基于多元关系建模的少样本分类算法研究

重庆大学硕士学位论文

（学术学位）

学生姓名：XXX

指导教师：XXX 教授

学科门类：工 学

学科名称：软件工程

研究方向：计算机视觉

答辩委员会主席：X X 教授

授位时间：2024 年 6 月

摘 要

待补全

关键词：少样本分类；关系建模；对比学习；语义信息表示

Key words:

目 录

图目录

表目录

1 问题重述

本章内容共分为四节，第一节介绍本文的研究背景及意义；第二节总结少样本分类算法的国内外研究现状，并对其面临的挑战进行分析；第三节介绍本文的研究内容与创新点；第四节对本文组织结构进行概括。

1.1 问题一

在当今时代，深度学习技术已在诸如图像分类、目标检测、实例分割等人工智能领域中取得显著成就[1–8]，在某些特定任务中的表现达到甚至超越了人类的水平。然而，这些技术的成功在很大程度上依赖于大规模标注数据集的支撑。一旦没有足够数量的标注样本，很多深度学习模型便会因为只在少量样本数据上进行训练而出现过拟合或欠拟合现象，进而导致无法达到良好的性能表现。

与深度学习模型不同，人类在成长过程中学习积累了大量知识后，面对新物体或新场景时能够总结以往的知识与经验，通过少量样本便可迅速准确地识别新类别。例如，某人已经认识了“猫”、“狗”、“马”等动物，而从未见过“水豚”这类动物，但通过观察几张甚至一张“水豚”的图片，便可对其准确识别，而深度学习模型则可能需要使用数百乃至上千张图片进行训练才能达到相同的识别效果。为了模拟人类认识新类别的过程，少样本分类（Few-Shot Classification, 简称 FSC）应运而生。少样本分类致力于模拟人类的知识迁移能力，期望模型在具有大量标注数据的基类数据上训练之后，能够将所学知识迁移至新类别上，实现用少量标注样本进行有效学习。

作为目前计算机视觉领域的热门研究方向之一，无论是在学术探索还是实际应用方面，少样本分类都具有深远意义。首先，在学术探索方面，少样本分类打破了传统深度学习依赖大规模标注数据集进行训练的范式，推动了包括元学习、迁移学习、模型正则化等在内的一系列理论和方法的发展，为解决深度学习任务中的数据稀缺问题提供了新的视角和方法论。并且少样本分类强调模型的泛化能力，为提高模型在仅有少量标注数据类别上的泛化性能，多种理论和算法被提出，这同样可被其他学习任务借鉴使用。另外，在现实场景中，诸多任务无法获取大量标注数据，少样本分类为这些任务提供了理论基础和技术支持。例如，医学图像分析、疾病诊断领域，标注数据获取困难且成本高昂。少样本分类技术可以利用有限的病例进行高效学习，辅助医生进行更准确的诊断。在生态研究和动物识别等领域，很多物种稀有导致难以收集大量样本，少样本分类可以帮助识别这类物种。

少样本分类需要在大量标注数据的基类上训练之后，在仅有少量标注数据的新类上执行分类任务，因此如何迁移学习到的知识成为了关键。虽然基类与新类的样本类别不同，但其数据间却共享着一些深层次、多样化的关系，这些关系表

现为样本间的相似性、差异性，以及语义空间与视觉空间的联系性等。通过在基类数据上对这些关系进行建模，可以更好地理解与挖掘数据间的内在联系，从而迁移在基类上学习到的知识，提升模型在新类数据上的表现。本文旨在研究少样本数据集中的多元关系，通过建模多粒度样本关系，提升视觉特征提取网络的特征提取能力和视觉特征的判别性；通过引入语义信息，建模语义-视觉多空间关系，使得模型可以获取标注样本的多种模态信息，提高模型的泛化能力。本文充分挖掘并利用样本数据的多元关系，为少样本分类问题提供了新的研究视角，有望为少样本分类领域的学术研究与实际应用进程起到一定程度的促进作用。

1.2 问题二

1.2.1 研究现状

近年来，已有很多少样本分类方法被提出，按其技术方案可以大致分为五类，分别是：基于元学习的少样本算法、基于度量的少样本算法、基于数据增强的少样本算法、基于特征学习的少样本算法和基于语义的少样本算法，以下将分别对其进行介绍。

（1）基于元学习的少样本算法

基于元学习的少样本分类算法[9–12]，其核心思想是在训练阶段便模拟少样本测试任务，在从基类数据集采样的大量少样本分类任务中学习到元知识，元知识可以迁移到其他少样本任务，从而使模型在遇到新任务时能够通过极少量的样本训练便快速调整参数并达到较好的分类性能。例如，Finn 等人[9] 提出了模型无关的元学习算法（Model-Agnostic Meta-Learning，简称 MAML）。MAML 设计了一种优化算法，通过找到一组初始化模型参数，使用少量梯度下降便能够使其适应新的任务。Lee 等人[10] 则是使用支持向量机（Support Vector Machine，简称 SVM）代替 MAML 方法中的线性分类器，并结合了一个可微分二次规划求解器使得其能够端到端学习。Rusu 等人[11] 提出了一种在低维潜在空间进行模型元学习的方法 LEO，其将元学习问题转化为潜在空间中的优化问题，利用潜在空间的特征嵌入捕捉少样本任务间共享的结构性知识，促进不同任务间的知识转移。基于元学习的方法虽很符合少样本分类的特点，但其通常需要先对特征提取网络进行预训练，并在元学习阶段采样大量任务来微调网络，存在训练过程较为复杂的问题。

（2）基于度量的少样本算法

基于度量的方法[13–16] 为少样本分类问题提供了另一种解决思路，其旨在通过学习样本之间的距离或相似度度量来处理少样本问题。这类方法的核心思想是，如果能够合理地度量样本之间的距离或相似性，即便是只有少量的训练样本，也可以通过比较未知样本与已知样本之间的距离或相似度来进行有效的分类。基于度量的方法大多使用欧式距离、余弦相似度计算样本之间的距离，例如 Snell 等人[13] 提出的原型网络（Prototypical Networks，简称 ProtoNet）。ProtoNet 基于以下假设：在

特征空间中，每个类别都可以由其样本特征的平均值代表的一个原型来表示。在进行分类时，其会计算查询样本与每个类别原型之间的欧式距离，并将查询样本分类到最近的原型所代表的类别。Zhang 等人[15] 提出的 DeepEMD 方法为少样本分类引入了一种新的距离度量方式：推土机距离（Earth Mover's Distance，简称 EMD）。DeepEMD 将一张图像分为不同的图像块，对其进行特征提取并利用推土机距离作为度量标准来比较不同图像之间的相似度。另外，Sung 等人[17] 提出的关系网络（Relation Networks，简称 RelationNet）则是通过学习一个深度度量来评估样本之间的关系得分，进而通过关系得分进行分类。与之前方法不同，此关系得分是通过网络学习到的，而不是设计的固定距离度量方式。基于度量的少样本算法简单高效，其难点主要在于如何建立一个合适的度量方式来衡量样本之间的距离或相似度。

（3）基于数据增强的少样本算法

增加样本数量来应对标注样本不足的问题，是少样本分类最直观的解决方案，因此，基于数据增强的少样本算法被提出[18,19]。少样本分类中，每个类别样本数目极少，模型很容易产生过拟合问题，该类算法通过增加训练样本的数量和多样性来帮助模型学习到更加鲁棒的特征表示，从而减少过拟合的风险。例如，Chen 等人[18] 提出了一种名为图像变形元网络（Image Deformation Meta-Networks，简称 IDeMe-Net）的新颖框架。IDeMe-Net 训练一个网络，该网络能够通过线性地融合一组图像生成变形图像，从而产生额外的标注样本，增加模型的训练样本。Li 等人[19] 提出的对抗性特征幻觉网络（Adversarial Feature Hallucination Networks，简称 AFHN）则是在特征层次对样本数量进行增加。AFHN 方法利用生成对抗网络（Generative Adversarial Nets，简称 GAN）[20] 来生成新的样本特征，从而解决训练样本特征稀缺的问题。另外，还有部分方法将语义信息作为条件并使用生成模型合成额外的训练样本或特征，由于此类方法使用到了语义信息，因此将其划分为基于语义的少样本算法，将在后续进行介绍。基于数据增强的方法更符合解决少样本分类问题的直觉，但其需要增加很多样本以缓解过拟合问题，并且如何确保所增加样本的多样性也是一大挑战。

（4）基于特征学习的少样本算法

近年来，少样本分类的特征学习阶段越来越受到重视，并出现了一系列基于特征学习的少样本算法[21–30]。这些方法直接使用整个基础数据集以和普通分类任务一致的方式来训练模型，直接执行分类任务或者增加额外的辅助任务以获得特征提取能力出色的特征提取网络。Tian 等人[21] 总结了基于元学习以及度量学习方法的不足，并开创性地提出 RFS 方法。RFS 在整个基类数据集上执行分类任务训练网络来学习良好的特征嵌入，在测试阶段，RFS 冻结特征提取网络参数并使用其提取图像特征，并随后添加逻辑回归分类器进行少样本分类任务。通过此种简单的方式便可得到一个优质的特征提取网络，并能够达到良好的少样本分类结果。在此基础上，一些其他人的工作[22,23] 进一步证明了此类方法的有效性。另外，还有一

些工作在分类任务的基础上添加额外的辅助任务进一步提升特征提取网络的泛化性。例如，Zhang 等人[24] 提出使用方向梯度直方图（Histogram of Oriented Gradient，简称 HOG）和局部二值模式 (Local Binary Patter，简称 LBP) 来提取手工特征并用来指导特征提取网络的优化，缓解了模型的过拟合问题。其他一些工作[25–30] 则是利用自监督或者对比学习任务作为辅助任务来提升模型的特征提取能力以及泛化能力，从而达到良好的少样本分类表现。相比于基于元学习、度量和数据增强的方法，基于特征学习的方法对少样本分类提供了一种更为简单的解决方案，但目前部分方法仅使用分类损失训练网络或者直接使用一些对比学习的方法辅助训练，没有对样本关系进行充分挖掘，限制了模型性能。

