcal{L}_t^{\text{cls}}, \mathcal{L}_t^{\text{rec}}) \leftarrow$ 双边 MAE (x, y)
    $(\hat{x}_1, \hat{x}_2) \leftarrow$ 掩码并重构 (x, r_1, r_2)
    $\mathcal{L}_t^{\text{det}} \leftarrow$ 计算细节损失 $(\hat{x}_1, \hat{x}_2)$
    通过最小化公式 3.12 中的 $\mathcal{L}_t$ 训练 Θ^t
  end while
end for

图3.3举例说明了这一点。

$\mathcal{L}_t^{\text{cls}}$、重建损失 $\mathcal{L}_t^{\text{rec}}$ 以及细节损失 loss $\mathcal{L}_t^{\text{det}}$ 的加权和构成了训练的总损失：

$$\mathcal{L}_t = \lambda_{\text{cls}}\mathcal{L}_t^{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_t^{\text{rec}} + \lambda_{\text{det}}\mathcal{L}_t^{\text{det}}. \tag{3.12}$$

方法的伪代码在算法1中给出。

第三节 实验结果与分析

一、 性能指标与实现

数据集以及设置： 方法在三个数据集上进行了实验：CIFAR-100[96]、ImageNet-Subset 和[97]，以评估方法的性能。对于 CIFAR-100 和 ImageNet-Subset，分别在包含 10 个、20 个和 50 个任务的场景中进行了测试，每个任务的类别数相同。同时评估了 ImageNet-Full 的 10 任务设置，其中每个任务都包含 100 个新类别。

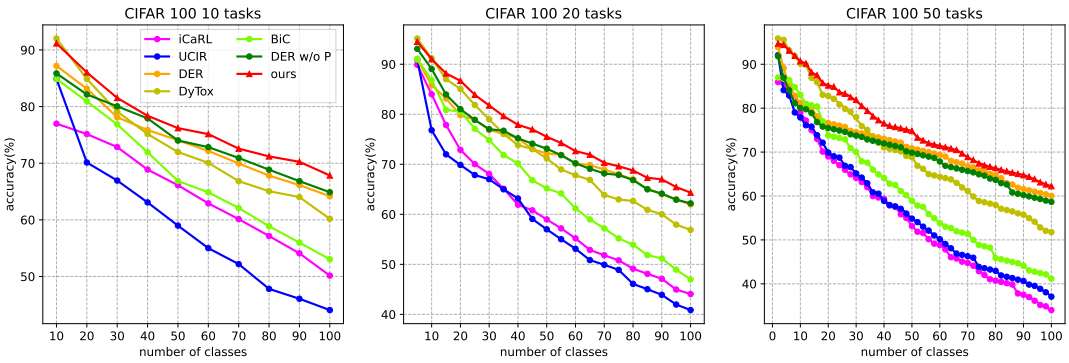

图 3.4 在 CIFAR-100 数据集 10 个、20 个和 50 个任务场景下增量任务性能的变化。

为了衡量训练期间完成所有任务后的总体准确率，方法报告了每个任务后所学任务的平均准确率，以及在增量学习结束时所有任务的准确率。

实现细节： 对于所有的数据集使用了相同的网络。

模型从零开始训练以防止数据泄露，批量大小为 1024，使用初始学习率为 1×10^{-4} 和带有余弦衰减的 Adam[98]。式3.12的损失权重设置为 $\lambda_{cls} = 0.01$，$\lambda_{rec} = 1.0$，$\lambda_{det} = 1.0$。遮蔽比例设置为 $r = 0.75$，$r_1 = 0.75$ 以及 $r_2 = 0.4$。每一个任务训练 400 轮。对于文献中基于示例的方法，为每个类别存储 20 个样本。

编码器使用了 5 个 Transformer 块，解码器使用了 1 个 Transformer 块。所有的 Transformer 块拥有相同的编码维度 384 以及 12 个自注意力头。这种设计不同于原始的 MAE，因为其更加轻量。方法保存图像块所占用的内存量与其它每类存储 20 幅完整图像的方法相同。例如，选择 80 幅图像，使用 0.75 的遮蔽比例随机保存每幅图像中 25% 的图像块（因此只占用与 20 幅完整图像相同的空间）。细节模块使用 3 层 MLP 实现，维度为 384。

二、 与最先进方法的对比

在本节中将目前最先进的方法与本章方法进行了对比，包括 DER[99] 和 DyTox[100]。在所有的图和表中，“DER w/o P” 表示不含剪枝的 DER[99]，因此其在不同任务中能够拥有更多参数。DyTox[100] 同样使用了 Transformer 结构，使用其官方代码库以复现结果。

CIFAR-100： 表3.1中给出了平均准确率、最后一个任务后的准确率以及平均遗忘程度。显然，在每种设置下，本章方法都远优于其他方法。对于较长的任务序列，双边 MAE 能够充分利用自监督重建机制与更为丰富的重放数据，从而

表 3.1 CIFAR-100 数据集 10 个、20 个和 50 个任务场景下的平均准确率（%）、最后阶段准确率（%）以及遗忘程度 F（%）。

方法	10 任务			20 任务			50 任务		
	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓
iCaRL[8]	65.27	50.74	31.23	61.20	43.75	32.40	56.08	36.62	36.59
LUCIR[91]	58.66	43.39	35.67	58.17	40.63	37.75	56.86	37.09	38.13
BiC[76]	68.80	53.54	28.44	66.48	47.02	29.30	62.09	41.04	34.27
PODNet[15]	58.03	41.05	41.47	53.97	35.02	36.70	51.19	32.99	40.42
DER w/o P[99]	75.36	65.22	15.02	74.09	62.48	23.55	72.41	59.08	26.73
DER[99]	74.64	64.35	15.78	73.98	62.55	23.47	72.05	59.76	26.59
DyTox[100]	75.47	62.10	15.43	75.10	59.41	21.60	73.89	57.21	24.22
本章方法	**79.12**	**68.40**	**12.17**	**78.76**	**65.22**	**14.39**	**76.95**	**63.12**	**18.34**

在知识保持与新任务适应之间实现更优平衡。得益于其结构设计，双边 MAE 相较于现有方法表现出更强的抗遗忘能力，遗忘率明显更低。

总体准确率变化如 图3.4 所示。在使用相同大小的重放存储空间的前提下，本章方法在三个典型类增量学习场景中均展现出稳定优势，在最后一项任务后的准确率平均比 DyTox 提高约 6%，进一步证明了所提出方法在长期任务序列下的有效性与鲁棒性。

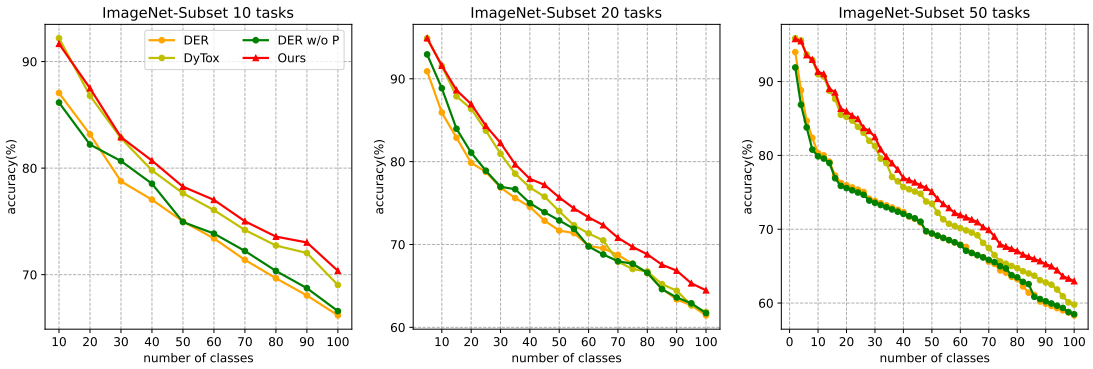

图 3.5 在 ImageNet-Subset 数据集增量任务性能的变化。

ImageNet-Subset 与 ImageNet-Full： 在表3.2和表3.3中分别报告了 ImageNet-Subset 和 ImageNet-Full 的性能。在包含 10 个、20 个和 50 个任务的设置中，方法在最后一个任务后的准确率绝对增益分别比 DyTox[100] 高出 1.19%、2.53% 和 2.85%。每个阶段的平均准确率较高，遗忘率较低，这也证明了方法在减轻遗忘方面的有效性。同样在图3.5中说明了 ImageNet-Subset 的性能变化。在第一个任务中，方法与 DyTox 的准确率相似，但在后面的任务中，本章方法超过了所有其他方法，尤其是在长任务序列中。在更大规模的 ImageNet-Full 中，双边 MAE

表 3.2 在 ImageNet-Subset 数据集 10 个、20 个和 50 个任务场景下的平均准确率（%）、最后阶段准确率（%）以及遗忘程度 F（%）。

方法	10 任务			20 任务			50 任务		
	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓
BiC[76]	64.96	55.07	31.32	59.40	49.35	34.70	53.75	44.56	40.23
PODNet[15]	63.44	51.75	35.63	55.11	45.37	41.70	51.72	42.94	44.65
DER w/o P[99]	77.18	66.70	14.86	72.70	61.74	20.76	70.44	58.87	24.20
DER[99]	76.12	66.06	15.09	72.56	61.51	20.46	69.77	58.19	25.35
DyTox[100]	77.15	69.10	14.66	73.13	61.87	17.32	71.51	60.02	20.54
本章方法	**79.54**	**70.29**	**12.04**	**75.20**	**64.40**	**14.89**	**74.42**	**62.87**	**17.22**

表 3.3 ImageNet-Full 数据集 10 个增量任务场景下的结果。

方法	top-1		top-5	
	平均↑	最后↑	平均↑	最后↑
iCaRL[8]	38.40	22.70	63.70	44.00
Simple-DER	66.63	59.24	85.62	80.76
DER w/o P[99]	68.84	60.16	88.17	82.86
DER[99]	66.73	58.62	87.08	81.89
DyTox[100]	71.29	63.34	88.59	84.49
本章方法	**74.76**	**66.15**	**91.43**	**87.13**

在所有指标上都明显超过其他方法约 3%。

三、 消融实验

不同组成部分的消融实验： 双边 MAE 包括自监督重建任务、重放数据生成以及用于图像级和嵌入级融合的双边 MAE 分支。在表3.4中对这三个因素进行了分析。方法中的这三个主要部分具有不同的功能，它们相互配合，使性能比基线提高了约 6%。观察到: (a) 更高质量的重放数据对性能有直接贡献，采用 $r = 0.75$ 的掩码比率能以与基线相同的存储成本获得 4 倍的重放数据。(b) 重建损失是一种有效的自我监督，能使平均准确率提高约 2%。(c) 双边架构通过提高重放数据生成质量以及引入图像和嵌入级监督取得良好的效果。

遮蔽比例： MAE[17] 的一个关键参数是遮蔽比例。在 r 的选择上需要权衡：过大的 r（如 0.95）会导致重建效果不佳，从而影响重放数据的质量，造成更严重的遗忘。然而，过小的 r 产生的额外重放数据量有限（例如，当 r 为 0.10 时，只能承受约 11% 的额外重放数据）。表3.5中的结果表明，对于双边 MAE 而言，$r = 0.75$ 是一个很好的折衷。为了验证生成的重放数据的质量，结果中还包含使用原始图像代替生成图像进行重放的准确率。表3.5第 2 行和第 4 行的结果显示，

表 3.4 在包含 10 项任务的 CIFAR-100 设置中，对提出方法的每个组成部分进行的消融实验。“重放” 表示使用 MAE 生成的数据进行重放，“重构” 表示应用自监督重建损失，“双分支” 表示引入 MAE 的细节分支。

方法	重放	重构	双分支	平均	最后
基准				73.40	62.31
改进	✓			75.88	64.35
	✓	✓		77.48	66.54
	✓	✓	✓	79.12	68.40

表 3.5 对遮蔽率和生成数据质量的消融实验。实验在包含 10 项任务的 CIFAR-100 设置上进行。在最后一行中重放了部分真实图像，其存储容量与使用比例为 $r_1 = 0.75$ 时相当。

r	数据源	平均	最后
0.60	生成	77.50	67.37
0.75	生成	79.12	68.40
0.90	生成	77.12	67.02
N/A	真实	79.57	68.87

方法获得了高质量的图像，与重放真实图像相比，准确率相差不到 0.5%。

频域中的细节损失： 通过将嵌入从空域转换到频域来实现细节损失。这样做的目的是为了集中处理高频信息，这与 MAE 细节分支的学习目标相匹配。如表3.6所示，进行这样的转换是有益的，因为它在最后一项任务中带来了超过 2% 的增益。

细节损失中关于 r_1 和 r_2 的消融实验： 在所有实验中，将 r_1 设为 0.75 作为参考，同时改变用于计算细节损失真值的 r_2。对 r_2 的权衡是，如果 r_2 较大，则按照遮蔽比例 r_1 和 r_2 重建的结果差异较小，因此，监督信号中关于细节损失的信息很少，细节分支的影响也会减小。另一方面，过小的 r_2（如 0.10）会保留主分支重建图像的大部分残留部分，这可能会导致对主分支的监督较弱，减慢其训练速度。

在图3.6中展示了一系列 r_2 值的结果。这些结果表明，约 0.40 的 r_2 值可以很好地为 MAE 细节分支提供监督。

模型与示例的大小： 为了比较不同方法的有效性，通常使用参数数量相同或相似的模型，并使用相同数量的示例。

表 3.6 对于细节头的消融实验。实验在包含 10 项任务的 CIFAR-100 设置上进行，以百分比的形式报告了 top-1 准确率。“域” 表示损失应用于哪一个域。

域	平均	最后
空域	77.45	65.93
频域	79.12	68.40

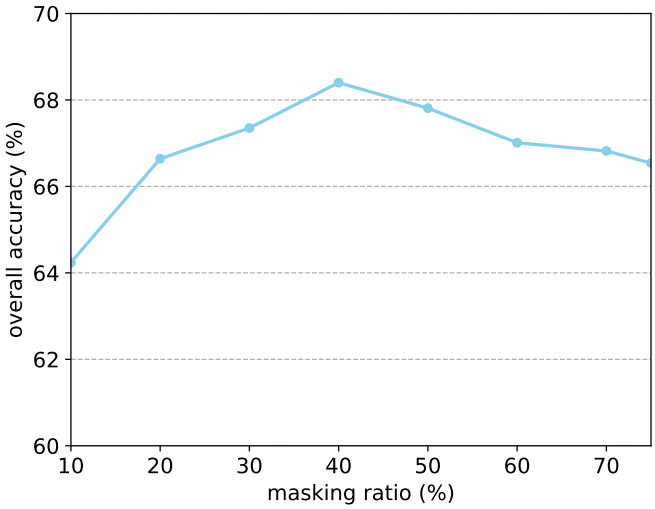

图 3.6 关于 $\mathcal{L}_t^{\text{det}}$ 中遮蔽比例 r_2 的消融实验。另一个遮蔽比例 r_1 设置为 75% 作为参考。

在方法中，对原始 MAE 进行了调整，使其更加轻量，参数数量与 DyTox 相当甚至更少，如表3.7所示。将屏蔽率默认设置为 75%，每类保存 80 个示例，因此模型和示例的存储大小相似，因为存储的图像块所需的空间与基线完全相同。