（5）基于语义的少样本算法

受到人类认知新类别时语义信息可以提供帮助作用的启发，研究者开始将语义信息引入到少样本分类算法中。基于语义的少样本算法通常使用 Word2Vec[31]、GloVe[32]、BERT[33] 等自然语言模型或者 CLIP[34] 等多模态模型的文本编码器来将类别名称转化为语义特征，并使用其对视觉特征进行补充以使得模型能够获取样本的多种模态信息，丰富了样本特征所包含的信息，进而可以提高少样本分类准确率。根据其利用语义信息的方式不同，本文将其大致分为两类，分别是基于特征生成的方法和基于语义修正的方法。

基于特征生成的方法：此类方法将语义信息作为生成模型的条件生成额外的样本以提升样本多样性，从而缓解分类器仅用少量样本训练容易出现过拟合的问题。例如，Chen 等人[35] 将编码器提取的多层视觉特征映射到语义空间，在语义空间使用语义信息对映射后的视觉特征进行特征增强后再用一个解码器将其映射回视觉空间并得到增强后的特征，使用增强后的特征与原始特征共同训练分类器从而达到特征增强的目的。Zhang 等人[36] 提出的 STVAE 模型则是使用不同维度的先验知识（包括视觉先验和语义先验）分别作为变分自编码器（Variational Auto Encoder，简称 VAE）的生成条件生成特征并对其进行融合得到最终的生成特征，将生成的特征作为额外的训练样本以增加样本多样性。

基于语义修正的方法：此类方法的核心思想在于通过约束或者融合的方式利用语义信息对视觉特征进行修正，以优化模型对样本的理解和分类能力，提升模型的泛化能力。例如，Xing 等人[37] 设计了一种自适应融合机制，该机制能够根据需要学习的新图像的类别自适应地融合视觉信息与语义信息，从而捕获视觉、语义两种模态空间的互补信息，增强模型在新类别上的识别能力。另外，Chen 等人[38] 则是将语义信息作为额外输入，与样本图像一同输入模型，并设计了空间维度以及通道维度两种互补机制，以利用语义特征作为提示自适应地调整视觉特征提取网络以及对视觉特征进行补充，从而获得更全面的样本特征，提升模型的少样本分类准确率。

总的来说，基于语义的少样本算法引入了语义信息，能够对视觉信息进行补

充，丰富了模型获取信息的来源，但如何更简单有效地利用语义信息需要进一步探讨。

1.2.2 研究挑战

通过对国内外研究现状进行分析，本文认为当前少样本分类问题还存在着以下挑战：

（1）少样本分类中，训练一个强大的特征提取网络十分重要，它决定了特征的判别性以及模型的泛化性。然而，目前部分少样本方法对于特征学习阶段关注不够，或直接使用一些通用的特征学习方法训练模型，使得在基类上训练的模型在新类上的特征提取能力不足。因此，如何使用基类数据训练一个迁移能力强、泛化性能好的特征提取网络是当前少样本分类面临的一个挑战。

（2）由于在新类执行少样本分类任务时，采样的任务标注样本数量极少，仅仅根据少量样本的视觉特征可能无法捕获类别的代表性特征，因此很多方法引入语义信息以对视觉信息进行补充，进而提高模型在新类上的泛化能力。但如何设计一种简单有效的手段既能够利用语义信息丰富样本特征的信息来源，又不需要复杂的训练流程及模块设计仍需要进一步探讨。

1.3 问题三

鉴于当前少样本分类问题中存在的模型特征提取能力不够强、少量样本视觉特征不具有代表性的问题，本文致力于通过充分挖掘数据间的多元关系对其进行解决。本文通过研究基于多元关系建模的少样本分类算法，以对基类与新类共享的深层次数据关系进行挖掘，从而理解数据间的内在联系，将在基类数据上学习到的知识更好地迁移至新类，提升模型在数据匮乏的新类上的分类性能。基于此，本文分别对多粒度样本关系以及语义-视觉多空间关系进行了建模，充分有效地利用了样本间的不同关系以及语义信息，提升了模型的特征提取能力和泛化能力。本文研究内容详细介绍如下：

（1）基于多粒度样本关系建模的少样本分类研究

针对少样本分类模型特征提取能力不足的问题，本文提出了一种多粒度样本关系对比学习（Multi-Grained Sample Relation Contrastive Learning，简称 MGSRCL）方法，旨在对不同粒度的样本关系进行建模以提升模型的知识迁移能力。MGSRCL 方法将样本关系细致地划分为三种：同一样本不同变换版本之间的样本内关系、同类样本的类内关系和不同类样本的类间关系。通过对不同样本关系针对性地设计对比学习任务，MGSRCL 合理地对多种粒度的样本关系进行约束和优化，提升了模型所提取特征的判别性和泛化性。在 miniImageNet[14]、tierdImageNet[39]、CIFAR-FS[40] 和 CUB-200-2011[41] 四个少样本分类基准数据集上的大量实验表明，MGSRCL 方法通过充分挖掘样本关系提升了模型的特征提取能力，达到了优异的分类准确率。

（2）基于语义-视觉多空间关系建模的少样本分类研究

针对仅根据少量样本的视觉特征无法捕获类别代表性特征的问题，本文进一步引入语义信息作为视觉信息的补充，提出了语义-视觉多空间映射适配（Semantic-Visual Multi-Space Mapping Adapter，简称 SVMSMA）模型。该模型利用自然语言模型或多模态模型的文本编码器提取语义信息，将其通过语义-视觉多空间映射网络映射到视觉空间，并设计跨模态分类和特征对齐策略，使模型能够对语义信息与视觉信息的关系进行建模，丰富了样本特征的信息来源，使其更具有代表性，从而增强了模型对新类别的适应性和泛化能力。本方法同样在四个少样本分类基准数据集进行了大量实验，在 MGSRCL 模型的基础上取得了进一步的性能提升。

本文提出的 MGSRCL 模型与 SVMSMA 模型分别从多粒度样本关系和语义-视觉多空间关系两个角度出发，对少样本分类中的多元关系进行了深入挖掘与研究。MGSRCL 模型通过对多种粒度的样本关系进行不同的对比学习任务，对数据间的多种样本关系进行有效建模，提升了模型的特征提取能力；SVMSMA 模型则进一步引入类别的语义信息，使用两种跨模态任务对数据在语义与视觉不同空间之间的关系进行建模，提高了模型的泛化能力。通过这两个模型，本文有效地利用了数据中的多元关系，取得了较好的少样本分类结果。

本文的创新之处主要体现在以下两个方面：

（1）多粒度样本关系的深入挖掘：重新对样本关系进行了思考与划分并提出了基于样本内关系、类内关系和类间关系的多粒度样本关系对比学习方法，充分利用了样本之间复杂且多样的关系，为少样本分类提供了一个有效的特征学习方法。

（2）语义-视觉多空间关系的建模：通过融合语义信息和视觉信息，提出了一种简单有效的少样本分类模型，该模型可以通过跨模态的特征学习和原型对齐，有效利用语义信息对视觉信息进行补充，从而进一步提升少样本分类的性能。

1.4 本文的组织结构

本文的组织结构图如图1.1所示，共分为 5 个章节，各章节的介绍如下：

第一章：绪论。介绍少样本分类的研究背景和意义，并分析总结少样本分类算法的国内外研究现状及存在的挑战。最后对本文的研究内容和组织结构进行概述。

第二章：相关研究技术与理论。首先对少样本分类进行了进一步详细介绍，然后介绍本文方法中所使用到的对比学习技术以及语义信息表示，最后对本文实验所使用到的数据集及评价指标进行介绍。

第三章：基于多粒度样本关系建模的少样本分类研究。首先对部分现有少样本特征学习算法的不足进行分析，提出了基于多粒度样本关系对比学习的少样本特征学习算法（MGSRCL），随后详细介绍了针对不同粒度样本关系的建模方法，

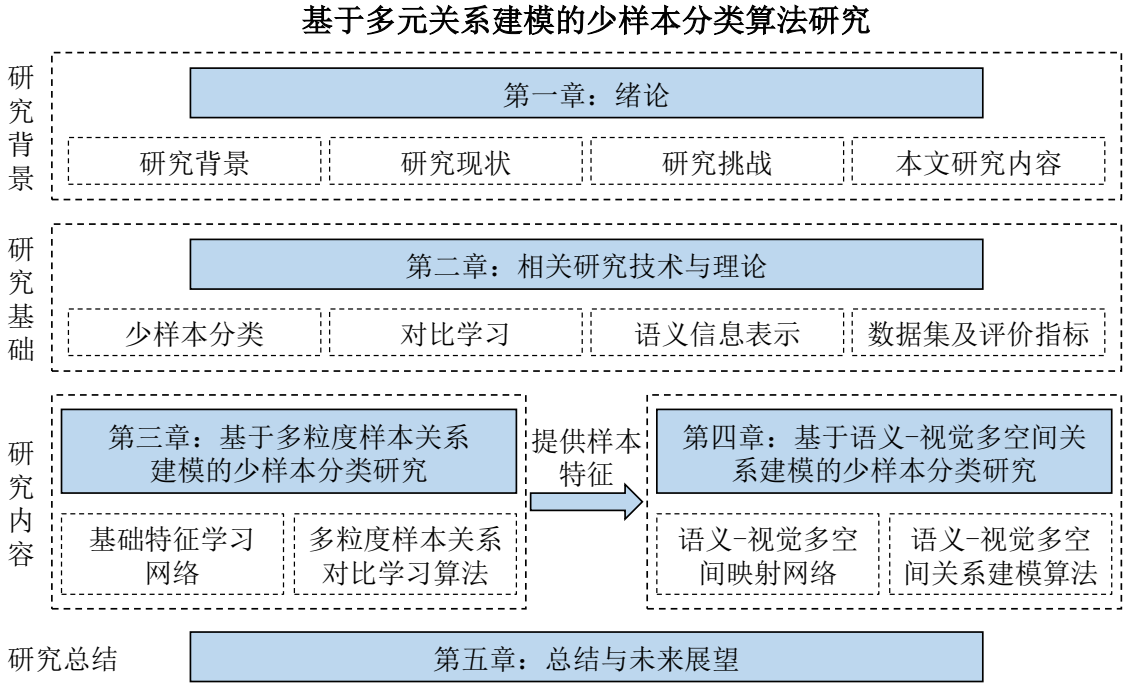

图 1.1 本文组织结构图。

Fig. 1.1 The organizational structure of the paper.