关于有效缓冲区大小的消融实验： 在表3.8中，使用相同的缓冲区大小，通过遮蔽 DyTox 中输入图像的图像块将方法与 DyTox 进行了比较。所有三行都使用相同大小的内存来存储示例。在使用 80 个遮蔽率为 75% 的示例时，DyTox（最后一行）的性能比使用 20 个完整图像示例时更好。它的性能仍然劣于本章方法，这表明性能提升并不仅仅来自于额外的示例，还来自于将 MAE 与细节分支整合到双边架构中。

重建分析： 在图3.7中展示了在 ImageNet-Subset 的包含 10 个任务的设置中进行图像重建的结果。左栏显示的是从任务 1、4、7 和 10 中随机选取的图像。双边 MAE 以与任务无关的方式学习重建图像，这有助于在学习之前就为未来的任务生成合理的结果。MAE 的细节分支学习重建高频细节，以补充主分支。主分支的结果有时缺乏特定样本的特征，但在提出的细节分支的帮助下，重建结果

表 3.7 模型大小的比较。这里比较了两个版本的双边 MAE 模型和对比模型。实验在包含 10 个任务的 CIFAR-100 上进行。

方法	参数量 (M)	平均↑	最后↑	F↓
DER w/o P	112.27	75.36	65.22	15.02
DyTox	10.73	75.47	62.10	15.43
本章方法 (MLP size = 1536)	12.89	79.12	68.40	12.17
本章方法 (MLP size = 768)	9.35	78.36	67.52	12.90

表 3.8 在内存使用量相同的情况下对于有效缓冲区大小的消融实验。Dytox† 表示直接对存储的图像示例应用遮蔽比例 $r = 75\%$，以便使 DyTox 的示例数量和存储大小与方法的设置相同，从而实现公平比较。

方法	存储	存储大小	25% 比例	图像	准确率 (%)
本章方法	80	1x	✓		68.40
DyTox	20	1x		✓	62.10
DyTox†	80	1x	✓		65.46

更加准确，并能提供更好的重放数据。

表 3.9 在 CIFAR-100 包含 10 个任务的设置中，双边 MAE 框架与其它节省内存的方法进行了比较。存储表示每个类所需的存储空间（以 KB 表示）。

指标	存储	平均↑	最后↑
Latent replay (CVPRW'20)[101]	-	62.44	51.30
MCIL (CVPR'20)[92]	60	63.25	53.12
Down-scaled (TNNLS'21)[102]	60	67.04	55.40
JPEG compression (ICLR'22)[103]	60	72.34	61.32
CIM (CVPR'23)[104]	60	75.30	63.05
本章方法	60	79.12	68.40

更通用的表征有助于 CIL： 这里按照 PASS[66] 的方式计算了不同方法的特征空间密度指标[105]：$\pi = \pi_{intra} / \pi_{inter}$，其中 π_{intra} 表示同一类别中的平均余弦相似度，π_{inter} 表示不同类别中的平均余弦相似度。特征空间密度的增加与数据偏移情况下更强的泛化能力相关[66]。随后比较了所有任务训练后的特征空间密度，如上图3.8所示。

很明显，本章的方法得出的密度显著高于其他方法。

与其它高效重放方法的消融实验： 在表3.9比较了框架中 MAE 生成的重放样本与各种节省内存的方法生成的样本，这些方法分别基于隐重放（latent replay）、合成示例（synthesized exemplars）、缩放（down-scaling）、JPEG 图像压缩（JPEG image compression）和 CIM（前景提取和背景压缩）。所有这些方法都使用了相

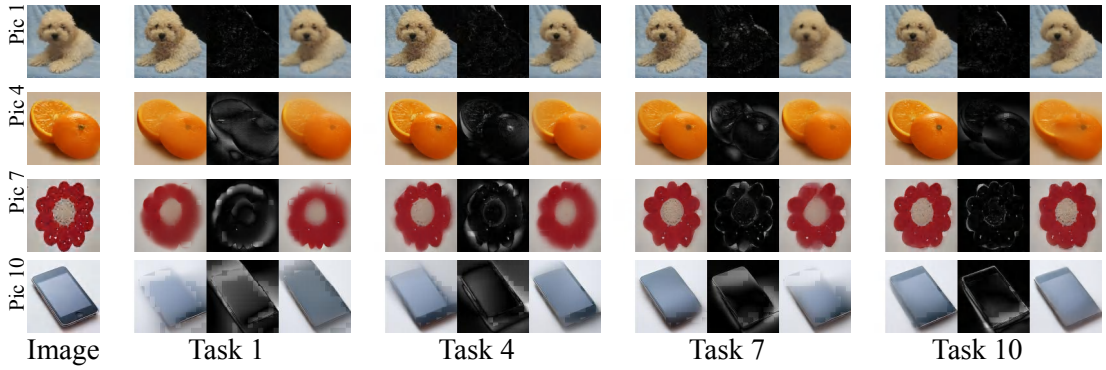

图 3.7 在包含 10 个任务的设置中对 ImageNet 子集中的图像进行重建得到的结果。从任务 1、4、6 和 10 中选取的四幅原始图像如左图所示。其余各栏显示的是使用双边 MAE 组合（左）、仅使用 MAE 的细节分支（中）和仅使用 MAE 的主分支（右）重建的图像。可以看到训练进行到后期模型仍可以保持一定的图像重构能力。

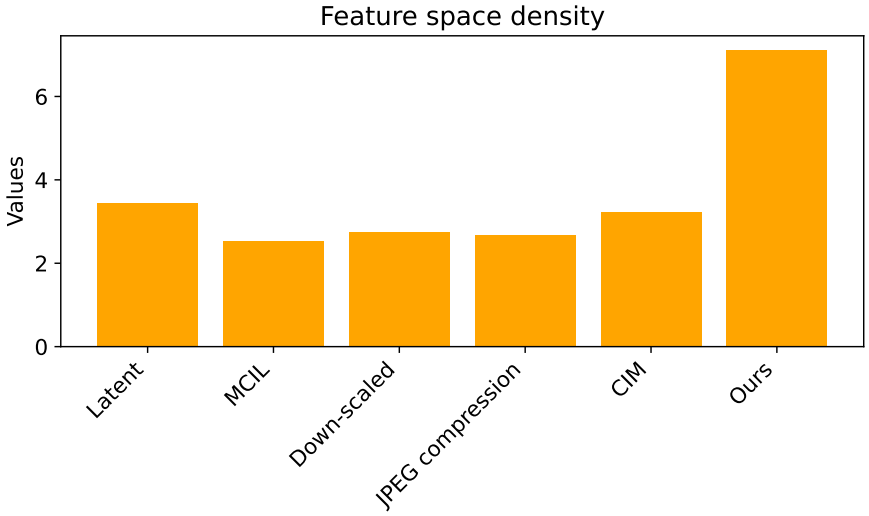

图 3.8 对于特征空间密度 π 的比较。

同的存储量（除了 Latent Replay 使用了一个带有 450 万参数的 GAN），而本章的方法则始终获得更高的性能。

第四节 本章小结

本章介绍了基于掩码自编码器类增量学习的研究动机与贡献。基于重放的方法通过存储过去任务中的示例或生成合成样本来减轻灾难性遗忘，但其在图像质量和生成模型的稳定性方面存在不足。针对这些问题，方法提出了基于掩码自编码器的重放方法，该方法在示例存储和生成数据的质量上具有显著优势。与传统的生成式方法相比，MAE 提供了一种更稳定的重放方式，并能有效减轻遗忘现象。

第四章　基于细粒度知识选择的无示例类增量学习

上文介绍了一种基于掩码自编码器的类增量学习方法——双边 MAE 框架。该方法的核心思想是利用 MAE 的特性高效存储示例，并通过双边 MAE 融合机制缓解灾难性遗忘。实验结果表明，该方法在多个增量学习基准任务上优于现有最先进方法。通过消融实验分析了各组件对整体性能的贡献，验证了双边 MAE 在类增量学习场景下的有效性。

尽管双边 MAE 框架在性能上取得了良好效果，但仍存在对旧类表示依赖较强、对知识表达粒度控制不足等问题，影响了模型的可扩展性和泛化能力。因此，为进一步提升模型在无示例增量学习场景下的适应能力，本章提出了一种基于细粒度知识选择的无示例类增量学习方法。该方法引入分块级知识选择机制与原型恢复机制，从历史任务中提取并维护关键知识，在促进模型对新类别适应能力的同时，有效保留对旧类别的判别能力。实验结果表明，所提方法在多个标准增量学习基准任务上均显著优于现有方法，且消融实验验证了各模块对整体性能的积极贡献。

第一节　研究动机与贡献

近年来，深度神经网络在计算机视觉任务中取得了显著的进展[106–108]，尤其是在图像分类、目标检测和语义分割等任务上展现了卓越的性能。然而，现有的深度学习方法在处理无示例类增量学习任务时仍然面临重大挑战。任务要求模型在学习顺序任务时不保留旧任务样本，这种设置符合现实世界中数据隐私保护和存储限制的需求，但同时也会导致严重的灾难性遗忘问题。因此，如何在不依赖旧任务样本的情况下有效保留已学知识，是研究中的核心问题。

为了解决灾难性遗忘，许多研究者提出了不同的方法。其中，知识蒸馏是一种常见且有效的策略，主要通过约束当前模型与旧模型之间的表示差距来减少遗忘。然而，在当前设置下，现有知识蒸馏方法存在两个主要问题。任务稳定性与可塑性的矛盾：知识蒸馏倾向于保持模型的稳定性，但这会限制模型对新任务的学习能力，影响其可塑性。分类器偏差问题：由于增量学习过程中旧任务样本不可用，导致模型在新任务上训练时，旧类别的决策边界发生偏移，进

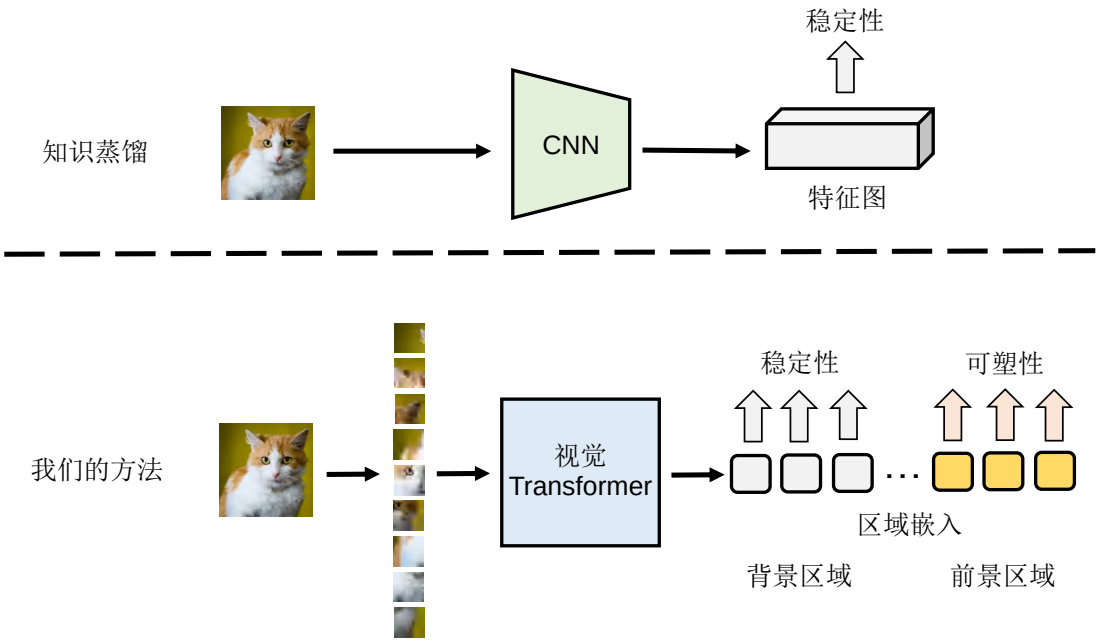

图 4.1 传统的知识蒸馏与基于视觉 Transformer 架构的分块级细粒度知识选择方法之间的比较。它将图像视为一个整体，而分块嵌入使方法能够在不同的局部区域之间实现更好的稳定性和可塑性平衡。此外，还提出了一种与任务无关的细粒度原型恢复方法，以更好地重现旧知识。

而影响模型在旧任务上的表现。

针对上述挑战，提出了一种新颖的框架，充分利用视觉 Transformer 的分块表示特性，并引入分块级知识选择与原型恢复机制，以有效缓解灾难性遗忘问题，如图 4.1所示。

方法提出了分块级知识选择机制。不同于传统知识蒸馏方法在整个图像级别进行蒸馏，利用视觉 Transformer 对输入图像的分块表示能力，根据每个分块与 [CLS] 标记嵌入的相似度来衡量其任务相关性。对于前景分块，减少蒸馏的正则化强度，增强其可塑性，以便更好地学习新任务信息；对于背景分块，强化正则化约束，使其在不同任务之间保持稳定的特征表示，从而增强模型的知识迁移能力。其次，方法提出了一种基于原型恢复的分类器重放策略。为了缓解分类器偏差问题，方法引入了一种基于类别原型的恢复方法。首先，计算每个样本到其类别中心（原型）的偏移距离，并进行正则化，以维持类别内数据的分布一致性。随后，利用当前任务的原型信息来恢复旧类别的原型，从而更准确地重建旧任务的决策边界，避免传统高斯分布假设导致的偏差问题。

此外，方法还引入了通用属性知识的引导，增强了模型在不同任务间的泛

化能力。通用属性知识在这里指的是类别间共同具备的前景与背景间的语义区分。通过将这些通用属性融入模型，促进知识迁移并减少遗忘。方法在多个基准数据集（CIFAR-100、ImageNet-Subset 等）上进行了广泛实验。结果表明，所提出的方法相比于现有方法，在保持旧任务知识的同时，能够更好地学习新任务，在多个主干网络（如 ViT、ResNet-50、MobileNetV2）上均取得了更好的性能。

第二节 细粒度知识选择方法

一、 基础知识

问题定义与分析。类增量学习顺序地学习不同的任务。这些任务中的每一个都与先前的任务没有重叠的类。令 $t \in \{1,2,..T\}$ 表示增量学习任务，其中 T 是所有任务的数量。训练数据 D_t 包含 C_t 个类别，具有 N_t 个训练样本 $\{(x_t^i, y_t^i)\}_{i=1}^{N_t}$。$x_t^i$ 表示图像，$y_t^i \in C_t$ 是其类别标签。