最后通过在四个基准数据集的大量实验证明了 MGSRCL 模型的有效性。

第四章：基于语义-视觉多空间关系建模的少样本分类研究。首先对基于语义的少样本分类算法进行分析介绍，提出了基于语义-视觉多空间关系建模的少样本特征适配算法（SVMSMA），然后介绍了 SVMSMA 模型的模型框架以及提出的跨模态分类和跨模态特征对齐模块，最后对所提方法进行了实验分析。

第五章：总结与未来展望。总结并分析了本文提出的基于多元关系建模的少样本分类算法研究的成果及不足，并对未来的研究方向与内容进行了展望。

2 模型的假设

本章内容共分为五节，第一节详细介绍少样本分类任务；第二节介绍与第三章方法相关的对比学习工作；第三节介绍第四章方法应用到的语义信息表示及一些用来获得语义信息的自然语言处理模型和多模态模型；第四节对本文所使用的数据集和评价指标进行介绍，第五节对本章进行小结。

2.1 少样本分类

少样本分类，旨在模拟人类识别新类别的过程，希望模型在拥有大规模标注数据的类别上进行训练之后，能够总结并迁移所学知识到新的类别，以实现在新类别上仅用少量标注数据进行训练便能够达到良好效果的目的。与常规分类任务将数据集划分为训练集与测试集不同，少样本分类数据集被划分为基类数据和新类数据，两者类别互不相交。其中，基类数据与普通分类任务的训练集一致，所有数据均可以被用来训练模型，无论是以元学习还是以普通分类任务的训练方式。而新类数据则是用来测试模型性能，在少样本分类的测试过程中，会在新类数据集上随机采样大量分类任务，每个任务的数据又被划分为支持集与查询集，如图2.1所示。其中，支持集数据为带有标注的样本，可用来微调整个模型和重新训练分类器，而查询集作用则是类似普通分类任务中的测试集，用来评估模型准确率。根据采样任务中类别数目 N 和样本数目 K 的多少，其又可被称为 N-way K-shot 任务，N 通常取 5，K 通常取 1 或 5。最终，通过对大量采样任务分别进行评估，并计算这些任务的平均准确率作为模型性能的最终评价指标。

图 2.1 少样本分类测试任务示意图。

Fig. 2.1 Illustration of few-shot classification testing tasks.

2.2 对比学习

在计算机视觉领域，特征学习的方法越来越多样化。其中，对比学习以其独特的学习机制，即通过比较样本之间的相似性和差异性来提取鲜明且有区分度的特征表征，近年来受到了广泛的关注和研究。在图像处理任务中，对比学习已经证明了其在提高模型泛化能力和识别精度方面的显著效果，并被广泛应用到少样本分类问题中。根据对比学习是否使用数据集标签信息，可以将其分为无监督对比学习和有监督对比学习，以下将分别进行介绍。

2.2.1 无监督对比学习

无监督对比学习不依赖于标注数据，它通常采用正负样本对的形式来构建训练任务。正样本对通常来自于同一实例的不同视角（例如，同一图像的不同数据增强版本），而负样本对则来自于不同实例。模型的目标是使得正样本对在表示空间中彼此接近，而负样本对彼此远离。该过程一般通过最小化 InfoNCE 损失函数实现，该损失函数如下式所示，

$$\mathcal{L} = -\log \frac{\exp(cos(f(x), f(x^+))/\tau)}{\exp(cos(f(x), f(x^+)) + \sum_{j=1}^{N} \exp(cos(f(x), f(x^-))}. \tag{2.1}$$

其中，图像 x 经由网络 $f(\cdot)$ 后映射到特征空间，x^+，x^- 分别代表 x 的正样本以及负样本，N 为负样本数量。$cos(\cdot)$ 是余弦相似度，$\exp(\cdot)$ 为以 e 为底的指数函数。

Chen 等人[42] 提出了一个简单有效的无监督对比学习框架-SimCLR，旨在通过比较不同视角下图像的特征表示来学习强大的特征提取网络。SimCLR 的核心思想是利用数据增强来产生正样本对，即从同一张图像中通过随机的数据增强操作（如裁剪、颜色变换等）生成两个视角，然后使来自同一图像的特征相互靠近，同时使得来自不同图像的特征尽可能地远离。尽管 SimCLR 在无监督特征学习方面取得了显著的成果，但其有一个明显缺点，即 SimCLR 的效果很大程度上依赖于对比损失函数中大量不同的负样本对，为了达到最佳性能，需要批次大小很大，这对计算资源的要求较高。He 等人[43] 提出的 MoCo 算法通过引入一个动态字典来存储样本特征表示解决了此问题。这个字典是一个队列，新的样本特征进入队列时，旧的样本特征被移除，以保持队列的固定大小。MoCo 可通过此字典高效地采样大量负样本，因此不再需要使用很大的批次便可达到最佳效果。这些无监督对比学习方法特别适合于数据量大但未标注的场景，能够有效地利用大量未标注数据来学习有意义的特征表示。

2.2.2 有监督对比学习

虽然无监督对比学习为使用大量无标注数据训练一个好的预训练模型提供了有效途径，但因为其在样本建模过程中将样本 x 与其负样本距离推远，而负样本中可能包含 x 的同类样本，这可能会学习到错误的样本关系。因此，Khosla 等人[44]

提出了有监督对比学习（Supervised Contrastive Learning，简称 SupCon）对这个问题进行解决。SupCon 是对比学习的一种变体，它结合了监督信号来进一步提升学习效率和特征表示的质量。与无监督对比学习相比，有监督对比学习在构造正负样本对时利用了标签信息，以确保模型不仅学会区分不同的样本，而且能够区分不同的类别，如图2.2所示（此图来源于 SupCon[44]）。SupCon 不仅保留了无监督对比学习中正样本对的概念，更进一步地，将属于同一类别的不同样本也视为正样本对，负样本对则是来自不同类别的样本，以此强化模型对不同类别间差异的识别能力。这种方法有效地缩小了同类样本间的表征距离，同时增强了不同类别间表征的区分度，有助于提升模型在复杂视觉任务中的表现。

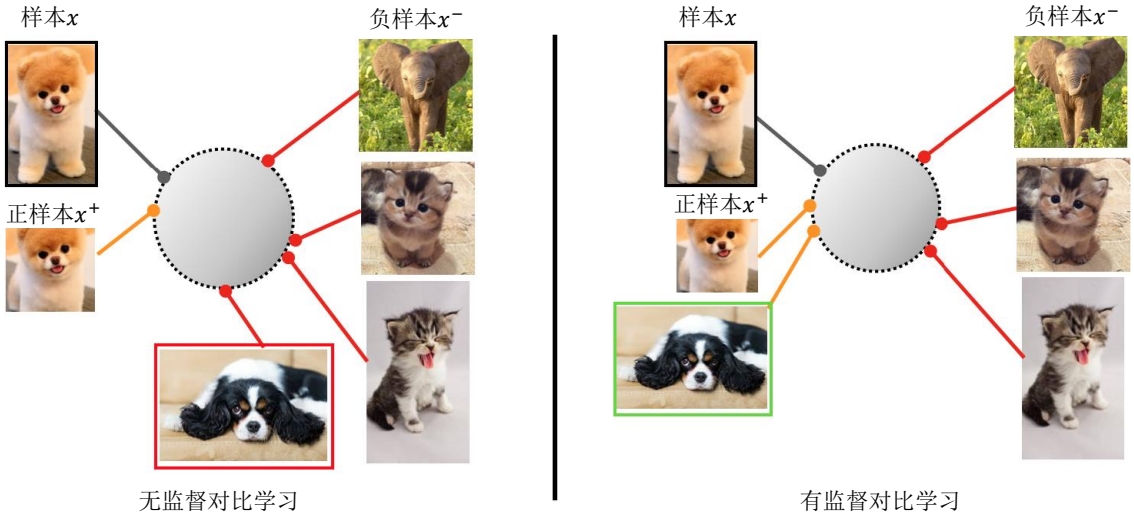

图 2.2 无监督对比学习与有监督对比学习。

Fig. 2.2 Unsupervised Contrastive Learning VS Supervised Contrastive Learning.

2.3 语义信息表示

目前，在少样本分类问题中，很多工作开始使用语义信息以对视觉信息进行补充，使用的语义信息一般为用自然语言处理（Natural Language Processing，简称 NLP）模型或多模态模型的文本编码器提取的语义特征。提取语义特征时，会将类别名称或提示文本与类别名称进行拼接之后的文本输入文本编码器，然后得到编码器输出的语义向量作为语义特征。以下对少样本分类中经常使用的语义特征提取模型进行介绍。

2.3.1 Word2Vec

Mikolov 等人[31,45] 提出的 Word2Vec 是一种广泛使用的自然语言处理技术，它从大量文本语料中以无监督的方式学习语义知识，旨在将词汇映射到稠密向量空间中，其中语义相似的词汇会在向量空间中彼此接近。Word2Vec 包含两种训练模型：连续词袋（Continuous Bag-of-Words，简称 CBOW）模型和跳跃（Continuous

Skip-gram，简称 Skip-Gram）模型，如图2.3所示（此图来源于 Word2Vec[31]）。

CBOW 模型：CBOW 模型通过上下文（周围的词汇）来预测当前词，如图2.3（左）所示。具体来说，它将上下文中的多个词汇作为输入，并尝试预测在这些上下文词汇中间的目标词汇。这个模型特别适合处理较小的数据集。

Skip-Gram 模型：与 CBOW 相反，Skip-Gram 模型使用一个词来预测其周围的上下文，如图2.3（右）所示。给定一个特定的词，目标是预测在一个特定范围内的前后词汇。Skip-Gram 模型在处理大数据集时表现更好，尤其是对罕见词汇的表示更为有效。

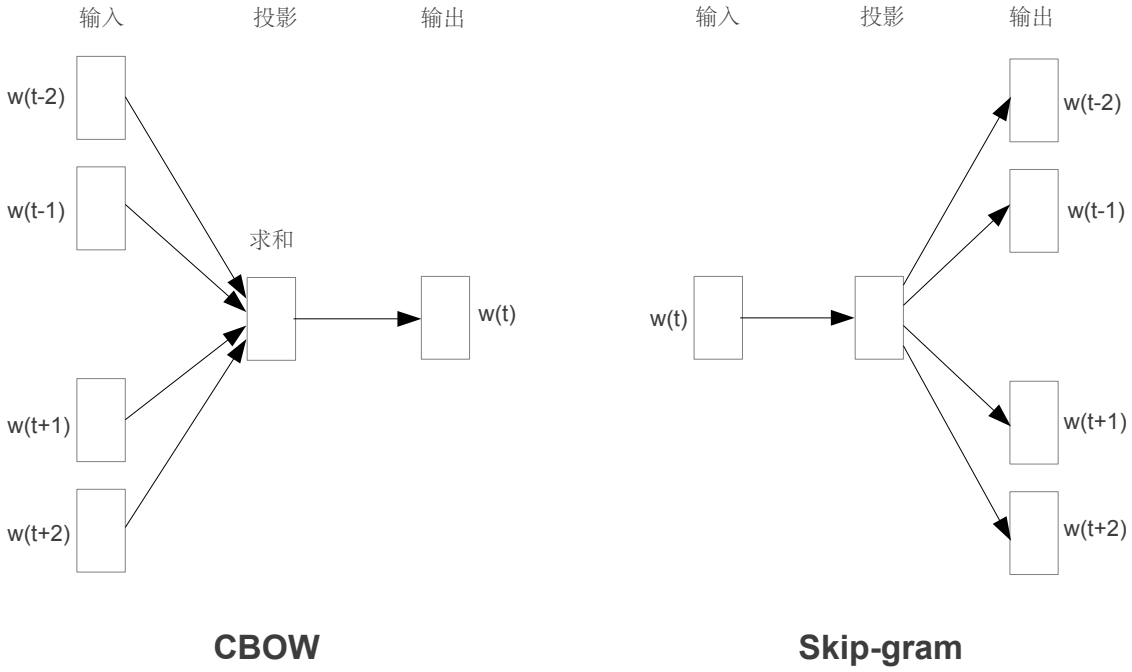

图 2.3 CBOW 模型与 Skip-Gram 模型示意图。

Fig. 2.3 Illustration of CBOW model and Skip-Gram model.

Word2Vec 的核心优势在于它能够捕捉到词汇之间的细微语义关系，并通过向量运算来揭示词汇之间的语义相似性。这使得 Word2Vec 在诸多自然语言处理任务中得到广泛应用，包括文本相似性度量、情感分析、机器翻译以及作为深度学习模型的预训练层等，另外，很多计算机视觉任务也使用 Word2Vec 来提取语义特征以对视觉特征进行修正或补充。

2.3.2 GloVe

Pennington 等人[32] 提出的 GloVe（Global Vectors for Word Representation）也是一种用于词嵌入的无监督学习算法。该模型旨在将单词映射到一个向量空间中，使得这些向量能够捕捉到词与词之间的共现关系，从而反映出词义的复杂模式和结构。GloVe 模型的关键创新在于它结合了两种主流的词表示方法的优点：基于

全局矩阵分解（Global Matrix Factorization）的方法和基于局部上下文窗口（Local Context Window）的方法。

GloVe 的核心思想是首先构建一个全局词共现矩阵，记录整个语料库中各个词之间的共现次数，然后通过优化一个目标函数来学习词向量。这个目标函数旨在让共现次数的对数值与相应词向量的点积尽可能接近，同时引入偏置项来进一步提升模型的灵活性和准确性。具体来说，GloVe 构建一个大型的词-词共现矩阵，矩阵中的每个元素代表了两个词在一定窗口大小内共同出现的次数。这一步捕获了全局的共现统计信息。然后其定义了一个特殊的损失函数，该损失函数不仅关注词对之间的共现概率，而且关注共现概率的比例，这有助于捕获词义之间更细微的差别。这个损失函数同时考虑到了共现次数的稀疏性和不均匀性。通过最小化损失函数，模型学习到的词向量能够反映出词与词之间的共现概率，这意味着词向量空间中的距离可以表示词义之间的相似度。这一步既利用了局部信息（通过具体的共现频率），也综合了全局信息（通过整个语料库的统计数据）。

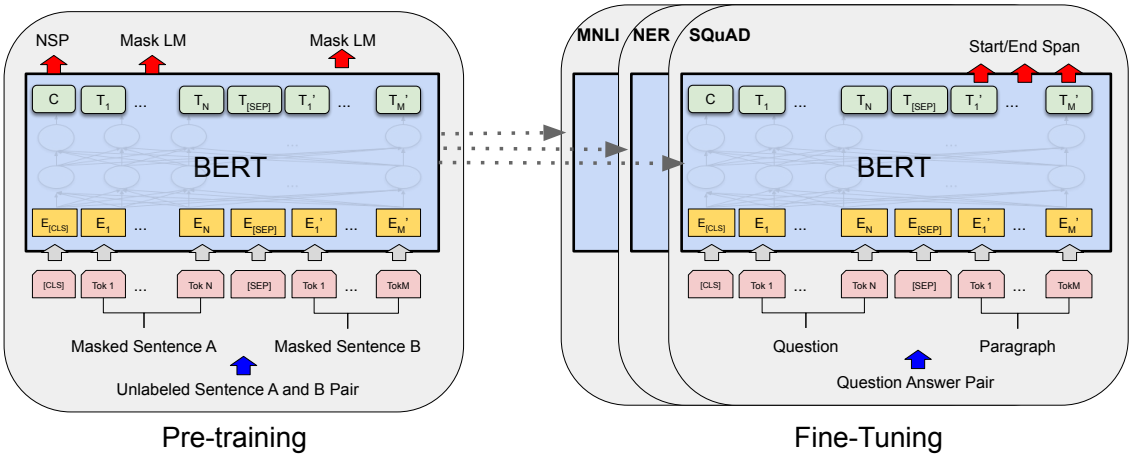

图 2.4 BERT 的整体预训练和微调过程。

Fig. 2.4 Overall pre-training and fine-tuning procedures for BERT.

2.3.3 BERT

Devlin 等人[33] 提出的 BERT（Bidirectional Encoder Representations from Transformers）模型是一种革命性的自然语言处理模型。该模型利用了 Transformer 架构的双向编码器，能够理解语言的深层语义和上下文关系。BERT 的创新之处在于其基于 Transformer 模型的编码器，使得它能够同时考虑词语左侧和右侧的上下文信息，这与以往的单向模型或浅层双向模型不同，使其能够更准确地理解词义。如图2.4所示（此图来源于 BERT[33]），BERT 模型首先在大规模的文本语料库上进行预训练，学习通用的语义表示，然后针对具体的 NLP 任务进行微调，这一过程极大提升了模型在特定任务上的性能。由于其良好的性能与开创性，后续又出现了诸如 SBERT[46]、RoBERTa[47]、ALBERT[48] 等改进工作。BERT 模型通过两种类型

的预训练任务学习语义表示：

（1）掩码语言模型（Masked Language Model，简称 MLM）：在训练过程中，BERT 会随机遮蔽模型输入句子中的一部分词语（使用 [MASK] token 代替原有输入），然后让模型预测这些遮蔽的词语，这可以迫使模型学习到词语的双向上下文关系。另外，为了解决模型微调期间从未看到 [MASK] token 的问题，BERT 模型不总是直接用 [MASK] token 代替所选单词，而是将所选单词 80% 的概率替换为 [MASK] token，10% 的概率用一个随即单词替换所选单词，剩下 10% 的概率则是保持其不变。

（2）下一句预测（Next Sentence Prediction，简称 NSP）：由于很多 NLP 下游任务都是基于理解两个句子之间的关系，如问答和自然语言推断，因此 BERT 设计了一个下一句预测的任务。给定两个句子 A 和 B，模型需要预测 B 是否是 A 的下一句，这可以帮助模型理解句子间的关系。

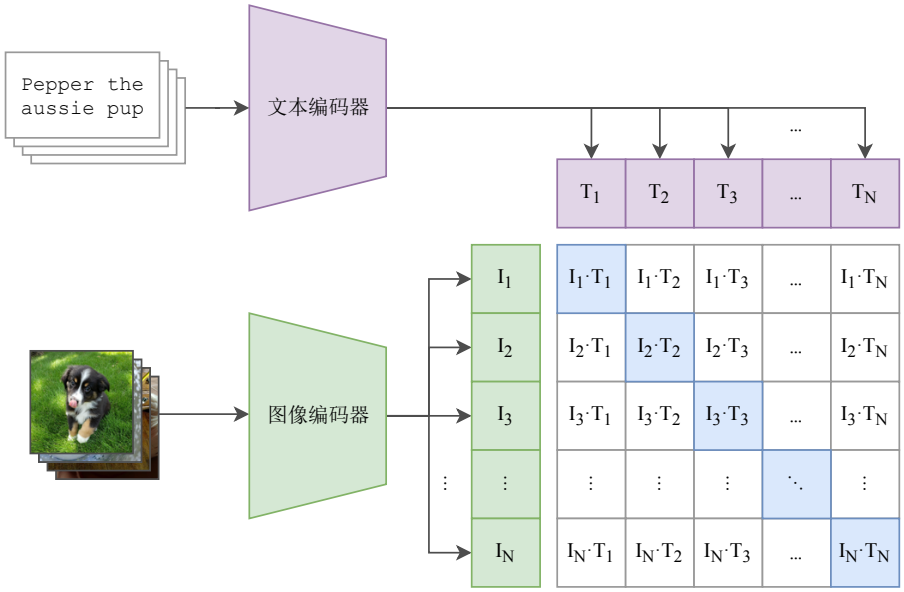

图 2.5 CLIP 模型预训练示意图。

Fig. 2.5 Illustration of pre-training CLIP model.