大多数类增量学习的深度网络可以分为两个部分：特征提取器 F_θ 和分类器 G_ϕ，其中分类器会随着每个新任务 $t+1$ 的增加而扩展，以包括类别 C_{t+1}。输入 x 通过特征提取器 F_θ 转换为深度特征向量 $z = F_\theta(x) \in \mathbb{R}^d$，然后使用统一的分类器 $G_\phi(z) \in \mathbb{R}^{|C_t|}$ 以学习一个关于类别 C_t 的概率分布，以预测 x 的标签。

类增量学习要求模型在任何训练任务中都能对来自先前任务的所有已学习样本进行分类。换言之，模型在执行任务 t 时应该保留对属于任务 $t' < t$ 的类别样本进行分类的能力。考虑到这些要求，非示例类增量学习施加了一个额外的约束，即模型必须在不使用任何来自先前任务样本的情况下学习每个新任务。大多数相关方法都以一个基本目标进行监督，该目标是最小化定义在当前训练数据 D_t 上的损失函数（例如交叉熵损失函数）。

视觉 Transformer 架构。

DyTox[100] 已经证明，视觉 Transformer 在 CIL 中是有效的，因为其动态任务相关的标记可以轻松地适应不同的任务。在本章中，发现了视觉 Transformer 的一个重要特性，它可以促进 CIL 并自适应地减轻新任务中的遗忘：即图像的分块级表示。以下是对视觉 Transformer 过程的回顾：

视觉 Transformer 首先将输入图像 x 裁剪成 $K \times K$ 个不重叠的分块，用 N 表示完整图像 x 中的分块数量。在此操作之后，这些分块通过一个 MLP 层映射到维度为 d 的视觉嵌入。将这些嵌入与形状为 $\mathbb{R}^d$ 的类别标记 [CLS] 连接起来，得到一个大小为 $\mathbb{R}^{(N+1)\times d}$ 的张量。

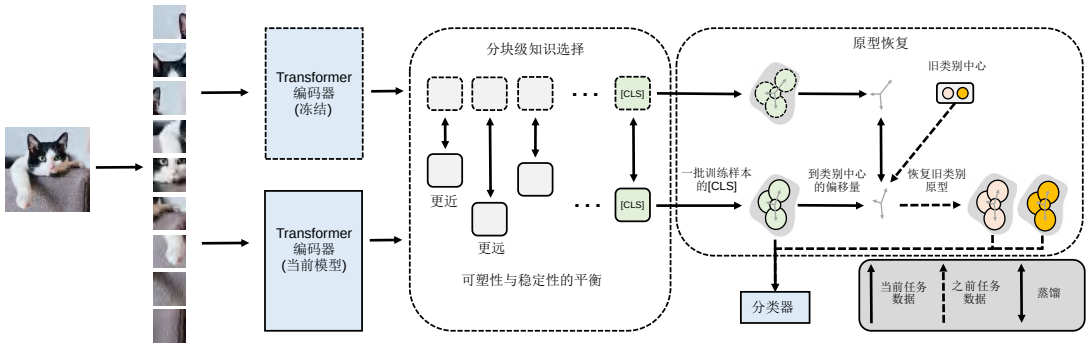

图 4.2 方法细粒度知识选择和恢复框架的示意图。分块嵌入通过不同的权重进行正则化。首先训练当前网络以保持与旧网络相似的原型分布，并用旧类中心和当前任务原型恢复旧原型。这两种原型被送入分类器，以减少任务偏差。

在对原始分块位置进行位置编码后，输入被传递到视觉 Transformer 的编码器。

每个编码器 Transformer 块有两个顺序部分：自注意力层和前馈层。在每个部分之前都会应用层归一化。

这些操作保持相同的嵌入尺寸，即 $\mathbb{R}^{(N+1)\times d}$，并为每个区域生成分块级表示。

从处理结果中得到的类别标记嵌入可以用于分类，通过交叉熵损失 $\mathcal{L}_t^{\mathrm{CIL}}$ 进行训练。采用线性分类器和 softmax 操作来预测每个学习类别的概率。

二、 分块级知识选择

在学习任务 t 时，仅有 D_t 对模型可用。先前的 EFCIL 方法中的基础知识蒸馏，如 PASS[66] 和 SSRE[109]，直接使用当前数据来维持任务间的稳定性。这种操作没有考虑当前任务样本与旧任务样本之间的语义差距，从而导致在减轻对旧任务的遗忘效果不佳。为了解决这个问题，这里重新思考了在分块化图像中使用知识蒸馏的方法。一个自然的想法是，在应用知识蒸馏时为每个分块分配不同的权重，因为它们对分类任务的重要性不同：前景区域的分块通常包含更多与任务相关的上下文，而背景中的分块则主要具有任务无关的像素和随机信息。

考虑到在 D_t 上平衡模型的可塑性和稳定性的双边策略，将其分为两步实施：(a) 定义一个度量标准，用于评估每个分块与当前任务 t 的相关性，(b) 对每个分块应用分块级知识选择，使用这些分块特定的权重来进行当前模型和旧模型之间的知识蒸馏。方法整体框架见 图4.2。

为了简化问题而不失一般性，采用 [CLS] 标记的嵌入 $P_{t,cls}$ 和每个图像分块 $P_{t,i}$ 之间的 L2 距离来计算它们对任务 t 的重要性：

$$W_i = \frac{1}{||P_{t,cls} - P_{t,i}||_2 + \epsilon}. \tag{4.1}$$

为接近 [CLS] 标记的分块分配更大的 W_i，并将每个 W_i 除以它们的最大值进行归一化，得到 w_i。ϵ 设置为 1×10^{-8} 以避免分母中出现零值。分块级知识选择 $\mathcal{L}_t^{pks}$ 被定义为：

$$\mathcal{L}_t^{pks} = \sum_{i=1}^{N} w_i ||P_{t,i} - P_{t-1,i}||_2 + ||P_{t,cls} - P_{t-1,cls}||_2, \tag{4.2}$$

$P_{t,i}$ 表示任务 t 的模型计算得到的第 i 个分块的嵌入表示，计算当前模型 F_{θ_t} 和旧模型 $F_{\theta_{t-1}}$ 之间的嵌入表示的 L2 距离。

三、 原型恢复

首先描述原型偏移和类别中心的定义。特征提取器 F_θ 从任务 t 的输入图像 X_t 中计算出表示 $z \in \mathbb{R}^d$，该表示用于通过分类器 G_ϕ 预测类别标签。对于视觉 Transformer，采用 [CLS] 标记作为图像分类的表示。设 $N_{t,k}$、$X_{t,k}$ 和 $\mu_{t,k}$ 分别表示任务 t 中类别 k 的样本数量、图像集以及类别中心。而 $\mu_{t,k} = \frac{1}{N_{t,k}} \sum_{i=1}^{N_{t,k}} F_\theta(X_{t,k}^i)$，即对该类别的所有样本取平均值。

为了引入更真实且无遗忘的样本级原型，以减轻分类器的偏差，通过使用类别中心和当前样本来恢复旧任务的原型。将原型重放分为两个步骤：1) 引入监督机制，使这些原型偏移与任务无关；2) 利用这一特性恢复旧的样本级原型。**任务无关原型偏移的监督**。首先，考虑当前任务和上一个任务的模型，即 F_{θ_t} 和 $F_{\theta_{t-1}}$，并应用偏移正则化。令 bs 表示批大小，考虑批次中的训练样本 $(x_i^t, y_i^t), i = 1, 2, ..., bs$，并将它们随机划分为两个大小相同的子集 S_1 和 S_2，每个子集的大小为 $\lfloor \frac{bs}{2} \rfloor$。然后，计算两个子集中样本的原型偏移：对于 S_1，计算 $\{O_{t,i} = F_{\theta_t}(x_i^t) - \mu_{t,y_i^t} | i \in S_1\}$；对于 S_2，计算 $\{O_{t-1,i} = F_{\theta_{t-1}}(x_i^t) - \mu_{t,y_i^t} | i \in S_2\}$。采用旧模型 $F_{\theta_{t-1}}$ 来计算子集 S_2 的原型偏移，这样可以利用旧模型中包含的偏移分布的先前知识。从中随机采样 $\lfloor \frac{bs}{2} \rfloor$ 对原型偏移，并最小化它们之间的均方误差。

$$\mathcal{L}_t^{pr} = \frac{1}{sz} \sum_{(i_k, j_k) \in Idx} \mathcal{L}_{mse}(O_{t,i_k}, O_{t-1,j_k}), \tag{4.3}$$

Algorithm 2 训练过程的伪代码

输入: 任务数量 T，第 t 个任务的训练样本 $D_t = (x_i, y_i)_{i=1}^{N}$，任务 t 中类别 k 的类别原型 $\mu_{t,k}$（在训练过程中维护），初始参数 $\Theta_0 = \{\theta_0, \phi_0\}$ 包含视觉 Transformer 的特征提取器 F_{θ_t} 和分类器 G_{ϕ_t} 的参数。CE 代表交叉熵损失。

输出: 模型 Θ_T

for $t \in \{1, 2, ..., T\}$ **do**
  $\Theta_t \leftarrow \Theta_{t-1}$
  while 未收敛 **do**
    从 D_t 中采样 (x, y)
    $P_{t,i}, P_{t-1,i} \leftarrow F_{\theta_t}(x), F_{\theta_{t-1}}(x)$
    $\mathcal{L}_t^{pks} \leftarrow$ 计算 $(P_{t,i}, P_{t-1,i})$ 依照 Eq. (4.2)
    $O_t, O_{t-1} \leftarrow P_{t,y} - \mu_{t,y}, P_{t-1,y} - \mu_{t,y}$
    $\mathcal{L}_{t,pr} \leftarrow \mathcal{L}_{mse}(O_t, O_{t-1})$ 依照 Eq. (4.3)
    $F_{t_{old}, y_{old}} \leftarrow O_t + \mu_{t_{old}, y_{old}}$ 依照 Eq. (4.4)
    $\mathcal{L}_t^{CIL} \leftarrow \mathcal{L}_t^{CE}\ (G_{\phi_t}(F_{t,y}, F_{t_{old}, y_{old}}), y, y_{old})$
    通过最小化 Eq. (4.6) 中的 $\mathcal{L}_t^{\text{all}}$ 更新 Θ_t
  end while
end for

其中 $sz = \lfloor \frac{bs}{2} \rfloor$，$Idx = (i_1, j_1), \ldots, (i_{sz}, j_{sz})$（$i_k \in S_1$，$j_k \in S_2$），$O_{t,i_k}$ 表示来自 S_1 的第 i_k 个原型偏移，而 O_{t-1,j_k} 具有类似的含义。

旧任务原型的恢复。使用当前样本 x_i^t 的原型偏移量以从旧任务中恢复原型:

$$F_{t_{old}, k_{old}} = \mu_{t_{old}, k_{old}} + (F_{\theta_t}(x_i^t) - \mu_{t, y_i^t}). \tag{4.4}$$

公式 4.4 中的第二项是当前样本的计算偏移。样本 (x_i^t, y_i^t) 在当前批次中随机选择。它可以融入公式 4.5 中如下，

$$\mathcal{L}_t^{\text{CIL}} = \mathcal{L}_t^{CE}(G_{\phi_t}(F_{t,y}, F_{t_{old}, y_{old}}), y, y_{old}), \tag{4.5}$$

其中 $\mathcal{L}_t^{CE}$ 是交叉熵损失。整体算法在 Alg. 2 中进行了说明。

四、 学习目标

整体学习目标结合了分类损失、样本原型一致性损失和分块级知识选择:

$$\mathcal{L}_t^{\text{all}} = \mathcal{L}_t^{\text{CIL}} + \lambda_{pks}\mathcal{L}_t^{\text{pks}} + \lambda_{pr}\mathcal{L}_t^{\text{pr}}. \tag{4.6}$$

第三节 实验结果与分析

数据集。方法在三个数据集上进行实验：CIFAR100、TinyImageNet 和 ImageNet-Subset，这些数据集在先前的研究中广泛使用。对于每个实验，首先从数据集

表 4.1 在不同任务数量下，CIFAR100、TinyImageNet 和 ImageNet-Subset 上的平均准确率和最终准确率。基于重放的方法存储每个先前类别的 20 个样本，用符号 † 标记。最佳总体结果以粗体显示。

数据集			CIFAR100						TinyImageNet						ImageNet-Sub	
设置			5 任务		10 任务		20 任务		5 任务		10 任务		20 任务		10 任务	
	方法	Para.(M)	平均↑	最后↑	平均↑	最后↑	平均↑	最后↑	平均↑	最后↑	平均↑	最后↑	平均↑	最后↑	平均↑	最后↑
E=20	iCaRL-CNN†	11.2	51.07	40.12	48.66	39.65	44.43	35.47	34.64	22.31	31.15	21.10	27.90	20.46	50.53	41.08
	iCaRL-NCM†	11.2	58.56	49.74	54.19	45.13	50.51	40.68	45.86	33.45	43.29	33.75	38.04	28.89	60.79	51.90
	LUCIR†	11.2	63.78	55.06	62.39	50.14	59.07	48.78	49.15	37.09	48.52	36.80	42.83	32.55	66.16	56.21
	EEIL†	11.2	60.37	52.35	56.05	47.67	52.34	41.59	47.12	34.24	45.01	34.26	40.50	30.14	63.34	54.19
	RRR†	11.2	66.43	57.22	65.78	55.74	62.43	51.35	51.20	42.23	49.54	40.12	47.46	35.54	67.05	58.22
E=0	LwF_MC	14.5	45.93	36.17	27.43	50.47	20.07	15.88	29.12	17.12	23.10	12.33	17.43	8.75	31.18	20.01
	EWC	14.5	16.04	9.32	14.70	8.47	14.12	8.23	18.80	12.71	15.77	10.12	12.39	8.42	-	-
	MUC	14.5	49.42	38.45	30.19	19.57	21.27	15.65	32.58	17.98	26.61	14.54	21.95	12.70	35.07	22.65
	IL2A	14.5	63.22	55.13	57.65	45.32	54.90	45.24	48.17	36.14	42.10	35.23	36.79	28.74	-	-
	PASS	14.5	63.47	55.67	61.84	49.03	58.09	48.48	49.55	41.58	47.29	39.28	42.07	32.78	61.80	50.44
	SSRE	19.4	65.88	56.33	65.04	55.01	61.70	50.47	50.39	41.67	48.93	39.89	48.17	39.76	67.69	57.51
	本章方法	9.3	**68.17**	**59.02**	**70.13**	**57.90**	**66.86**	**54.25**	**54.88**	**44.97**	**52.72**	**43.35**	**51.68**	**41.94**	**70.18**	**61.42**

中选择部分类别作为基础任务，然后将剩余的类别均匀分配到每个顺序任务中。这个过程可以表示为 $F + C \times T$，其中 F、C、T 分别表示基础任务中的类别数、每个任务中的类别数以及任务数。对于 CIFAR100 和 ImageNet-Subset，采用三种配置：$50 + 5 \times 10$、$50 + 10 \times 5$、$40 + 20 \times 3$。对于 TinyImageNet，设置为：$100 + 5 \times 20$、$100 + 10 \times 10$ 和 $100 + 20 \times 5$。

对比方法。将本章方法与其他非示例类增量学习方法进行比较：SSRE[109]、PASS[66]、IL2A[110]、EWC[79]、LwF-MC[8] 和 MUC[111]。这里还与几种基于示例的方法进行比较，如 iCaRL（最近均值和 CNN）[8]、EEIL[112] 和 LUCIR[91]。

实现细节。关于视觉 Transformer 的结构，在编码器中使用了 5 个 Transformer 块，在解码器中使用了 1 个，比 Vit-Base 的原始版本要轻量得多。所有 Transformer 块的嵌入维度为 384，具有 12 个自注意力头。为每个任务训练了 400 个轮次。在任务 t 之后，为每个类别保存一个平均原型（类别中心）。在实验中，将 λ_{pks} 和 λ_{pr} 设置为 10。报告 CIL 任务的三种常见指标：在学习最后一个任务后，所有已学习任务的平均和最终 top-1 准确率，以及迄今为止学习的所有类别的平均遗忘率。用 Acc_i 表示在任务 i 之后所有已学习类别的准确率。然后，平均准确率定义为 $Avg_{acc} = \frac{\sum_{i=1}^{T} Acc_i}{T}$，最终准确率为 Acc_T。设 $a_{m,n}$ 表示在学习任务 m 后任务 n 的准确率。任务 i 在学习任务 k 后的遗忘度量 f_k^i 被计算为 $f_k^i = \max_{t \in 1,2,\ldots,k-1}(a_{t,i} - a_{k,i})$。平均遗忘率 F_k 定义为 $F_k = \frac{1}{k-1}\sum_{i=1}^{k-1} f_k^i$。所有实验均运行三次，并报告平均性能。

表 4.2 方法各个组成部分的消融实验。实验在 CIFAR-100 上进行，报告 top-1 准确率（以百分比表示）。用 PKS 和 PR 分别表示分块级知识选择和原型恢复。

方法	PKS	PR	5 任务	10 任务	20 任务
PASS	-	-	55.67	49.03	48.48
基准 (ViT)			56.44	51.90	51.20
	✓		58.27	56.82	52.87
		✓	57.78	54.61	51.51
	✓	✓	59.02	57.90	54.25

表 4.3 与其他方法的平均遗忘率比较。实验在 CIFAR100、TinyImageNet 和 ImageNet-Subset 上进行，任务数量为 5、10 和 20。

数据集	**TinyImageNet**			**ImageNet-Subset**
方法	**5 任务**	**10 任务**	**20 任务**	**10 任务**
LwF_MC	54.26	54.37	63.54	56.07
EWC	67.55	70.23	75.54	71.97
MUC	51.46	50.21	58.00	53.85
IL2A	25.43	28.32	35.46	32.43
PASS	18.04	23.11	30.55	26.73
SSRE	**9.17**	14.06	14.20	23.22
本章方法	11.45	**12.21**	**12.82**	**18.39**

一、 与最先进方法的对比

在表 4.1 中，将本章方法与几种非示例和基于示例的方法进行了比较。在非示例设置中，本章方法在所有三个数据集的不同数据划分设置（5/10/20 任务）下，均优于所有先前的相关方法。

以 20 个任务的结果为例，本章方法在 CIFAR100 的 20 任务设置下（最终平均准确率）超越了最佳非示例方法 SSRE 3.31%。此外，本章方法甚至取得了比使用存储样本来减轻遗忘的所有基于示例的方法更高的准确率。这一现象在分辨率更大的数据集（如 TinyImageNet 和 ImageNet-Subset）中仍然保持不变。对于表 4.3 中报告的平均遗忘率，本章方法优于大多数非基于示例的方法。在 ImageNet-Subset 数据集上，差距（高达 4.17%）更加明显。这从另一个角度证明了方法在增量训练过程中的优越性能。

在图 4.3 中展示了动态准确率曲线，结果显示，方法（红色部分）在所有训练阶段的下降速度更慢。

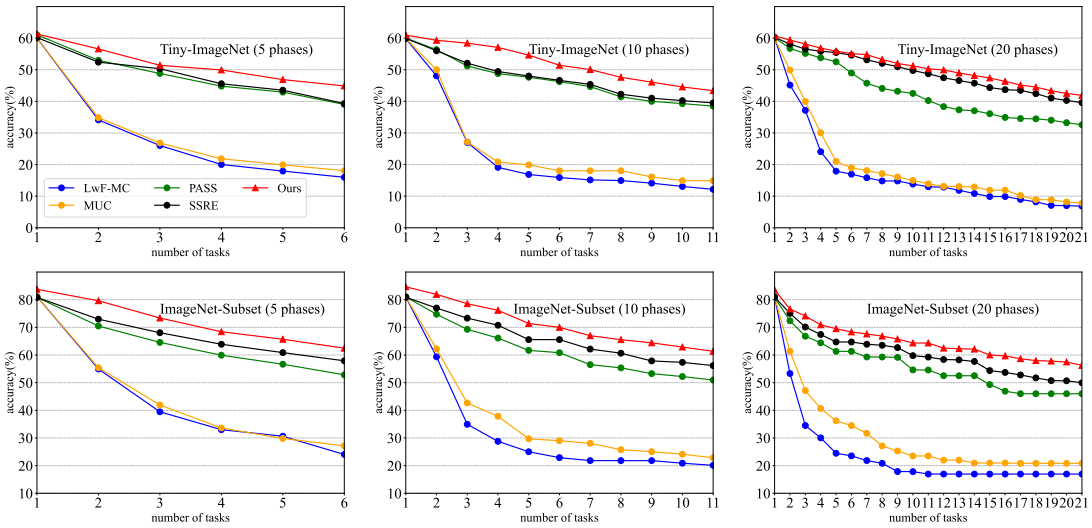

图 4.3 在 TinyImageNet 和 ImageNet-Subset 上不同任务数量的结果表明，本章方法优于其他方法。且在整体训练过程中保持优势。

表 4.4 在 CIFAR-100 上应用 Dytox 的结果，包括平均准确率（%）、最终准确率（%）和遗忘率 F（%），针对 10 任务和 20 任务的场景。

设置	10 任务			20 任务		
	平均↑	最后↑	F↓	平均↑	最后↑	F↓
DyTox	75.47	62.10	15.43	75.10	59.41	21.60
本章方法	78.35	66.47	13.12	77.63	63.90	15.76

二、 消融实验

各组成部分。方法由两个组成部分构成：分块级知识选择和原型恢复。在表 4.2 中分析了每个方面的影响。一种基线是使用基础知识蒸馏和原型增强在 PASS 中进行训练的。同时将视觉 Transformer（ViT）训练视为另一种基线。观察发现：(a) 分块级知识选择显著提高了性能，提升幅度为 3.71%。(b) 原型恢复也带来了一定的提升，提供了更真实的原型重放。(c) 这两个因素可以相互协作，实现更高的性能。这验证了在非示例类增量学习设置中，分块级知识选择和原型恢复两者的重要性。

方法在 PASS 中引入了两个模块来进行该主干网络的实验：原型增强和自监督，如表 4.2 第二行所列。例如，与 PASS 中的原始网络（即 ResNet18）相比，本章的框架在性能上相似或略高于前两行的结果，表明视觉 Transformer 可以作为进一步研究的新基线。这也展示了提出的两个模块对视觉 Transformer 的影响，而非更强的基线。由于 SSRE 中的动态结构重组是针对卷积层设计的，因此没有进行相关实验。

表 4.5 实验在 CIFAR-100 上进行，报告 top-1 准确率（以百分比表示）。第一行将所有 W_i 设置为 1，第二行则使用与分块嵌入和 [CLS] 嵌入的距离成比例的 W_i。

W_i	5 任务	10 任务	20 任务
1	56.18	51.99	50.53
$\|\|P_{t,cls} - P_{t,i}\|\|_2$	53.81	49.78	48.31
Eq. 4.1 (本章方法)	59.02	57.90	54.25

分块级知识选择的进一步研究。由于分块级知识选择是基于视觉 Transformer 的基础知识蒸馏的简单但有效的扩展，故将其应用于 DyTox 和基于示例的任务，以验证其在更多视觉 Transformer 方法和问题设置中的通用性。在表 4.4 的每个实验中，按照 DyTox 的要求存储每个学习类别的 20 个示例，以便进行公平比较。基础知识蒸馏被分块级知识选择所替代。观察到，所提出的方法在平均/最终平均准确率和遗忘率方面显著优于原始 DyTox。这进一步证明了细粒度分块蒸馏方面的有效性。与基础知识蒸馏相比，方法通过对不同分块使用自适应权重来保留知识，为模型在可塑性和稳定性的权衡之间提供了更大的灵活性。

此外，考虑到分块级知识选择方法在蒸馏过程中对每个分块嵌入使用不同的权重，此外还进行了比较实验，以评估该策略在两种不同设置中的有效性。第一个实验将所有蒸馏权重 W_i 设置为 1，第二个实验则通过 $W_i = ||P_{t,cls} - P_{t,i}||_2$ 来计算权重，以替代 Eq. 4.1。

根据表 4.5 中第一个设置的结果，发现对所有分块的蒸馏施加相同的权重导致性能比第三行结果更差。这表明这种严格的限制虽然保留了更多关于先前任务的信息，但却损害了当前任务的学习能力。此外，对距离 [CLS] 嵌入更远的嵌入施加更大的蒸馏权重（与方法相反）导致的结果也比基线更差。这两个设置的结果证明了在与任务相关的分块上保持可塑性以及在与任务无关的分块上保持稳定性的重要性。

分块级知识选择的可视化。这里可视化了一些示例，并展示了在不同任务中分块级知识选择所应用的实际权重，如图 4.4 所示。这些图像选自 ImageNet-Subset 的 10 任务设置。与基础知识蒸馏对每个分块使用相同权重不同，方法可以自适应地选择一些背景补丁以保持稳定性，同时在前景分块上提供更多可塑性，以学习与任务相关的知识。

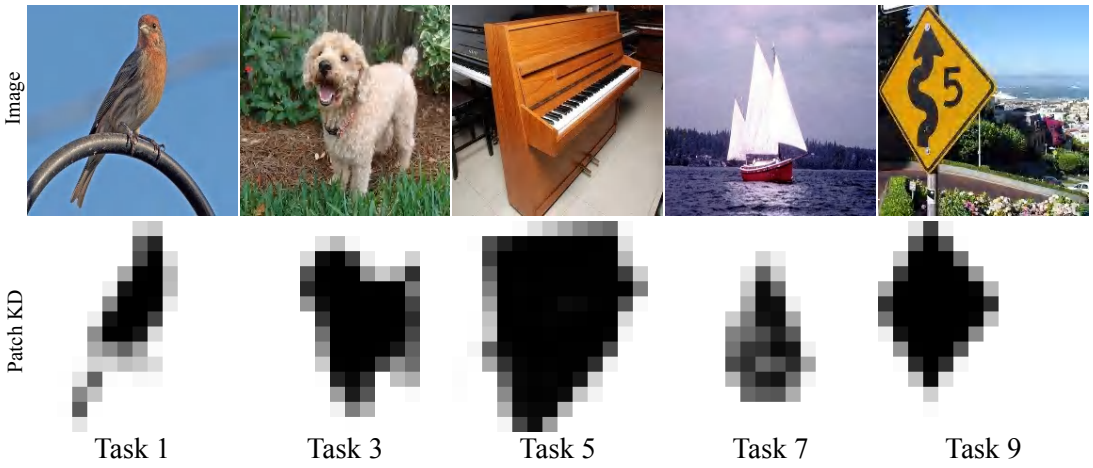

图 4.4 分块级知识选择的可视化及其与基础知识蒸馏的比较。白色分块表示在保持稳定性方面具有更大蒸馏权重。

第四节 本章小结

本章介绍了一种基于视觉 Transformer 的无示例类增量学习框架，旨在减少灾难性遗忘和分类器偏差。该框架通过分块嵌入技术，在不同区域之间实现稳定性与可塑性的平衡，并采用独特策略分别处理与任务相关和与任务无关的区域。此外，为了有效保留旧任务的决策边界而不引入分类器偏差，框架引入了一种新的原型恢复模块。实验结果表明，该方法在增量学习任务中具有良好的性能，为未来研究提供了一种潜在的基线方案。

第五章　基于任务自适应显著性监督的无示例类增量学习

上文分别介绍了两种用于类增量学习的方法：双边 MAE 框架通过利用掩码自编码器有效存储与融合知识，有效缓解了灾难性遗忘问题；基于细粒度知识选择的方法则通过关键知识提取与原型恢复机制，提升了模型对新类别的适应能力，同时保持对旧类别的判别性能。这两种方法在多个基准任务中均取得了优异的性能，验证了其在类增量学习场景下的有效性。

然而，当前方法在处理显著性特征保持与迁移时仍存在一定局限，特别是在跨任务过程中，显著性区域易发生漂移，进而影响知识的有效继承。为此，本章进一步提出了一种基于任务自适应显著性监督的无示例类增量学习方法，旨在系统性缓解显著性漂移与灾难性遗忘问题。该方法综合引入任务自适应显著性监督、边界引导的中层显著性漂移正则化以及辅助低层监督机制，以增强模型对关键区域的识别与保持能力。此外，通过引入显著性噪声注入机制，进一步提升模型在面对跨任务变化时的鲁棒性与泛化能力。

第一节　研究动机与贡献

深度神经网络在许多计算机视觉任务上取得了最先进的性能。然而，大多数这些任务都只考虑一个静态世界，其中任务是明确定义且稳定的，并且所有训练数据都在单个训练会话中可用。真实世界由动态变化的环境和数据分布组成，尤其考虑到训练大型卷积神经网络的计算负担，这些因素重新引发了对增量学习新任务同时避免灾难性遗忘的研究兴趣[73, 113]。

类增量学习是考虑向已训练的模型中添加新类别的可能性的研究场景。大多数类增量学习方法依赖于一个内存缓冲区，用于存储来自过去任务的示例[8, 15, 76, 112]。在本章中考虑的是无示例类增量学习，这是一个更具挑战性的设置，不保留来自先前任务的任何数据。这是一种现实场景，由于隐私担忧或对数据长期存储的限制，这一情况备受关注。然而，无法保留来自过去任务的示例显著加剧了灾难性遗忘的问题。

有几项最近的研究工作考虑了该问题。DeepInversion[114] 反转训练的网络，从随机噪声生成图像作为示例，并将其与当前任务样本混合进行训练。SDC[115]