2.3.4 CLIP

近年来，Radford 等人[34] 提出的 CLIP（Contrastive Language-Image Pre-training）模型受到了很多研究者的关注，并促进了多模态大模型和一些其他任务的发展。CLIP 旨在通过大规模的图文对比学习来同时理解图像和文本，并建立它们之间的联系。CLIP 模型的创新之处在于其跨模态能力，它不仅能理解图片内容，也能理解与图片内容相对应的文本描述，从而在多种视觉任务上展示出了卓越的性能和强大的泛化能力，并提供了一个充分建模文本关系的文本编码器。

如图2.5所示（此图来源于 CLIP[34]），CLIP 由两部分组成：一个图像编码器

和一个文本编码器。图像编码器负责提取图像的视觉特征，而文本编码器则提取文本的语义特征。这两个编码器可以是任何形式的神经网络。在原始 CLIP 模型中，图像编码器基于 Vision Transformer（ViT）或 ResNet 架构，而文本编码器基于 Transformer 架构。CLIP 的训练过程涉及大量图像和文本对的对比学习。具体来说，模型训练的目标是最大化相匹配的图像和文本对之间的相似度，同时最小化不匹配对的相似度。这种训练方式使得 CLIP 学习到的特征表示能够跨越视觉和语义的界限，理解两种模态之间的对应关系。

2.4 数据集及评价指标

本文使用四个少样本分类基准数据集对模型性能进行评估，包括三个普通少样本分类数据集：miniImageNet[14]、tieredImageNet[39]、CIFAR-FS[40]，以及一个细粒度数据集：CUB-200-2011（CUB）[41]，以下对其进行分别介绍，并对少样本分类评价指标进行描述。

表 2.1 miniImageNet、CIFAR-FS 和 CUB 的数据集划分。

Table 2.1 Dataset partition of miniImageNet, CIFAR-FS and CUB.

数据集	类别数目			
	训练集	验证集	测试集	总数
miniImageNet	64	16	20	100
CIFAR-FS	64	16	20	100
CUB	100	50	50	200

表 2.2 tieredImageNet 的数据集划分。

Table 2.2 Dataset partition of tieredImageNet.

类别层级	类别数目			
	训练集	验证集	测试集	总数
超类	20	6	8	34
子类	351	97	160	608

2.4.1 数据集

miniImageNet 数据集[14] 和 tieredImageNet 数据集[39] 均为 ImageNet[49] 的子集。其中，miniImageNet 数据集包含 100 个类别，每个类别有 600 张图像。本文遵循 Ravi 等人[50] 提出的划分准则，训练集、验证集和测试集分别包含 64、16 和 20 个类别。tieredImageNet 数据集则包含 34 个超类（608 个子类），分为 20 个训练类别（351 个子类）、6 个验证类别（97 个子类）和 8 个测试类别（160 个子类）。CIFAR-FS 数据集[40] 源自 CIFAR-100 数据集，该数据集包含 64 个训练类别、16 个验证

类别和 20 个测试类别，每个类别同样有 600 张图像。Caltech-UCSD Birds(CUB）-200-2011（简称 CUB）数据集[41] 则是一个包含不同种类的鸟类细粒度图像数据集，包含 11788 个图像样本，分为 200 个类别。根据 Triantafillou 等人[51] 的划分准则，该数据集包含 100 个训练类别、50 个验证类别和 50 个测试类别。各数据集划分如表2.1和2.2所示。

2.4.2 评价指标

对于所有数据集，本文评估 5-way 1-shot 以及 5-way 5-shot 少样本分类任务性能。在一次模型评估中，本文方法采样 2000 个少样本分类任务，并计算了 95% 置信区间的平均分类准确率作为模型的评价指标。在一个少样本分类任务中，每个类别的支持集样本数目为 1 或 5（根据任务决定），查询集样本数目为 15，与其他方法[21,25] 保持一致。

2.5 本章小结

本章首先详细介绍了少样本分类任务的定义及其训练测试过程。然后对后续研究工作所涉及到的相关技术进行了介绍，其中包括第三章所使用到的对比学习技术，根据是否使用数据集标签信息将其分为无监督对比学习和有监督对比学习进行了详细阐述；以及第四章所使用到的语义信息表示，介绍了如何提取语义信息表示和少样本分类中常用的语义特征提取模型。最后，介绍了本文方法所使用到的少样本分类数据集和评价指标。

3　基于多粒度样本关系建模的少样本分类研究

本章研究基于多粒度样本关系建模的少样本特征学习算法，通过挖掘多种粒度的样本关系并对其进行建模从而增强模型的特征提取能力，进而提升少样本分类任务的准确率。本章内容共分为四节，第一节介绍研究动机和方法概述；第二节介绍本章提出的基于多粒度样本关系对比学习的少样本特征学习算法；第三节给出实验设置和结果分析；第四节对本章进行小结。

3.1 引言

3.1.1 研究动机

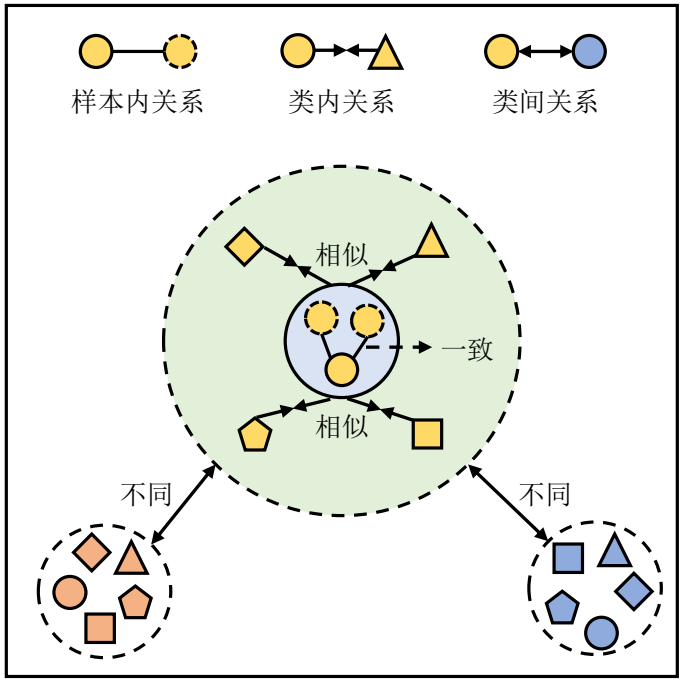

图 3.1 样本关系示意图。在该图中，不同形状与颜色分别代表不同样本与类别。同一样本的不同变换由相同的颜色和形状表示。样本关系包括三种类型: 样本内关系、类内关系和类间关系。本章所提方法约束同一样本不同变换版本在语义内容上保持一致，同类样本保持相似，非同类样本保持不同。

Fig. 3.1 Illustration of sample relations. In this figure, different shapes and different colors represent different samples and different classes, respectively. Different transformations of the same sample are represented by the same color and shape. The sample relations contain three types: intra-sample relation, intra-class relation and inter-class relation. The approach proposed in this chapter enforces different transformations to be consistent in semantic content, homogenous samples to be similar, and inhomogeneous samples to be different.

3.1.2 方法概述

3.2 基于多粒度样本关系对比学习的少样本特征学习算法

在本节中，首先对少样本分类任务及其符号定义进行介绍；然后对所提出的基于多粒度样本关系对比学习的少样本特征学习模型进行简要介绍；接下来详细介绍了所提模型的各个模块及其损失优化；最后介绍了模型总体优化目标以及模型推理过程。

3.2.1 符号定义

在本章中，少样本分类任务的基类数据集和新类数据集分别表示为：

$$\begin{aligned}\mathcal{D}_{base} &= \{(x,y)|x \in X^{base}, y \in Y^{base}\},\\ \mathcal{D}_{novel} &= \{(x,y)|x \in X^{novel}, y \in Y^{novel}\}.\end{aligned} \tag{3.1}$$

其中，$\mathcal{D}_{base}$ 所包含的类别 $\mathcal{C}_{base}$ 和 $\mathcal{D}_{novel}$ 所包含的类别 $\mathcal{C}_{novel}$ 不相交。另外，x、y 分别表示样本图像和样本标签；X^{base}、Y^{base} 和 X^{novel}、Y^{novel} 分别表示基类数据和新类数据的样本图像集合和标签集合。

$\mathcal{D}_{base}$ 用于在预训练阶段训练一个具有良好泛化性能的模型，$\mathcal{D}_{novel}$ 用于测试过程采样大量 N-way K-shot 少样本分类任务并计算平均准确率来评估模型性能。每个少样本分类任务 $\mathcal{T}$ 包括一个支持集 $\mathcal{S}_{\mathcal{T}}$ 和一个查询集 $\mathcal{Q}_{\mathcal{T}}$，

$$\mathcal{T} = \{\mathcal{S}_{\mathcal{T}}, \mathcal{Q}_{\mathcal{T}}\}. \tag{3.2}$$

其中，$\mathcal{S}_{\mathcal{T}}$ 包含来自 N 个类别的 $N \times K$ 个标注样本，而 $\mathcal{Q}_{\mathcal{T}}$ 包含来自相同 N 个类别的 $N \times Q$ 个样本，并且 $\mathcal{S}_{\mathcal{T}}$ 和 $\mathcal{Q}_{\mathcal{T}}$ 中的样本是没有交集的。在测试阶段，针对每个采样的少样本分类任务使用 $\mathcal{S}_{\mathcal{T}}$ 重新训练一个分类器，使用 $\mathcal{Q}_{\mathcal{T}}$ 来评估分类器性能。

3.2.2 整体框架

本章重新审视了对比学习中的样本关系，并根据样本关系粒度的不同将其划分为三种类型：同一样本在不同变换下的样本内关系（intra-sample relation）、同类样本的类内关系（intra-class relation），以及不同类样本的类间关系（inter-class relation）。基于此，本章提出了一种新颖的多粒度样本关系对比学习方法（Multi-Grained Sample Relation Contrastive Learning，简称 MGSRCL），通过对少样本分类中不同粒度的样本关系进行建模从而获得了一个强大的特征提取网络。如图3.2所示，MGSRCL 模型包含三个主要部分：基础特征学习网络（Base Feature Learning Network，简称 Base）、变换一致性学习（Transformation Consistency Learning，简称 TCL）模块和类对比学习（Class Contrastive Learning，简称 CCL）模块。具体而言，基础特征学习网络是通过一般图像分类任务训练的神经网络。TCL 模块旨