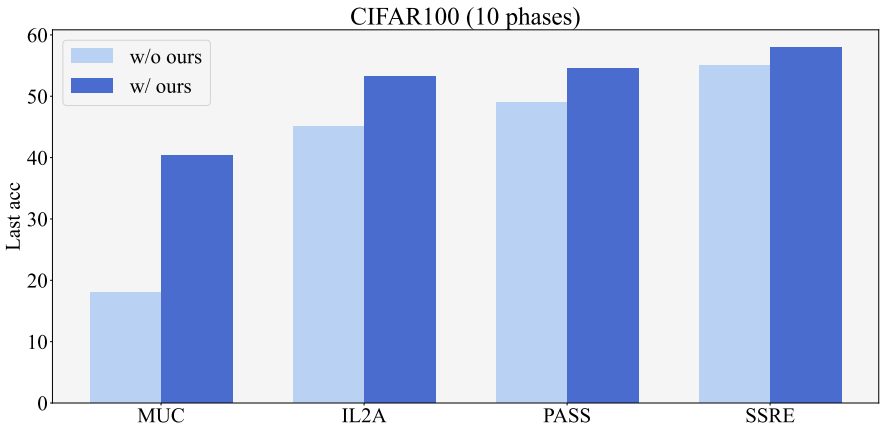

图 5.1 TASS 方法可以直接应用于许多最近的无示例类增量学习方法，从而显著提升分类准确性，并减少灾难性遗忘。

通过假设可以使用新数据近似和估计先前任务中类别的语义漂移，更新每个学习类别的原型。其他先前的研究工作提出了表示学习方法，用于克服灾难性遗忘[66, 110]。如 IL2A[110] 指出，学习更好的表示可以减少在转移到新任务时的表示偏差。加入自监督学习任务，例如 Barlow Twins[61] 和旋转预测[66]，也被提出以实现更稳定的表示并减轻遗忘。卷积神经网络天然地学习关注对其训练的任务具有区分性的特征。在无示例类增量学习中，灾难性遗忘也会发生，因为模型关注的显著特征漂移至新任务特定的特征。在学习新任务时，标准的正则化方法很少能防止这种显著性漂移。一种直接的约束显著性的方法是对样本的显著性图进行蒸馏。然而，在问题设置中，由于无法保存来自先前任务的样本，情况变得更加复杂。另一种方法是在当前任务样本和先前任务注意力之间应用显著性蒸馏[116]。在增强显著性一致性时，这种方法受到当前类别和旧类别之间的语义差距的影响。缺乏显著性约束可能导致注意力在未来任务中向背景漂移，从而导致遗忘。此外，仅仅在注意力上应用蒸馏无法提供可塑性，容易受到注意力遗忘的影响，这是知识遗忘的一个关键因素。相比之下，任务自适应显著性监督方法旨在保持显著性集中于增量学习的任务上，同时保持其可塑性和稳定性。通过监督注意力，它提升了许多先前方法的性能，如图 5.1所示。

具体而言，TASS 整合了三个部分来解决这个问题。首先，使用膨胀的边界图以防止模型中间层跨越物体边界的显著性漂移。由于显著性漂移通常发生在跨任务时，通过膨胀边界监督鼓励模型集中于显著的前景区域，减少了显著性向背景转移的可能性，从而使模型能够自适应地选择与任务相关的前景内的注意力区域。其次，为了同时增强模型跨任务的注意力稳定性，在类增量框架中

添加了一个与核心任务密切相关的与任务无关的低层辅助监督任务，这是因为图像分类已被证实有助于模型定位图像中最显著的区域。最后，提出了一种模块，将显著性噪声注入到特定的特征通道中，并训练网络对其进行去噪，帮助网络进一步抵抗跨任务的注意力漂移。引入了与任务无关的通用属性知识引导，帮助模型学习在不同类别间共享的显著性知识，从而在类增量学习过程中维护一种任务无关的稳定属性。

本工作的主要贡献为：

(i) 为无示例类增量学习设定下的任务自适应显著性监督提供了新的见解；还展示了缺乏或不足的显著性监督方法的负面影响，这说明了提出方法的优越性，并激发了减轻显著性漂移的需求。

(ii) 提出了由三个部分构成的任务自适应显著性监督，这些部分共同用于缓解显著性漂移问题。

(iii) 展示了 TASS 可以轻松集成到其他最先进的方法中，如 MUC[111]，IL2A[110]，PASS[66]，SSRE[109]，从而实现显著的性能提升。

(iv) 实验表明，TASS 在 CIFAR-100、Tiny-ImageNet 和 ImageNet-Subset 的基准测试中优于所有现有的 EFCIL 方法，甚至优于几种基于示例的方法。

第二节 任务自适应显著性监督

首先定义了无示例类增量学习场景。然后描述了提出的 TASS 方法，包括膨胀边界监督、辅助低层监督和显著性噪声注入。整体框架如图 第二节 所示。

一、 无示例类增量学习

类增量学习旨在顺序学习由不相交类别样本组成的任务。记 $t \in 1,2,\ldots,T$ 表示增量学习任务。每个任务的训练数据 D_t 包含 C_t 个类别，其中有 N_t 个训练样本 $(x_t^i, y_t^i)_{i=1}^{N_t}$，其中 x_t^i 是图像，$y_t^i \in C_t$ 是它们的标签。应用于类增量学习的大多数深度网络可以分为两个部分：一个特征提取器 F_θ 和一个通用分类器 G_ϕ，后者随着每个新任务 $t+1$ 的到来而增长，以包含类别 C_{t+1}。特征提取器 F_θ 首先将输入 x 映射到一个深度特征向量 $z = F_\theta(x) \in \mathbb{R}^d$，接着统一的分类器 $G_\phi(z) \in \mathbb{R}^{|C_t|}$ 产生类别 C_t 上的概率分布，用于对输入 x 进行预测。

类增量学习要求模型能够在任何训练任务中正确分类来自先前任务的所有样本——换言之，在学习任务 t 时，模型不能忘记如何对来自任务 $t' < t$ 的类别

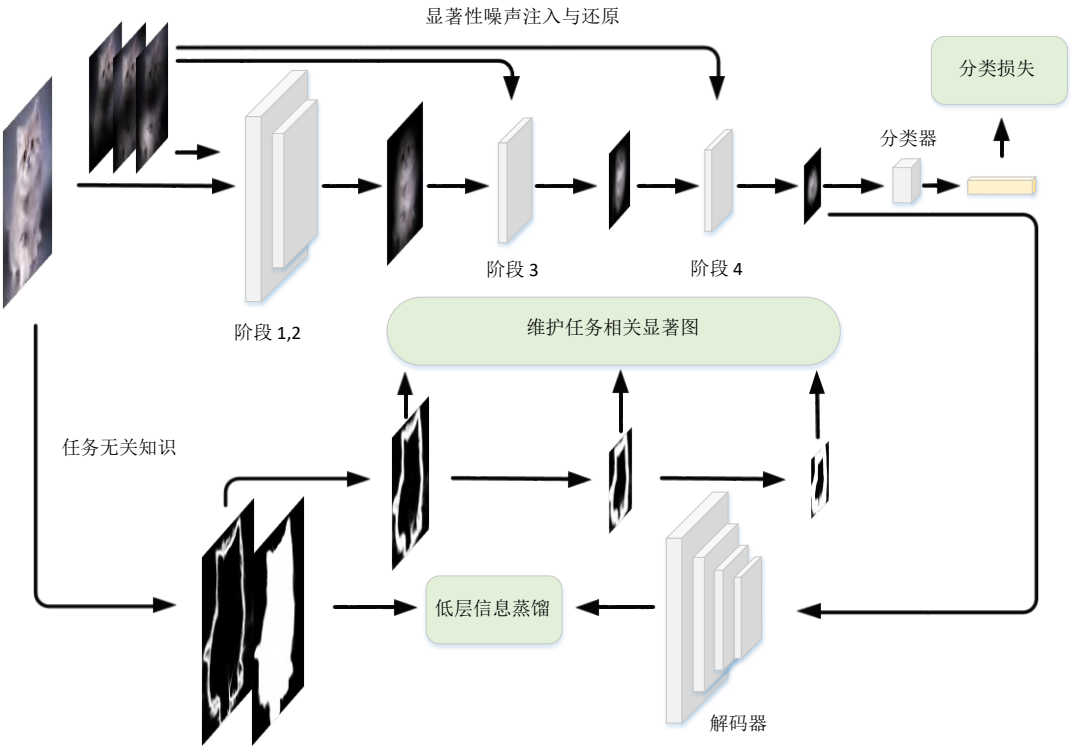

图 5.2 任务自适应显著性监督的整体框架。应用一个低层模型生成显著性图和边界图。边界图通过膨胀和下采样，以在编码器的不同阶段提供监督。在编码器之后附加一个解码器用于低层蒸馏，作为固定的、与任务无关的显著性引导。为了防止在后续训练阶段出现显著性漂移，在每个编码器阶段引入显著性噪声。模型被训练进行去噪，并减少当前数据在未来阶段的显著性漂移。

样本进行分类。无示例类增量学习进一步限制模型在学习每个新任务时不能访问先前任务的样本。一般而言，学习目标为最小化在当前训练数据 D_t 上定义的损失函数 $\mathcal{L}$：

$$\mathcal{L}_t^{\mathrm{CIL}}(x,y) = \mathcal{L}_{\mathrm{ce}}(G_{\phi_t}(F_{\theta_t}(x)),y) + \mathcal{L}_t^{\mathrm{M}}, \tag{5.1}$$

其中 $\mathcal{L}_{\mathrm{ce}}$ 是标准的交叉熵分类损失，$\mathcal{L}_t^{\mathrm{M}}$ 是一种特定于方法的损失，用于在增量学习过程中减轻遗忘。注意，若没有 $\mathcal{L}_t^{\mathrm{M}}$，公式 5.1 就会简化为对任务 t 进行微调。

二、 边界引导的中层显著性漂移正则化

简单地在任务之间蒸馏注意力并未考虑任务自适应的注意力。由一个模型生成每个输入图像 x 的低层表示（在实验中是显著性图和边界图）。使用 CSNet[117] 生成显著区域和边界图，因为它轻量且高效，但框架中可以使用任何生成显著性图的模型。为了在主干网络的这些中间层引入注意力的可塑性，使

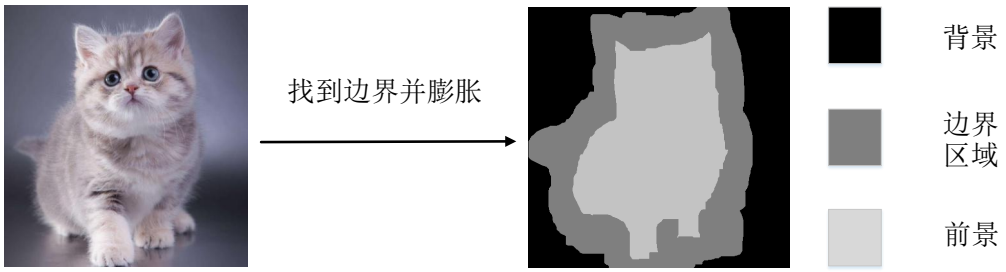

图 5.3 对边界图进行膨胀操作，并在 CNN 主干的三个阶段应用二元交叉熵损失，以防止中层注意力漂移到边界区域。

用生成的边界图作为一种自适应监督，如图 5.3 所示。在物体边界上添加惩罚项，以避免注意力漂移到背景中。首先使用 0.5 作为阈值将生成的边界图二值化，然后通过以下方式膨胀边界图：

$$B_d(x) = \text{Dilate}(A_b(x), d), \tag{5.2}$$

其中 $A_b(x)$ 是图像 x 的生成边界图，它由对显著性图使用拉普拉斯滤波器转换而来，d 表示应用在边界图上的膨胀半径，用于控制边界引导显著性的严格程度。

模型生成中层显著性图时，并非像上述描述的低层显著性图那样在每一层使用解码器，而是在 CNN 主干的三个阶段使用 Grad-CAM[89]（详见图 第二节）。这里还尝试了几种其他用于生成学生显著性图的方法，并在补充材料中对它们进行了报告。为了与 Grad-CAM 生成的显著性边界图进行比较，通过下采样将生成的膨胀边界图 $B_d(x)$ 匹配到这三个阶段的特征图尺寸。在膨胀边界区域上使用二元交叉熵损失进行监督。该损失定义如下：

$$\mathcal{L}_t^{\text{dbs}}(x) = -\frac{\sum_{j=1}^{N} B_d(x,j)\log(1 - S(x,j))}{\sum_{j=1}^{N} B_d(x,j)}, \tag{5.3}$$

其中，$S(x,j)$ 表示模型在图像 x 上的像素 j 处的显著性图，$B_d(x,j)$ 是像素 j 处的膨胀生成的边界图，N 是图像 x 中的像素数。仅在膨胀边界区域内计算此损失（即 $B_d(x,j) = 1$ 的地方）。这种损失有助于使学生显著性图避免与膨胀的教师边界区域相交。

三、 辅助低层监督

在类增量学习期间，提出从所有增量分类任务共享的低层静态任务中学习稳定特征。低层视觉任务，如显著目标检测，需要输入图像的有用表示。通过

跨任务学习这些特征表示，模型可以专注于输入图像的关键区域，并利用学到的稳定特征减少表示漂移，因为低层特征在任务之间变化非常小。

显著性图预测与图像分类相关，因为前景在很大程度上决定了结果，而背景相对不重要。在学习具有新类别的新任务时，新类别图像的背景可能包含引入不必要噪声的新视觉概念，导致遗忘关键的先前知识。显著性引导训练[118]展示了显著性特征对学习分类任务的有效性。对显著区域边界的额外监督可以帮助显著目标检测任务的分割和定位[119–123]。这两种任务之间的积极互动为与主分类任务相关的特征带来更丰富的注意力。它可以以静态知识的形式在类增量任务之间提供积极的指导。一些示例在图 5.5 中有所呈现。

将低层视觉任务作为网络的辅助监督，用于丰富与任务无关的注意力。边界图是通过对估计的显著性图使用拉普拉斯滤波器来计算。方法在主干网络 F_θ 后添加一个解码器 D_ψ[124]，用于预测输入图像的低层显著性图和边界图。将预测值和目标之间的平均 L2 距离作为低层显著性蒸馏损失：

$$\mathcal{L}_t^{\text{lms}}(x) = \frac{||D_\psi(F_\theta(x)) - A(x)||_2}{\sqrt{N}}, \tag{5.4}$$

其中，$A(x)$ 表示输入 x 上的目标低层图，包括显著性图 $A_s(x)$ 和边界图 $A_b(x)$。$D_\psi(F_\theta(x))$ 是解码器生成的组合的显著性图和边界图，N 是显著性图中的像素数。