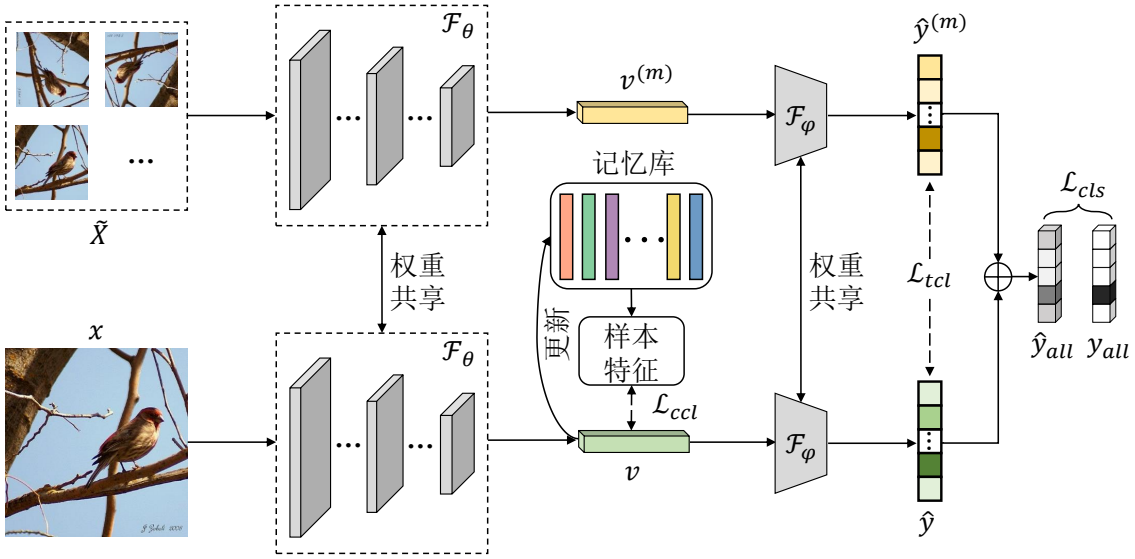

图 3.2 多粒度样本关系对比学习模型（MGSRCL）示意图。它包含一个特征提取网络 $\mathcal{F}_\theta$ 和一个分类器 $\mathcal{F}_\varphi$。在此图中，v 和 $v^{(m)}$ 代表原始图像 x 及其第 m 个变换版本 $x^{(m)}$ 的特征，其中 $x^{(m)} \in \widehat{X}$。$\bigoplus$ 是一个连接操作，用于对原始图像的预测输出 $\widehat{y}$ 与 M 个变换的预测输出 $\widehat{y}^{(1)}, ..., \widehat{y}^{(m)}, ..., \widehat{y}^{(M)}$ 进行连接。记忆库（Memory Bank）用于存储特征。$\mathcal{L}_{cls}$，$\mathcal{L}_{tcl}$ 和 $\mathcal{L}_{ccl}$ 分别是分类损失、变换一致性学习（TCL）损失和类对比学习（CCL）损失。为了便于阅读，此图中没有展示自监督模块。

Fig. 3.2 Illustration of Multi-Grained Sample Relation Contrastive Learning (MGSRCL) model. It contains a feature extraction network $\mathcal{F}_\theta$ and a classifier $\mathcal{F}_\varphi$. In this figure, v and $v^{(m)}$ represent the features of the original image x and its m-th transformed version $x^{(m)}$, where $x^{(m)} \in \widehat{X}$. $\bigoplus$ is a concatenation operator for the predicted output $\widehat{y}$ of the original image and the predicted outputs $\{\widehat{y}^{(1)}, ..., \widehat{y}^{(m)}, ..., \widehat{y}^{(M)}\}$ of M transformations. Memory bank is used to store the features. $\mathcal{L}_{cls}$, $\mathcal{L}_{tcl}$, and $\mathcal{L}_{ccl}$ are the classification loss, transformation consistency learning (TCL) loss, and class contrastive learning (CCL) loss, respectively. For the sake of legibility, the self-supervised module is not shown in this image.

在确保同一样本的不同变换版本具有一致的语义内容。而 CCL 则用于确保同类样本具有相似的语义内容，以及非同类样本具有不同的语义内容。接下来，本节将对 MGSRCL 方法的每个部分进行更为详细的阐述。

3.2.3 基础特征学习网络

如图3.2所示，特征提取网络，表示为带有参数 θ 的 $\mathcal{F}_\theta$，被用于提取图像特征。设 $(x, y) \in \mathcal{D}_{base}$ 表示从 $\mathcal{D}_{base}$ 中采样的图像及其对应的标签。图像 x 的特征向量 v 可以通过 $\mathcal{F}_\theta$ 获得：$v = \mathcal{F}_\theta(x)$。然后，使用参数为 φ 的分类器 $\mathcal{F}_\varphi$，将特征向量 v 投影到标签空间，以获得预测的置信度分数 p：$p = \mathcal{F}_\varphi(v)$。最后，通过在 p 上应用 Softmax 函数，可以得到预测概率输出 $\widehat{y}$：$\widehat{y} = \text{Softmax}(p)$。基础特征学习网络的参数 θ 和 φ 通过最小化整个基类数据集 $\mathcal{D}_{base}$ 上的分类损失 $\mathcal{L}_{cls}$ 来进行优化，其可以

表示为以下公式，

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{D}_{base}|} \sum_{\{x,y\}\in\mathcal{D}_{base}} y \log \widehat{y}. \tag{3.3}$$

为了防止在训练集上过拟合，许多方法[25,26,29]引入了变换样本参与训练，并使用自监督学习技术预测在训练过程中对图像执行了哪种变换以增强网络的特征提取能力。遵循这些方法，本文也添加了一个由多层感知机（Multilayer Perceptron，简称 MLP）构成的自监督（Self-Supervised，简称 SS）模块。设 $\widetilde{X} = \{\widetilde{x}^{(1)}, ..., \widetilde{x}^{(M)}\}$ 为一张图像的变换版本集合，其中 M 表示变换样本的总数，$\widetilde{x}^{(m)}$ 表示图像的第 m 个变换版本。$\widetilde{X}$ 可以通过在图像上应用一系列变换（如裁剪、调整大小、旋转等数据增强操作）获得。变换后的图像 $\widetilde{X}$ 和原始图像 x 同时输入模型，用于分类和自监督任务。自监督任务的目标是识别图像进行了哪种变换，其损失 $\mathcal{L}_{ss}$ 表示为以下公式，

$$\mathcal{L}_{ss} = -\frac{1}{|\mathcal{D}_{base}|}\frac{1}{M+1} \sum_{x\in\mathcal{D}_{base}} \sum_{m=0}^{M} s^{(m)} \log \widehat{s}^{(m)}, \tag{3.4}$$

其中 $\widehat{s}^{(m)}$ 和 $s^{(m)}$ 分别表示自监督任务中第 m 个变换版本的预测概率输出和真实标签。$s^{(0)}$ 是原始图像 x 的自监督标签。此外，增加了变换样本之后的分类损失可以重新定义为以下公式，

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{D}_{base}|}\frac{1}{M+1} \sum_{x\in\mathcal{D}_{base}} \sum_{m=0}^{M} y^{(m)} \log \widehat{y}^{(m)}, \tag{3.5}$$

$\widehat{y}^{(m)}$ 表示分类任务中的预测概率输出，$y^{(m)}$ 表示分类任务的真实标签。

最后，基础特征学习网络的损失 $\mathcal{L}_{base}$ 可以写为分类损失 $\mathcal{L}_{cls}$ 和自监督损失 $\mathcal{L}_{ss}$ 之和，

$$\mathcal{L}_{base} = \mathcal{L}_{cls} + \mathcal{L}_{ss}. \tag{3.6}$$

3.2.4 多粒度样本关系对比学习算法

（1）变换一致性学习

一个样本图像与其变换版本包含完全相同的对象和背景，仅因为进行了数据增强而使得图像在旋转角度、明暗、颜色等方面发生变化，但其内在的类别属性和语义内容应保持不变。为了实现这一目标，本文设计了一个变换一致性学习（Transformation Consistency Learning，简称 TCL）模块，以约束同一样本不同变换版本的样本内关系。TCL 模块通过约束一个样本和其变换版本的预测输出相同来确保它们具有一致的语义内容。这是因为预测输出反映了样本在每个类别中的预测概率，这些概率不仅表示了模型对于样本属于各个类别的置信度，而且深入地揭示了样本的本质属性——语义内容。

本章方法将一个样本与其变换版本同时输入网络，并在预测标签输出层面计

算它们的 TCL 损失。这里，本文使用 Jensen-Shannon 散度[52,53] 作为 TCL 损失，它能够衡量两个概率分布的差异，通过最小化两个预测标签的输出，可以使其概率分布一致，从而达到使样本和其变换版本具有一致语义内容的目的。TCL 损失可以写为以下公式，

$$\mathcal{L}_{tcl} = \frac{1}{|\mathcal{D}_{base}|} \sum_{x \in \mathcal{D}_{base}} \frac{1}{M} \sum_{m=1}^{M} JS(\widehat{y}_{\tau_1}, \widehat{y}_{\tau_1}^{(m)}), \tag{3.7}$$

其中 $\widehat{y}_{\tau_1}$ 和 $\widehat{y}_{\tau_1}^{(m)}$ 分别是原始图像和第 m 个变换图像的平滑标签输出。它们通过以下公式获得，

$$\widehat{y}_{\tau_1} = \text{Softmax}(p/\tau_1), \tag{3.8}$$

此公式中 $p = \mathcal{F}_{\varphi}(\mathcal{F}_{\theta}(x))$，$\tau_1$ 是一个温度参数，本文在实验中将其设置为 4.0。使用平滑标签输出的原因在于不同变换的输出不仅需要在最大预测概率的类别上保持一致，而且需要在所有其他类别上也保持一致，以确保它们具有完全相同的语义内容，而平滑标签输出可以提供更多关于概率分布差异的信息。

（2）类对比学习

同类样本虽然图像内包含了同一个类别的物体，但物体及其背景与同一图像不同变换版本相比差异性较大，因此其预测概率输出之间差异也会较大。如果强行将其预测输出进行对齐，可能会使得网络为了学习此种强关系而导致模型崩塌。但在另一方面，同类样本间距离比不同类样本间距离更近是毋庸置疑的。因此，本文采用类对比学习（Class Contrastive Learning，简称 CCL）以一种相对距离的形式约束同类样本的类内关系和不同类样本的类间关系。CCL 模块通过最大化同类样本特征的相似性，同时最小化不同类样本特征的相似性来在特征空间拉近同类样本，推远不同类样本。

与之前对比学习不同，CCL 模块为了将样本和其他每个不同类间的距离推远，对于每张图像都需要该图像的一个同类样本以及其他每个类别的不同类样本（之前对比学习通常随机采样,这使得每个批次计算损失时不同类样本可能仅来自部分不同类别）。为了实现这一目标并加快训练速度，本文使用了一个记忆库（Memory Bank）来存储和从中采样图像特征，记忆库存储了所有图像的特征。在一个批次中，CCL 模块从记忆库中为每类图像随机采样一个样本的特征。CCL 损失可以定义为，

$$\mathcal{L}_{ccl} = \frac{1}{|\mathcal{D}_{base}|} \sum_{x \in \mathcal{D}_{base}} -\log \frac{\exp(\frac{cos(v,v')}{\tau_2})}{\sum_{i=1}^{|\mathcal{C}_{base}|} \exp(\frac{cos(v,v_i)}{\tau_2})}, \tag{3.9}$$

其中 $|\mathcal{C}_{base}|$ 和 $|\mathcal{D}_{base}|$ 表示基类的类别数量和样本数量，v 和 v' 分别是某个样本及其同类样本的特征，v_i 代表来自第 i 类的样本的特征。这里 v' 和 v_i 是从记忆库中采样的。$cos(\cdot)$ 是余弦相似度，$\exp(\cdot)$ 为以 e 为底的指数函数。而 τ_2 是一个温度参

数，本文按照[42,44]的实验设置将其设为0.1。此外，记忆库的更新方式为，

$$v_k = r \times v_k + (1 - r) \times v_q, \tag{3.10}$$

v_q 和 v_k 分别代表在当前小批次中获得的图像特征以及在记忆库中存储的相同图像的特征，r 用于调整记忆库的更新速度，按照 IER 方法[25]的实验，本文将其设置为0.99。在训练阶段，记忆库每一轮训练过程都会完全更新一遍。

3.2.5 模型优化

结合公式3.6、3.7和3.9，本章提出的 MGSRCL 模型总体损失函数可以表示为以下公式，

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \alpha \cdot \mathcal{L}_{tcl} + \beta \cdot \mathcal{L}_{ccl}, \tag{3.11}$$

其中 α 和 β 是用于平衡不同损失的超参数，分别表示 TCL 模块和 CCL 模块的损失权重。

MGSRCL 模型通过在整个基类数据集上最小化上述损失函数对模型参数进行联合优化。通过建模多个粒度的样本关系，可以有效地增强模型的特征提取能力和泛化能力，帮助模型捕获更具判别性的特征，从而提高模型在新类 $\mathcal{D}_{novel}$ 上的分类性能。

3.2.6 模型推理

模型在基类数据集 $\mathcal{D}_{base}$ 训练完成之后，在测试阶段，将会冻结 MGSRCL 模型特征提取网络的所有参数，并通过解决来自新类 $\mathcal{D}_{novel}$ 的大量少样本分类任务来评估模型性能。在每个任务 $\mathcal{T}$ 的推理过程中，本文使用特征提取网络 $\mathcal{F}_\theta$ 来获得支持集 $\mathcal{S}_\mathcal{T}$ 和查询集 $\mathcal{Q}_\mathcal{T}$ 的图像特征。然后，本文使用 $\mathcal{S}_\mathcal{T}$ 的样本特征训练一个逻辑回归分类器 LC，并对 $\mathcal{Q}_\mathcal{T}$ 中的样本进行分类，最后将在多个少样本分类任务上的准确率平均值作为模型的评价指标。MGSRCL 模型的推理过程如图3.3所示。

3.3 实验设置及结果分析

在本节中

3.4 本章小结

本章

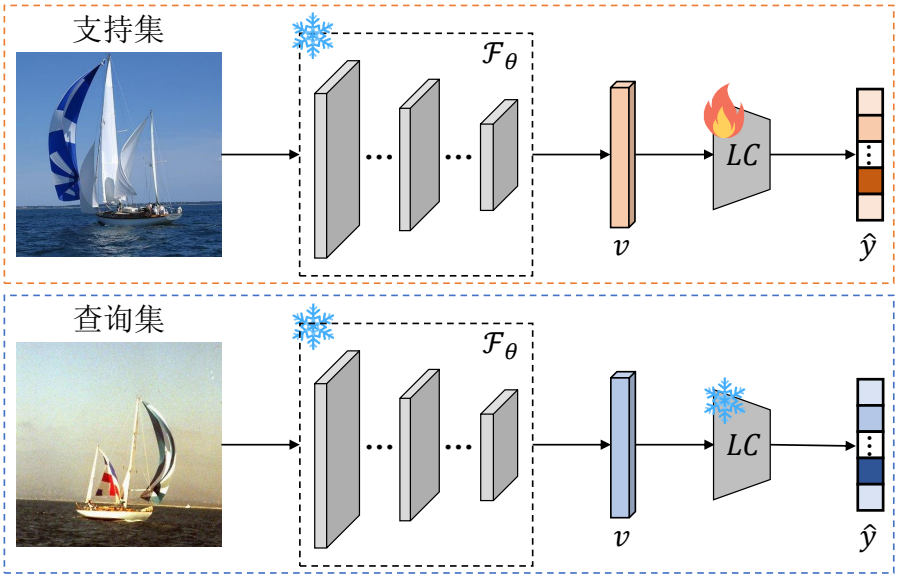

图 3.3 MGSRCL 模型推理过程示意图。推理过程中，使用冻结参数的特征提取网络 $\mathcal{F}_\theta$ 提取支持集与查询集的图像特征。其中，支持集特征被用来训练一个逻辑回归分类器 LC，查询集特征则是用来测试分类器性能。

Fig. 3.3 Illustration of MGSRCL model inference process. During the inference process, the feature extraction network $\mathcal{F}_\theta$ with frozen parameters is used to extract image features from both support set and query set. Herein, the support set features are utilized to train a logistic regression classifier LC, while the query set features are used to assess the classifier's performance.

4 基于语义-视觉多空间关系建模的少样本分类研究

上一章研究了基于多粒度样本关系建模的少样本特征学习算法，通过充分挖掘不同粒度的样本关系提高了网络所提取特征的质量。然而，该算法仅在视觉空间对多种样本关系进行了建模，忽略了数据集中所隐含的丰富语义信息，限制了模型通过基类数据进行训练来学习新类数据知识的能力。因此，在上一章的基础上，本章主要研究基于语义-视觉多空间关系建模的少样本特征适配算法，通过引入语义信息并与视觉信息进行建模从而丰富模型所获得的信息，增强模型的泛化能力。本章内容共分为四节，第一节介绍研究动机和方法概述；第二节介绍本章提出的基于语义-视觉多空间关系建模的少样本特征适配算法；第三节给出实验设置和结果分析；第四节对本章进行小结。

4.1 引言

4.2 基于语义-视觉多空间关系建模的少样本特征适配算法

4.3 实验设置及结果分析

4.4 本章小结

本章研究基于语义-视觉多空间关系建模的少样本特征适配算法，针对少样本分类中仅根据少量视觉特征无法捕获类别代表性特征的缺点，引入语义信息作为视觉信息的补充，通过对语义-视觉多空间关系进行建模，提出了语义-视觉多空间映射适配模型（Semantic-Visual Multi-Space Mapping Adapter，简称 SVMSMA），以丰富样本特征的信息来源，利用语义特征对视觉特征进行补充与修正，从而提升模型在新类上的泛化能力。SVMSMA 模型使用单/多模态映射网络将样本语义特征映射到视觉空间获得单/多模态映射特征，并通过跨模态分类（CMC）模块与跨模态特征对齐（CMFA）模块对映射网络进行优化，以使得语义特征与视觉特征建立联系。测试过程中，本章方法将支持集的视觉特征、单模态映射特征、以及多模态映射特征共同作为分类器的训练数据，达到了较仅使用单一特征时更好的分类结果。在 miniImageNet、tieredImageNet、CIFAR-FS 和 CUB-200-2011 数据集的大量实验表明了 SVMSMA 方法的有效性。

综上所述，本章提出的基于语义-视觉多空间关系建模的少样本特征适配算法通过对语义-视觉多空间关系进行建模，充分利用语义信息对视觉信息进行了补充，丰富了样本特征信息来源，提升了模型的泛化能力。

5 总结与未来展望

本章内容共分为两节，第一节对本文研究内容与方法进行总结；第二节介绍本文所提方法的局限性并对未来研究方向进行展望。

5.1 总结

少样本分类致力于模拟人类的知识迁移能力，期望模型在具有大量标注数据的基类数据上训练之后，能够将所学知识迁移至新类别，实现用少量标注样本进行有效学习。目前，少样本分类问题已取得一系列研究成果，但仍存在一些问题与挑战：特征提取网络迁移能力不够强；样本数量极少情况下无法捕获类别代表性特征。针对这些问题，本文分别从多粒度样本关系建模与语义-视觉多空间关系建模两个角度出发，对少样本分类中的多元关系进行了深入挖掘与研究。

（1）本文首先从多粒度样本关系建模的角度出发，开展了基于多粒度样本关系建模的少样本分类研究。针对少样本分类模型特征提取能力不足的问题，提出了一种基于多粒度样本关系对比学习的少样本特征学习算法：多粒度样本关系对比学习（Multi-Grained Sample Relation Contrastive Learning，简称 MGSRCL）模型，旨在通过对不同粒度的样本关系进行建模以提升模型的特征提取能力。MGSRCL 使用变换一致性学习来约束同一样本不同变换版本之间的样本内关系，通过使其预测概率分布相同令它们在语义内容上保持一致；使用类对比学习来约束同类样本的类内关系和不同类样本的类间关系，通过对其特征进行建模使同类样本语义内容相似、不同类样本语义内容不同。通过对多种粒度的样本关系细致地建模，MGSRCL 提升了模型的特征提取能力，达到了优异的少样本分类结果。在 miniImageNet、tieredImageNet、CIFAR-FS 和 CUB 四个少样本基准数据集上的大量实验证明了 MGSRCL 的有效性。另外，通过将 MGSRCL 模型作为预训练模型与其他方法结合，证明了所获得特征提取网络的可迁移性。