四、 显著性噪声注入

尽管应用低层蒸馏和膨胀边界监督在任务之间保持显著性表示，但模型仍可能在先前任务的样本中遗忘显著性。为解决这个问题，强制模型能够从注入的显著性噪声中恢复正确的显著性图。

在每个任务中，没有来自先前或未来任务的可用训练数据，因此无法直接知晓这些样本上的显著性漂移。方法没有使用真实显著性漂移信号来监督模型，而是在随机特征通道上引入显著性噪声。使用随机椭圆来近似未来任务中的潜在显著性漂移，并训练模型在每个阶段进行去噪。因此，模型可以学会有效减少实际的显著性漂移。

模型使用非常简单的方法生成椭圆形噪声。有六个参数维度：中心坐标 (x,y)、主轴和次轴长度 (a,b)、旋转角度 α 和掩码权重 w。这个过程的详细解释在补充材料中给出。借助膨胀边界监督，模型的每个阶段学会消除这种额外的显著性噪声，这有助于未来任务的泛化，并减轻先前任务中的显著性遗忘。

Algorithm 3 TASS 伪代码

输入: 任务数目 T，任务 t 的训练样本 $D_t = \{(x_i, y_i)\}_{i=1}^{N}$，初始参数 $\Theta^0 = \{\theta_0, \phi_0, \psi_0\}$ 包含特征提取器 F_θ、分类器 G_ϕ 和低层解码器 D_ψ 参数。
输出: 模型 Θ^T

for $t \in \{1, 2, ..., T\}$ **do**
  $\Theta^t \leftarrow \Theta^{t-1}$
  while 未收敛 **do**
    从 D_t 采样 (x, y)
    $\mathcal{L}_t^{\text{CIL}} \leftarrow$ 显著性噪声注入 (x, y)
    $\mathcal{L}_t^{\text{lms}} \leftarrow$ 低层多任务 $(x, A(x))$
    $S \leftarrow$ 计算 GradCAM 显著性 (x, y)
    $\mathcal{L}_t^{\text{dbs}} \leftarrow$ 膨胀边缘监督 $(S, A(x))$
    通过公式 Eq. 5.5最小化 $\mathcal{L}_t^{\text{all}}$ 更新 Θ^t
  end while
end for

五、 学习目标与训练算法

总体学习目标结合了低层多任务学习、膨胀边界监督和随机显著性噪声注入模块：

$$\mathcal{L}_t^{\text{all}} = \mathcal{L}_t^{\text{CIL}} + \mathcal{L}_t^{\text{lms}} + \mathcal{L}_t^{\text{dbs}}. \tag{5.5}$$

将该损失与公式 5.1 进行比较，对于 TASS，$\mathcal{L}_t^{\text{M}} = \mathcal{L}_t^{\text{CIL}} + \mathcal{L}_t^{\text{lms}}$，因此将显著性感知监督与交叉熵损失结合在一起。整个过程详见算法 3。

第三节 实验结果与分析

在本节中，首先描述了实验设置，然后将 TASS 与几个 EFCIL 基准方法进行比较。之后，对 TASS 的各个部分进行了进一步分析。

一、 实验设置

数据集。遵循三个基准数据集上 EFCIL 的标准实验协议，在 CIFAR-100、Tiny-ImageNet 和 ImageNet-Subset 上进行实验。在大多数实验中，在第一个任务中训练模型学习一半的类别，然后将剩余的类别均匀分配给后续每个任务。使用的约定是：$F + C \times T$ 表示第一个任务包含 F 个类别，接下来的 T 个任务每个包含 C 个类别。这是 EFCIL 中常见的配置，它在 PASS[66] 和 SSRE[109] 中都有使用。

最先进方法。由于本章方法专注于 EFCIL，这里主要与无示例的最先进方法进行比较：SSRE[109]、PASS[66]、IL2A[110]、EWC[79]、LwF-MC[8] 和 MUC[111]。为了

表 5.1 CIFAR-100 上不同任务数量下的平均 top-1 准确率、最终 top-1 准确率以及遗忘情况。基于重放的方法存储了每个先前类别的 20 个样本，用 † 标记。最佳整体结果用**粗体**标出。所有实验均运行三次，并报告所有指标的平均值。

	数据集	CIFAR100									TinyImageNet								
	设置	5 任务			10 任务			20 任务			5 任务			10 任务			20 任务		
	方法	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓
E=20	iCaRL-CNN†	51.07	40.12	42.13	48.66	39.65	45.69	44.43	35.47	43.54	34.64	22.31	36.89	31.15	21.10	36.70	27.90	20.46	45.12
E=20	iCaRL-NCM†	58.56	49.74	24.90	54.19	45.13	28.32	50.51	40.68	35.53	45.86	33.45	27.15	43.29	33.75	28.89	38.04	28.89	37.40
E=20	LUCIR†	63.78	55.06	21.00	62.39	50.14	25.12	59.07	48.78	28.65	49.15	37.09	20.61	48.52	36.80	22.25	42.83	32.55	33.74
E=20	EEIL†	60.37	52.35	23.36	56.05	47.67	26.65	52.34	41.59	32.40	47.12	34.24	25.56	45.01	34.26	25.91	40.50	30.14	35.04
E=20	RRR†	66.43	57.22	18.05	65.78	55.74	18.59	62.43	51.35	18.40	51.20	42.23	16.67	49.54	40.12	21.64	47.46	35.54	29.10
E=0	LwF_MC	45.93	36.17	44.23	27.43	50.47	17.04	20.07	15.88	55.46	29.12	17.12	54.26	23.10	12.33	54.37	17.43	8.75	63.54
E=0	EWC	16.04	9.32	60.17	14.70	8.47	62.53	14.12	8.23	63.89	18.80	12.71	67.55	15.77	10.12	70.23	12.39	8.42	75.54
E=0	MUC	49.42	38.45	40.28	30.19	19.57	47.56	21.27	15.65	52.65	32.58	17.98	51.46	26.61	14.54	50.21	21.95	12.70	58.00
E=0	IL2A	63.22	55.13	23.78	57.65	45.32	30.41	54.90	45.24	30.84	48.17	36.14	25.43	42.10	35.23	28.32	36.79	28.74	35.46
E=0	PASS	63.47	55.67	25.20	61.84	49.03	30.25	58.09	48.48	30.61	49.55	41.58	18.04	47.29	39.28	23.11	42.07	32.78	30.55
E=0	SSRE	65.88	56.33	18.37	65.04	55.01	19.48	61.70	50.47	18.37	50.39	41.67	17.25	48.93	39.89	22.50	48.17	39.76	26.74
E=0	**TASS** (本章方法)	**68.75**	**59.26**	**16.42**	**67.42**	**57.93**	**17.78**	**62.76**	**53.78**	**17.78**	**55.12**	**44.13**	**15.40**	**54.21**	**43.86**	**18.47**	**52.79**	**43.55**	**22.51**

展示方法的有效性，还将其性能与几种基于示例的方法进行比较，如 iCaRL（最近均值和 CNN）[8]、EEIL[112] 和 LUCIR[91]。这里还与集成了 SSRE 的 RRR[125] 进行比较，该方法专注于利用示例重放来保留显著性。

实现细节与性能指标。使用 ResNet-18 作为特征提取的主干网络。与 SSRE[109] 和 PASS[66] 两种最先进的 EFCIL 方法使用相同的基础网络。使用[124] 中的解码器来估计低层学生显著性图。所有实验都是从头开始使用 Adam 进行训练，共进行 100 个轮次，初始学习率为 0.001。学习率在第 45 和第 90 个轮次时按 10 的倍数减少。对于基于示例的方法，使用 herding[8] 来选择和存储每类 20 个样本，遵循常见的设置[8, 91]。将 RRR 与 SSRE 结合实现，以便与 TASS 进行公平比较。对于膨胀边界监督，将三个中层边界膨胀阶段中的 d 设定为图像尺寸的 5%、10% 和 15%。

这里报告了三种常见的类增量学习指标：平均和最终 top-1 准确率，以及截至任务 t 为止学习的所有类别的平均遗忘。记 Acc_i 为截至任务 i 为止学习的所有类别的准确率，平均准确率定义为 $Avg = \frac{\sum_{i=1}^{T} Acc_i}{T}$，最终准确率为 Acc_T。设 $a_{m,n}$ 表示在学习任务 m 后任务 n 的准确率，任务 i 在学习任务 k 后的遗忘度量 f_k^i 计算公式为 $f_k^i = \max_{t \in 1,2,\ldots,k-1}(a_{t,i} - a_{k,i})$。平均遗忘 F_k 定义为 $F_k = \frac{1}{k-1}\sum_{i=1}^{k-1} f_k^i$。

二、 与最先进方法的对比

在 CIFAR-100 上将 TASS 与最先进方法进行了比较，结果见表 5.1。TASS 的表现优于所有无示例方法。对于诸如 iCaRL[8]、EEIL[112] 和 LUCIR[91] 等基于示例的方法，本章方法仍然具有明显更好的性能。在更长的序列（即 10 任务和 20 任务）上，与其他 EFCIL 方法相比，本章方法在学习新类别时明显减少了遗

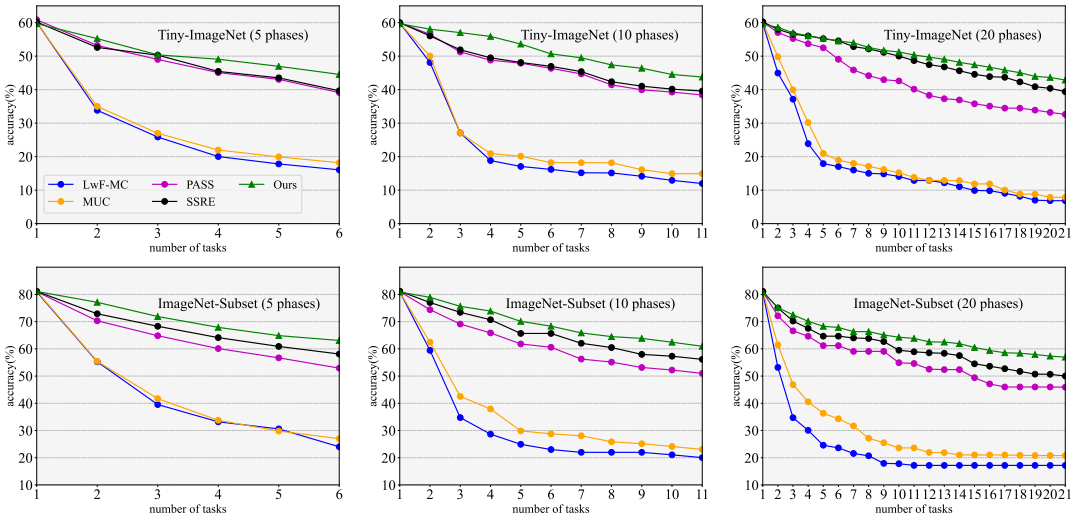

图 5.4 针对不同任务数量在 Tiny-ImageNet 和 ImageNet-Subset 上的结果。本章方法在效果上优于其他方法，特别是在更长的任务序列（即更多但更小的任务）上。

表 5.2 ImageNet-Subset 上不同任务数量下的平均 top-1 准确率、最终 top-1 准确率以及遗忘情况。所有实验均运行三次，并报告所有指标的平均值。

数据集	**ImageNet-Subset**								
设置	**5 任务**			**10 任务**			**20 任务**		
方法	平均↑	最后↑	F↓	平均↑	最后↑	F↓	平均↑	最后↑	F↓
LwF_MC	34.86	24.10	49.36	31.18	20.01	53.04	27.54	17.42	56.07
MUC	40.65	27.89	47.13	35.07	22.65	52.10	31.44	20.12	53.85
PASS	63.12	52.61	22.47	61.80	50.44	23.57	55.23	46.07	26.73
SSRE	69.54	58.46	17.22	67.69	57.51	18.60	61.23	50.05	23.22
TASS (本章方法)	**74.32**	**63.14**	**14.37**	**72.60**	**57.93**	**16.09**	**68.79**	**57.60**	**18.41**

忘。TASS 在最后一个任务上的表现比最佳方法 SSRE 高出约 3%。这种性能改进也可以从平均遗忘的角度观察到。

正如 Tiny-ImageNet 和 ImageNet-Subset 的表 5.2 和图 5.4 中所示，尽管本章方法在图 5.4 中第一个任务的 top-1 准确率类似，但在大多数中间任务和最终任务上表现更好。

在图 5.4 中更长序列的情况下，本章方法与最佳基准方法之间的差距在很大程度上保持一致，表明方法在减轻遗忘方面是有效的。与 CIFAR100 相比，表 5.2 中的性能提升在 Tiny-ImageNet 和 ImageNet-Subset 上更大，这表明本章方法能推广到具有更大图像和物体尺度的数据集。值得一提的是，TASS 也产生具有较小方差的结果。这可能是由于 TASS 减少了对背景区域的显著性漂移，其中可能包括随机噪声。

与其他 EFCIL 方法相结合。一些现有的 EFCIL 方法，如 PASS[66], IL2A[110] 和 SSRE[109]，专注于通过嵌入正则化来减少遗忘。考虑到显著性对图像分类的重

表 5.3 通过以即插即用的方式将 TASS 应用到其他 EFCIL 方法中，top-1 准确率的性能增益。绝对增益以（红色）表示。

数据集	CIFAR-100			Tiny-ImageNet		
方法	5 任务	10 任务	20 任务	5 任务	10 任务	20 任务
MUC	38.45	19.57	15.65	18.95	15.47	9.14
+TASS	49.17 (+10.72)	40.34 (+20.77)	37.86 (+22.21)	32.47 (+13.46)	30.13 (+14.66)	27.70 (+18.56)
IL2A	55.13	45.32	45.24	36.77	34.53	28.68
+TASS	58.74 (+3.61)	53.24 (+7.92)	53.07 (+7.83)	42.49 (+5.72)	41.34 (+6.81)	40.59 (+11.91)
PASS	55.67	49.03	48.48	41.58	39.28	32.78
+TASS	59.10 (+3.43)	54.45 (+5.42)	52.37 (+3.89)	44.05 (+2.47)	43.06 (+3.78)	42.57 (+9.79)
SSRE	56.33	55.01	50.47	41.45	40.07	39.25
+TASS	59.26 (+2.93)	57.93 (+2.92)	53.78 (+3.31)	44.13 (+2.68)	43.86 (+3.79)	43.55 (+4.30)

表 5.4 针对 TASS 各部分的消融。在 CIFAR-100 上进行了 10 任务设置的实验，报告了将 TASS 集成到 PASS 和 SSRE 中的 top-1 准确率（以百分比表示）。$\mathcal{L}_{\text{dbs}}$ (Eq. 5.3)、$\mathcal{L}_{\text{lms}}$ (Eq. 5.4) 和 SNI 分别表示 TASS 的三个组成部分：膨胀边界监督、低层多任务监督和显著性噪声注入。可以看到三部分都对整体结果有贡献。

方法 & 任务	$\mathcal{L}_{\text{lms}}$	$\mathcal{L}_{\text{dbs}}$	SNI	准确率
基准 (PASS)				49.0
变体	✓			51.2
	✓	✓		53.0
	✓	✓	✓	54.5
基准 (SSRE)				55.0
变体	✓			56.2
	✓	✓		57.3
	✓	✓	✓	57.9

要性，自然会考虑是否可以将 TASS 整合到这些方法中。表5.3中的结果显示了这种集成带来的性能增益。在许多情况下，添加 TASS 可使 MUC 的性能翻倍，并显著提升 IL2A 和 PASS。当将其融入最佳基线 SSRE 时，它会产生约 3% 的持续增益。这些结果清楚地表明，通过显式地减轻显著性漂移，TASS 与其他缓解遗忘的方法是互补的。它们还证明了显著性漂移作为 EFCIL 灾难性遗忘的原因的重要性。

三、 其他分析

在本节中更进一步研究提出的方法。若未特别指出，则使用集成到 SSRE[109] 中的 TASS 的结果。

消融实验。使用 CIFAR-100 上的 10 任务设置执行消融（见表5.4）。在 PASS[66]

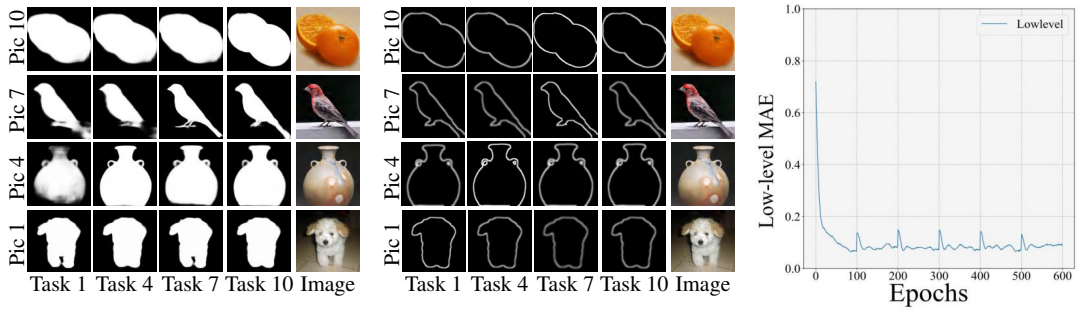

图 5.5 可视化学生编码器-解码器网络的显著性 (a) 和边界 (b) 图，其中包含增量学习不同阶段不同任务的原始图像。方法产生了稳定的低层结果，同时减少了分类中的遗忘。(c) 跨任务的学生和教师网络之间的 MAE 损失。

和 SSRE[109] 上进行消融。低层多任务监督是至关重要的，性能提升分别为 2.2% (PASS) 和 1.2% (SSRE)。膨胀边界监督进一步提升了约 1-2% 的性能。显著性噪声注入对两种方法都有帮助，并使 PASS 提升 1.5%。总体而言，TASS 相比于基线方法分别提升了 5.5% 和 2.9%。注意，SSRE 是先前最先进的方法，而 TASS 在很大程度上优于它。

低层多任务监督。

为分析提出的低层显著性监督的效果，在 5 任务和 10 任务设置下在 ImageNet-Subset 上进行了实验。首先在图5.5 (c) 中绘制了跨任务的损失。

在第一个任务中学习预测边界和显著性图后，该网络在其余的 5 任务序列中保持了良好的性能。这表明在持续学习过程中，低层任务是稳定的。进一步，在图 5.5(a-b) 中可视化了增量学习过程中的显著性图和边界图预测结果。给出了一些在学习不同任务后预测的边界和显著性图的示例。尽管 CIL 涉及不同类别的样本，但低层输出是相对稳定并且与类别无关的。由于模型能够跨任务地稳定预测这些低层特征，它可以为持续学习保留有用的先验知识。

表 5.5 10 任务 CIFAR-100 设置的平均准确率和最终准确率。

指标 & 方法	平均↑	最后↑
FeTrIL[126]	65.20	56.34
SOPE[127]	65.84	56.80
PRAKA[128]	68.86	59.20
TASS (SSRE)	67.42	57.93
TASS (PRAKA)	**69.70**	**60.04**

TASS 与更多的方法和基准。由于方法具有较强的泛化能力，这里还将 TASS 范式应用于 PRAKA[128]，从而进一步提升了性能。在表5.5中给出了具体的实验对

表 5.6 ImageNet-Full 上 10 任务设置的平均准确率和最终准确率。

指标 & 方法	平均↑
SOPE[127]	60.20
FeTrIL[126]	65.00
TASS (FeTrIL)	**66.03**

表 5.7 在 CIFAR-100 10 任务中，DINO 自注意力与学生模型显著性的平均 IoU(%)。

任务	1	4	7	10
SSRE	47.4	50.1	56.6	78.5
SSRE+TASS	75.2	82.3	88.5	90.1

比。还在上表5.5和表5.6中与 FeTrIL[126] 进行了比较。上表5.6中在 ImageNet-Full 上的实验显示了一致的提升。

显著性漂移的定量分析。

衡量了 DINO 中最后一层的自注意力图与 SSRE 和 SSRE+TASS 的 Grad-CAM 显著性图之间的交并比（IoU）。如表5.7所示，TASS 显著降低了学生模型中的显著性漂移，这再次显示了方法在 CIL 期间保持显著性的有效性。

第四节 本章小结

本章提出了一种用于 EFCIL 的任务自适应显著性引导方法（TASS）。该方法的核心思想是通过引导模型关注显著区域并抑制显著性漂移，从而有效减轻跨任务遗忘问题。实验结果表明，TASS 在多个基准数据集上均表现出优异的性能，超越了现有最先进的方法。此外，TASS 具有良好的兼容性，可以与其他方法结合使用，实现更显著的性能提升。

第六章 总结展望

第一节 工作总结

类增量学习是当前深度学习领域的一个重要挑战，模型在不断学习新任务的过程中，往往会遗忘旧任务的知识。针对这一问题，本文围绕知识蒸馏与任务自适应显著性建模，提出了一系列优化策略，以提升增量学习的稳定性与适应性，并在多个数据集上进行了实验验证。

首先，本文提出了一种基于掩码自编码器的高效类增量学习方法，通过随机遮蔽输入图像块进行自监督学习，以减少灾难性遗忘问题。相比于传统基于重放的 CIL 方法，该方法利用 MAE 的图像重建能力，在有限的内存中存储更多有效的示例。此外，为了提升重建图像的质量并增强模型的稳定性，本文设计了一种双边 MAE 结构，在图像级和嵌入级进行信息融合。实验表明，该方法在多个 CIL 基准数据集上取得了最先进的性能，证明了其有效性和优越性。

当设定迁移至不允许保存旧类别数据的无示例类增量任务时，现有的知识蒸馏方法在 EFCIL 任务中难以有效保持旧任务知识，且在新任务学习过程中容易发生显著性漂移。为此，本文提出了一种细粒度知识蒸馏策略，通过任务相关区域的信息增强模型对关键特征的保持能力，使其在增量学习过程中能够更有效地迁移知识。同时，针对模型在多个任务之间进行知识迁移时难以平衡旧知识保持与新知识学习的问题，本文设计了一种显著性增强的蒸馏训练策略，通过动态调整显著性区域，引导模型在学习新任务的同时尽可能保留已有知识。然而，仅依靠知识蒸馏仍然无法完全消除跨任务的显著性漂移，因此，本文进一步引入任务自适应显著性建模，以稳定任务间的注意力分布。提出分块级知识选择方法，以增强模型的稳定性和可塑性。该方法基于视觉 Transformer，通过度量任务相关性来选择关键知识进行蒸馏，并结合改进的原型恢复策略，有效缓解了分类器在增量学习中的偏差和遗忘。

在此基础上，本文提出了一种任务自适应显著性监督策略，以进一步缓解跨任务显著性漂移问题。TASS 通过膨胀边界监督来防止显著性在任务间扩散至不相关区域，利用低层辅助监督任务提高模型对显著区域的识别能力，并引入

显著性噪声去噪模块提升模型的鲁棒性，从而保证模型在增量学习过程中关注关键区域，减少灾难性遗忘。TASS 作为一种显著性建模方法，与细粒度知识蒸馏策略相辅相成，共同提升增量学习任务中的模型稳定性和适应性。

本文在 CIFAR-100、TinyImageNet 和 ImageNet-Subset 等公开数据集上进行了广泛实验，并通过分类精度、灾难性遗忘度量、显著性漂移分析和消融实验等多个方面验证了方法的有效性。实验结果表明，本文提出的方法在多个数据集上均优于现有主流增量学习方法，甚至超过部分基于示例的方法，证明了其在 EFCIL 任务中的潜在应用价值。围绕无示例类增量学习任务，从细粒度知识蒸馏与任务自适应显著性建模两个角度提出了优化方案，有效缓解了灾难性遗忘问题，并提升了模型的稳定性和适应性。

第二节　未来工作展望

如何在增量学习中有效迁移知识，是一个非常重要的研究方向，尤其是在处理无示例类增量学习问题时。虽然本文提出的优化策略已经在多个数据集上取得了不错的效果，但仍有一些方面可以改进和扩展。

当前的细粒度知识蒸馏方法虽然有效，但在任务间显著性区域的动态变化下，模型仍然可能丢失一些重要特征。因此，未来的研究可以探索如何在蒸馏过程中更好地捕捉这些关键特征，进一步提升知识的保持能力。比如，可以考虑结合自监督学习或对比学习的方法，增强模型在增量学习过程中对不同任务的适应能力。

关于任务自适应显著性建模，本文提出的 TASS 的适用性可能受到模型架构的影响。未来，可以尝试将这种方法扩展到其他类型的网络架构。此外，可以结合神经网络的可解释性研究，进一步分析不同任务之间的显著性迁移过程，以优化显著性建模的策略。

总的来说，未来的研究可以围绕知识迁移方法的精细化、显著性建模的普适性、增量学习在其他任务中的应用以及模型的长期稳定性等方面展开，希望本文能够推动这一领域的进一步发展。

参考文献

[1] LOPEZ-PAZ D, RANZATO M. Gradient Episodic Memory for Continual Learning. [C] // Advances in Neural Information Processing Systems: 2017.

[2] CHAUDHRY A, ROHRBACH M, ELHOSEINY M, et al. Efficient Lifelong Learning with A-GEM. [C] // Proceedings of the International Conference on Learning Representations: 2019.

[3] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming Catastrophic Forgetting in Neural Networks. [J]. Proceedings of the National Academy of Sciences, 2017.

[4] WANG Z, CUI Q, SHEN Y, et al. Learning to Prompt for Continual Learning. [J]. ArXiv preprint arXiv:2112.08654, 2022.

[5] WANG Z, SHEN Y, CUI Q, et al. DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning. [J]. ArXiv preprint arXiv:2204.04799, 2022.

[6] 朱飞, 张煦尧, 刘成林. 类别增量学习研究进展和性能评价. [J]. 自动化学报, 2023.

[7] 林钰尧, 杜飞, 杨云. 持续学习: 研究综述. [J]. 云南大学学报 (自然科学版), 2023.

[8] REBUFFI S.-A, KOLESNIKOV A, SPERL G, et al. ICaRL: Incremental Classifier and Representation Learning. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 2017.

[9] LI Z, HOIEM D. Learning Without Forgetting. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[10] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay. [C] // Advances in Neural Information Processing Systems: 2017.

[11] WU Y, CHEN Y, WANG L, et al. Memory replay GANs: Learning to generate images from few examples. [J]. ArXiv preprint arXiv:1804.03461, 2018.

[12] SMITH J, HSU Y.-C, BALLOCH J, et al. Always Be Dreaming: A New Approach for Data-free Class-Incremental Learning. [C] // IEEE/CVF International Conference on Computer Vision: 2021.

[13] GAO Z, ZHU X, ZHOU X, et al. R-EFCIL: Relation-guided exemplar-free class-incremental learning via dual-branch network. [J]. ArXiv preprint arXiv:2207.11283, 2022.

[14] RAMASESH V, YANG M, RUSSELL C, et al. Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting. [J]. ArXiv preprint arXiv:2010.01528, 2020.

[15] DOUILLARD A, CORD M, OLLION C, et al. PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[16] LESORT T, DOUILLARD A, GEPPERTH A, et al. Generative models for incremental learning. [J]. ArXiv preprint arXiv:1906.00689, 2019.

[17] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners. [J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[18] HOI S C, WANG J, ZHAO P, et al. Online Learning: A Comprehensive Survey. [J]. ArXiv preprint arXiv:1802.02871, 2018.

[19] PARISI G I, KEMKER R, PART J L, et al. Continual lifelong learning with neural networks: A review. [J]. Neural Networks, 2019.

[20] LESORT T, CASELLES-DUPRÉ H, GARCIA-ORTIZ M, et al. Continual learning for generative modeling: A review. [J]. Neural Networks, 2020.

[21] MUNDY A, HONG Y, RAWAT Y, et al. What is continual learning? A framework for continual learning research. [J]. ArXiv preprint arXiv:2202.00275, 2022.

[22] DE LANGE M, ALJUNDI R, MASANA M, et al. A continual learning survey: Defying forgetting in classification tasks. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[23] SCHWARZ J, CZARNECKI W M, TOMCZAK J M, et al. Progress & Compress: A Scalable Framework for Continual Learning. [C] // Proceedings of the International Conference on Machine Learning: 2018.

[24] ZENKE F, POOLE B, GANGULI S. Continual Learning Through Synaptic Intelligence. [C] // Proceedings of the International Conference on Machine Learning: 2017.

[25] CHEN R, CHEN G, LIAO X, et al. Class-incremental learning via prototype similarity replay and similarity-adjusted regularization. [J]. Applied Intelligence, 2024.

[26] HUANG W.-C, CHEN C.-F, HSU H. OVOR: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. [J]. ArXiv preprint arXiv:2402.04129, 2024.

[27] HU Y, YANG G, TAN Z, et al. Covariance-based Space Regularization for Few-shot Class Incremental Learning. [J]. ArXiv preprint arXiv:2411.01172, 2024.

[28] FARQUHAR S, GAL Y. Towards robust evaluations of continual learning. [J]. ArXiv preprint arXiv:1805.09733, 2019.

[29] HSU Y.-C, LIU Y.-C, RAMASAMY A, et al. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. [C] // NeurIPS Continual Learning Workshop: 2018.

[30] REBUFFI S.-A, KOLESNIKOV A, SPERL G, et al. ICaRL: Incremental classifier and representation learning. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 2017.

[31] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets. [C] // Advances in Neural Information Processing Systems: 2014.

[32] KINGMA D P, WELLING M. Auto-encoding variational Bayes. [J]. ArXiv preprint arXiv:1312.6114, 2013.

[33] SHIN H, LEE J K, KIM J, et al. Continual learning with deep generative replay. [C] // Advances in Neural Information Processing Systems: 2017.

[34] RAMAPURAM J, GREGOROVA M, KALOUSIS A. Lifelong generative modeling using dynamic expansion graphs. [J]. ArXiv preprint arXiv:1708.00704, 2020.

[35] YANG Y, REN D, PENG C, et al. Dynamic replay training for class-incremental learning. [C] // ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing: 2024.

[36] CHEN Y, TAN A Z, FENG S, et al. General federated class-incremental learning with lightweight generative replay. [J]. IEEE Internet of Things Journal, 2024.

[37] LIM W.-S, ZHOU Y, KIM D.-W, et al. MixER: Mixup-Based Experience Replay for Online Class-Incremental Learning. [J]. IEEE Access, 2024.

[38] RUSU A A, RABINOWITZ N C, DESJARDINS G, et al. Progressive Neural Networks. [C] // ArXiv preprint arXiv:1606.04671: 2016.

[39] HADSELL R, RAO D, RUSU A A, et al. Embracing change: Continual learning in deep neural networks. [J]. Trends in Cognitive Sciences, 2020.

[40] FERNANDO C, BANARSE D, BLUNDELL C, et al. PathNet: Evolution Channels Gradient Descent in Super Neural Networks. [J]. ArXiv preprint arXiv:1701.08734, 2017.

[41] MALLYA A, LAZEBNIK S. PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 2018.

[42] WEN H, PAN L, DAI Y, et al. Class incremental learning with multi-teacher distillation. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2024.

[43] HU Z, LI Y, LYU J, et al. Dense network expansion for class incremental learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2023.

[44] LIANG Y.-S, LI W.-J. Adaptive plasticity improvement for continual learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2023.

[45] GONG Y, YANG Z, LIU S, et al. Continual learning for text-independent speaker verification. [C] // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing: 2018.

[46] SUN C, GONG Y, WU J, et al. Lifelong learning for text classification with dynamic memory networks. [J]. ArXiv preprint arXiv:1906.06556, 2019.

[47] 殷瑞刚, 魏帅, 李晗, 等. 深度学习中的无监督学习方法综述. [J]. 计算机系统应用, 2016.

[48] JING L, TIAN Y. Self-supervised spatiotemporal feature learning via video clip order prediction. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[49] LIU X, HE K. Self-supervised learning: Generative or contrastive. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[50] DOERSCH C, GUPTA A, EFROS A A. Unsupervised visual representation learning by context prediction. [C] // Proceedings of the IEEE International Conference on Computer Vision: 2015.

[51] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles. [C] // European Conference on Computer Vision: 2016.

[52] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised Representation Learning by Predicting Image Rotations. [C] // Proceedings of the International Conference on Learning Representations: 2018.

[53] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization. [C] // European conference on computer vision: 2016.

[54] PATHAK D, KRAHENBUHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition: 2016.

[55] MISRA I, MAATEN L V D. Self-supervised learning of pretext-invariant representations. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[56] TIAN Y, KRISHNAN D, ISOLA P. What makes for good views for contrastive learning. [C] // Advances in Neural Information Processing Systems: 2020.

[57] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations. [C] // International conference on machine learning: 2020.

[58] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[59] GRILL J.-B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent: A new approach to self-supervised learning. [C] // Advances in neural information processing systems: 2020.

[60] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments. [C] // Advances in Neural Information Processing Systems: 2020.

[61] ZBONTAR J, JING L, MISRA I, et al. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. [C] // Proceedings of the International Conference on Machine Learning: 2021.

[62] BARDES A, PONCE J, LECUN Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. [C] // International Conference on Learning Representations: 2022.

[63] XIE Z, ZHANG Z, CAO Y, et al. SimMIM: A Simple Framework for Masked Image Modeling. [J]. ArXiv preprint arXiv:2111.09886, 2022.

[64] WEI C, FAN H, XIE S, et al. Masked feature prediction for self-supervised learning with transformers. [J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[65] ASSRAN M, CARON M, MISRA I, et al. Self-supervised learning from images with a joint-embedding predictive architecture. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[66] ZHU F, ZHANG X.-Y, WANG C, et al. Prototype Augmentation and Self-Supervision for Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[67] PHAM T, LIU J, YU X, et al. Dualnet: Continual learning, fast and slow. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[68] WANG X, ZHANG Y, HONG Y, et al. Meta-CL: Learning to Learn Continual Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[69] LASKIN M, SRINIVAS A, ABBEEL P. Curl: Contrastive unsupervised representations for reinforcement learning. [C] // International Conference on Machine Learning: 2020.

[70] CHEN X, HE K. Exploring simple siamese representation learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[71] CHA S, LEE J, SHIN J, et al. Co2l: Contrastive continual learning. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision: 2021.

[72] HINTON G, VINYALS O, DEAN J. Distilling the Knowledge in a Neural Network. [C] // Neural Information Processing Systems Deep Learning and Representation Learning Workshop: 2015.

[73] MCCLOSKEY M, COHEN N J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. [J]. Psychology of Learning and Motivation, 1989.

[74] FURLANELLO T, LIPTON Z C, TSCHANNEN M, et al. Born-Again Neural Networks. [C] // Proceedings of the 35th International Conference on Machine Learning: 2018.

[75] JUNG H, JU J, JUNG M, et al. Less-Forgetting Learning in Deep Neural Networks. [C] // Proceedings of the 31st AAAI Conference on Artificial Intelligence: 2016.

[76] WU Y, CHEN Y, WANG L, et al. Large Scale Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2019.

[77] AHN H, CHA S, MOON T. Knowledge Retention in Continual Learning by Knowledge Adaption and Augmentation. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision: 2021.

[78] PARK W, KIM D, LU Y, et al. Relational Knowledge Distillation. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2019.

[79] KIRKPATRICK J, PASCANU R, RABINOWITZ N, et al. Overcoming catastrophic forgetting in neural networks. [J]. Proceedings of the National Academy of Sciences, 2017.

[80] ALJUNDI R, LIN M, GOUJAUD B, et al. Memory aware synapses: Learning what (not) to forget. [C] // Proceedings of the European Conference on Computer Vision: 2018.

[81] RIEMER M, CASES I, AJEMIAN R, et al. Learning to learn without forgetting by maximizing transfer and minimizing interference. [C] // International Conference on Learning Representations: 2018.

[82] LIU Y, SCHIELE B, SUN Q. Adaptive aggregation networks for class-incremental learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[83] LI Z, ZHOU F, WU F, et al. Incremental learning with attention attractor networks. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision: 2019.

[84] CHAUDHRY A, DOKANIA P K, AJANTHAN T, et al. Riemannian walk for incremental learning: Understanding forgetting and intransigence. [C] // Proceedings of the European Conference on Computer Vision: 2018.

[85] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998.

[86] BORJI A. Salient object detection: A survey. [J]. Computational Visual Media, 2019.

[87] 孙涵, 刘译善, 林昱涵. 基于深度学习的显著性目标检测综述. [J]. Journal of Data Acquisition & Processing/Shu Ju Cai Ji Yu Chu Li, 2023.

[88] 袁野, 和晓歌, 朱定坤, 等. 视觉图像显著性检测综述. [J]. 计算机科学, 2020.

[89] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. [C] // Proceedings of the IEEE International Conference on Computer Vision: 2017.

[90] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. [C] // International Conference on Learning Representations: 2017.

[91] HOU S, PAN X, CHANGE LOY C, et al. Learning a unified classifier incrementally via rebalancing. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2019.

[92] LIU Y, SU Y, LIU A.-A, et al. Mnemonics Training: Multi-Class Incremental Learning Without Forgetting. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[93] ZHAO B, LIU Y, LI Z, et al. Maintaining discrimination and fairness in class incremental learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[94] JIN X, PENG B, WU Y, et al. Knowledge distillation via route constrained optimization. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[95] SHI X, YIN H, LIU Y, et al. Overcoming catastrophic forgetting via model adaptation. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[96] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images. [J]. 2009.

[97] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database. [J]. 2009.

[98] KINGMA D P, BA J. Adam: A method for stochastic optimization. [J]. International Conference on Learning Representations, 2015.

[99] YAN S, XIE J, HE X, et al. Dynamically Expandable Representation for Class Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2021.

[100] DOUILLARD A, RAMÉ A, COUAIRON G, et al. Dytox: Transformers for continual learning with dynamic token expansion. [C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 2022.

[101] LIU Y, SCHIELE B, SUN Q. Generative Feature Replay for Class-Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops: 2020.

[102] ZHAO B, LIAO Z, CHENG X, et al. Memory Efficient Class-Incremental Learning for Image Classification. [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

[103] WANG J, RAMANAN D. Memory Replay with Data Compression for Continual Learning. [C] // International Conference on Learning Representations: 2022.

[104] LUO X, HONG L, WANG X, et al. Class-Incremental Learning with Dual Memory. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2023.

[105] ROTH K, MILBICH T, GÖRNITZ N, et al. Revisiting training strategies and generalization performance in deep metric learning. [C] // Proceedings of the 37th International Conference on Machine Learning: 2020.

[106] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks. [C] // Advances in Neural Information Processing Systems: 2012.

[107] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 2016.

[108] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. [C] // International Conference on Learning Representations: 2021.

[109] ZHU K, ZHAI W, CAO Y, et al. Self-Sustaining Representation Expansion for Non-Exemplar Class-Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2022.

[110] ZHU F, CHENG Z, ZHANG X.-Y, et al. Class-Incremental Learning via Dual Augmentation. [J]. Advances in Neural Information Processing Systems, 2021.

[111] LIU Y, SCHIELE B, SUN Q. More classifiers, less forgetting: A generic multi-classifier paradigm for incremental learning. [C] // European Conference on Computer Vision: 2020.

[112] CASTRO F M, MARÍN-JIMÉNEZ M J, GUIL N, et al. End-to-end incremental learning. [C] // Proceedings of the European conference on computer vision: 2018.

[113] GOODFELLOW I J, MIRZA M, XIAO D, et al. An empirical investigation of catastrophic forgetting in gradient-based neural networks. [C] // International Conference on Learning Representations: 2013.

[114] YIN H, MOLCHANOV P, ALVAREZ J M, et al. Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[115] YU L, TWARDOWSKI B, LIU X, et al. Semantic Drift Compensation for Class-Incremental Learning. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020.

[116] DHAR P, SINGH R V, PENG K.-C, et al. Learning Without Memorizing. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2019.

[117] CHENG M.-M, GAO S, BORJI A, et al. A Highly Efficient Model to Study the Semantics of Salient Object Detection. [J]. IEEE TPAMI, 2021.

[118] ISMAIL A M, ZHOU Y, JUAN A, et al. Improving robustness against common corruptions with saliency-guided training. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision: 2021.

[119] FAN D.-P, JI G.-P, ZHOU T, et al. Advances in Salient Object Detection: A Survey. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[120] LI H, ZHOU T, LIU J, et al. Boosting Salient Object Detection with Anchor-based Diffusion. [J]. IEEE Transactions on Image Processing, 2023.

[121] LIN Z, WU J, ZHANG Z, et al. Learning Hierarchical Representations for Salient Object Detection. [J]. Pattern Recognition, 2023.

[122] MA R, SUN Q, ZHANG J, et al. Self-supervised Learning for Salient Object Detection. [J]. IEEE Transactions on Multimedia, 2024.

[123] ZHAO J.-X, LIU J.-J, FAN D.-P, et al. EGNet: Edge Guidance Network for Salient Object Detection. [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision: 2019.

[124] LI X, YOU A, ZHU Z, et al. Semantic Flow for Fast and Accurate Scene Parsing. [C] // ECCV: 2020.

[125] EBRAHIMI S, PETRYK S, GOKUL A, et al. Remembering for the right reasons: Explanations reduce catastrophic forgetting. [J]. Applied AI letters, 2021.

[126] PETIT G, POPESCU A, SCHINDLER H, et al. Fetril: Feature translation for exemplar-free class-incremental learning. [C] // WACV: 2023.

[127] ZHU K, ZHENG K, FENG R, et al. Self-Organizing Pathway Expansion for Non-Exemplar Class-Incremental Learning. [C] // ICCV: 2023.

[128] SHI W, YE M. Prototype Reminiscence and Augmented Asymmetric Knowledge Aggregation for Non-Exemplar Class-Incremental Learning. [C] // ICCV: 2023.

致谢

白驹过隙，转眼间我的硕士生涯即将落下帷幕。回首这段求学之路，内心充满感慨与感激。在此，我要由衷地感谢一直以来给予我支持与鼓励的家人，感谢在学术和生活中陪伴我成长的导师、朋友和同学们。

在这几年的求学过程中，我要特别感谢程明明老师以及刘夏雷老师的悉心指导与无私帮助。他们不仅在学术上给予了我极大的启发，耐心解答我的疑问，提供宝贵的建议，还在科研道路上给予了我坚定的信心和前行的动力。此外，感谢导师们为我提供了良好的研究环境和计算资源，让我能够专注于学术探索，不断提升自己的能力。

同样，我也要感谢实验室的同门伙伴们，我们在学术研究中相互交流、共同进步，在遇到科研瓶颈时彼此鼓励、携手攻克难题。这段时光不仅让我收获了知识，更让我结识了一群志同道合的朋友。我们在科研上并肩前行，这些将成为我人生中最宝贵的回忆。

此外，感谢所有曾经给予我帮助的老师和同学，正是你们的支持与鼓励，让我能够坚持不懈地追求自己的目标。硕士阶段的学习不仅让我掌握了科研方法，更让我学会了如何独立思考、如何面对挑战。

未来的道路仍然漫长，我将铭记这段求学岁月的点点滴滴，不断前行，迎接新的挑战，创造更加美好的未来！

个人简历

翟江天，出生于 2001 年 1 月 2 日。在 2022 年毕业于南开大学计算机科学与技术专业并获得学士学位。于 2022 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:

- **Jiang-Tian Zhai**, Xialei Liu, Andrew D. Bagdanov, Ke Li, Ming-Ming Cheng. Masked autoencoders are efficient class incremental learners [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.
- **Jiang-Tian Zhai**, Qi Zhang, Tong Wu, Xing-Yu Chen, Jiang-Jiang Liu, Ming-Ming Cheng. SLAN: Self-Locator Aided Network for Vision-Language Understanding [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.
- **Jiang-Tian Zhai**, Xialei Liu, Lu Yu, Ming-Ming Cheng. Fine-grained knowledge selection and restoration for non-exemplar class incremental learning [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2024.
- Xialei Liu*, **Jiang-Tian Zhai***, Andrew D. Bagdanov, Ke Li, Ming-Ming Cheng. Task-adaptive saliency guidance for exemplar-free class incremental learning [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.