（2）上述提出的 MGSRCL 方法虽然达到了优异的结果，但仍存在没有利用样本语义信息的问题。因此，以 MGSRCL 方法为基础，本文进一步进行了基于语义-视觉多空间关系建模的少样本分类研究。针对少量样本的视觉特征无法捕获类别代表性特征的问题，提出了一种基于语义-视觉多空间关系建模的少样本特征适配算法：语义-视觉多空间映射适配（Semantic-Visual Multi-Space Mapping Adapter，简称 SVMSMA）模型，旨在引入语义信息对视觉信息进行补充，丰富样本特征的信息来源以提升其多样性与代表性。SVMSMA 使用语义-视觉多空间映射网络将语义特征映射到视觉空间，并通过跨模态分类模块对单/多模态映射特征执行分类任务使其与视觉特征建立联系，以及跨模态特征对齐模块将映射特征与视觉特征原型进行对齐以获得更接近类别原型的特征。通过对语义-视觉多空间关系进行建

模，SVMSMA 丰富了样本特征的信息来源，提升了模型的泛化能力。在四个基准数据集上的实验证明了 SVMSMA 方法能够有效利用语义信息，在 MGSRCL 的基础上进一步提升少样本分类结果。

本文使用 MGSRCL 模型和 SVMSMA 模型分别对数据间的多种样本关系以及多种空间映射关系进行了建模，有效利用了数据中的多元关系，通过多粒度样本关系建模提升了视觉特征提取网络的特征提取能力，通过语义-视觉多空间关系建模提升了模型的泛化能力，从而取得了较好的少样本分类结果。

5.2 未来展望

本文分析了少样本分类面临的挑战，以数据中的多元关系为切入点，从多粒度样本关系建模与语义-视觉多空间建模两个角度入手，提出了多粒度样本关系对比学习模型和语义-视觉多空间映射适配模型来解决少样本分类问题，并取得了一定成果。但仍存在一定不足，后续可从以下几方面进一步研究：

（1）本文提出的 MGSRCL 方法中产生变换样本时使用的多是一些弱数据增强方法，其对特征提取网络性能提升产生的作用较为有限。目前诸如 Mixup、CutMix、以及 AugMix 等强数据增强已被证明了能够提高模型泛化能力，但由于其会将不同图像融合形成一张新的图像，这使得图像类别不再是单一标签，无法应用于 MGSRCL。因此，后续工作可以探讨如何将强数据增强方法引入所提出的方法，或者对方法进行改进以提高其适用性。

（2）本文通过使用 CLIP 模型的文本编码器作为语义特征提取网络，引入语义信息对视觉信息进行补充并取得了优异结果。在将类别名称输入文本编码器时，使用了 CLIP 原论文中提出的提示文本。但使用的提示文本是固定的，并不一定能够让模型输出对少样本分类任务来说最优的语义特征。因此后续可进一步研究其他提示文本或者将提示文本换成可学习参数，以获取最优的语义特征。

（3）本文中特征提取网络使用卷积网络，并没有使用近年来在很多视觉任务上取得良好表现的 Transformer 模型。这是因为 Transformer 模型一般需要大量的数据才能得到一个具有强大特征提取能力的预训练模型，而在少样本分类任务中，仅有 tieredImageNet 数据集规模较大，因此如何将 Transformer 模型引入少样本分类任务并取得像在其他任务上超越卷积网络的效果也是后续研究方向之一。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[3] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.

[4] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2017: 2961-2969.

[5] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报, 2014, 37(6): 1225-1240.

[6] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2020, 32 (4): 1201-1227.

[7] 蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述[J]. 自动化学报, 2021, 47(6): 1232-1255.

[8] 青晨, 禹晶, 肖创柏, 等. 深度卷积神经网络图像语义分割研究进展[J]. 中国图象图形学报, 2020, 25: 1069-1090.

[9] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//International Conference on Machine Learning. 2017: 1126-1135.

[10] Lee K, Maji S, Ravichandran A, et al. Meta-learning with differentiable convex optimization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10657-10665.

[11] Rusu A A, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization[C]// International Conference on Learning Representations. 2018.

[12] 李凡长, 刘洋, 吴鹏翔, 等. 元学习研究综述[J]. 计算机学报, 2021, 44: 422-446.

[13] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[J]. Advances in Neural Information Processing Systems, 2017, 30.

[14] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[J]. Advances in Neural Information Processing Systems, 2016, 29.

[15] Zhang C, Cai Y, Lin G, et al. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12203-12213.

[16] 刘鑫, 周凯锐, 何玉琳, 等. 基于度量的小样本分类方法研究综述[J/OL]. 模式识别与人工智能, 2021, 34: 909-923. DOI: 10.16451/j.cnki.issn1003-6059.202110004.

[17] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning

[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 1199-1208.

[18] Chen Z, Fu Y, Wang Y X, et al. Image deformation meta-networks for one-shot learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8680-8689.

[19] Li K, Zhang Y, Li K, et al. Adversarial feature hallucination networks for few-shot learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13470-13479.

[20] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27.

[21] Tian Y, Wang Y, Krishnan D, et al. Rethinking few-shot image classification: A good embedding is all you need?[C]//European Conference on Computer Vision. 2020: 266-282.

[22] Chen W Y, Liu Y C, Kira Z, et al. A closer look at few-shot classification[C]//International Conference on Learning Representations. 2019.

[23] Dhillon G S, Chaudhari P, Ravichandran A, et al. A baseline for few-shot image classification [C]//International Conference on Learning Representations. 2019.

[24] Zhang Y, Huang S, Zhou F. Generally boosting few-shot learning with handcrafted features[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3143-3152.

[25] Rizve M N, Khan S, Khan F S, et al. Exploring complementary strengths of invariant and equivariant representations for few-shot learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10836-10846.

[26] Ma J, Xie H, Han G, et al. Partner-assisted learning for few-shot image classification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10573-10582.

[27] Ouali Y, Hudelot C, Tami M. Spatial contrastive learning for few-shot classification[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2021: 671-686.

[28] Yang Z, Wang J, Zhu Y. Few-shot classification with contrastive learning[C]//European Conference on Computer Vision. 2022: 293-309.

[29] Chen D, Chen Y, Li Y, et al. Self-supervised learning for few-shot image classification[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. 2021: 1745-1749.

[30] Xie J, Long F, Lv J, et al. Joint distribution matters: Deep brownian distance covariance for few-shot classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7972-7981.

[31] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26.

[32] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.

2014: 1532-1543.

[33] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Vol. 1. 2019: 4171−-4186.

[34] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021: 8748-8763.

[35] Chen Z, Fu Y, Zhang Y, et al. Multi-level semantic feature augmentation for one-shot learning [J]. IEEE Transactions on Image Processing, 2019, 28(9): 4594-4605.

[36] Zhang Y, Huang S, Peng X, et al. Semi-identical twins variational autoencoder for few-shot learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[37] Xing C, Rostamzadeh N, Oreshkin B, et al. Adaptive cross-modal few-shot learning[J]. Advances in Neural Information Processing Systems, 2019, 32.

[38] Chen W, Si C, Zhang Z, et al. Semantic prompt for few-shot image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 23581-23591.

[39] Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification [C]//International Conference on Learning Representations. 2018.

[40] Bertinetto L, Henriques J F, Torr P H, et al. Meta-learning with differentiable closed-form solvers [C]//International Conference on Learning Representations. 2019.

[41] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[M]. California Institute of Technology, 2011.

[42] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International Conference on Machine Learning. 2020: 1597-1607.

[43] 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: 9729-9738.

[44] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[45] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [A]. 2013.

[46] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3982-3992.

[47] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[A]. 2019.

[48] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[C]//International Conference on Learning Representations. 2019.

[49] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2009: 248-255.

[50] Ravi S, Larochelle H. Optimization as a model for few-shot learning[C]//International Conference on Learning Representations. 2017.

[51] Triantafillou E, Zemel R, Urtasun R. Few-shot learning through an information retrieval lens[J]. Advances in Neural Information Processing Systems, 2017, 30.

[52] Endres D M, Schindelin J E. A new metric for probability distributions[J]. IEEE Transactions on Information Theory, 2003, 49(7): 1858-1860.

[53] Fuglede B, Topsoe F. Jensen-shannon divergence and hilbert space embedding[C]//IEEE International Symposium on Information Theory. 2004: 31.

附 录

A. 作者在攻读硕士学位期间的论文目录

[1] *** , *** , He T, et al. Mirrored EAST: An Efficient Detector for Automatic Vehicle Identification Number Detection in the Wild[J]. IEEE Transactions on Industrial Informatics, 2023. （中科院 SCI 一区）

[2] *** , Wang Y, Zhang Y, et al. Adversarial Bidirectional Feature Generation for Generalized Zero-Shot Learning Under Unreliable Semantics[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer Nature Switzerland, 2022: 639-654.（CCF-C）

[3] *** , Huangfu L, *** , et al. Rethinking the Sample Relations for Few-Shot Classification[J]. Image and Vision Computing. （中科院 SCI 三区，返修中）

B. 作者在攻读硕士学位期间参与的科研项目

[1] 国家自然科学基金面上项目，少样本学习特征生成与鲁棒性关键技术研究

[2] 重庆市自然科学基金面上项目，文本描述协同的双向生成式少样本学习研究

C. 学位论文数据集

<table>
<tr><td colspan="2">关键词</td><td>密级</td><td colspan="2">中图分类号</td></tr>
<tr><td colspan="2">少样本分类; 关系建模;
对比学习; 语义信息表示</td><td>公开</td><td colspan="2">TP</td></tr>
<tr><td>学位授予单位名称</td><td>学位授予单位代码</td><td colspan="2">学位类别</td><td>学位级别</td></tr>
<tr><td>***</td><td>***</td><td colspan="2">学术学位</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名</td><td>并列题名</td><td colspan="2">论文语种</td></tr>
<tr><td colspan="2">基于多元关系建模的少样
本分类算法研究</td><td>/</td><td colspan="2">汉语</td></tr>
<tr><td>作者姓名</td><td>***</td><td colspan="2">学号</td><td>***</td></tr>
<tr><td colspan="2">培养单位名称</td><td colspan="3">培养单位代码</td></tr>
<tr><td colspan="2">***</td><td colspan="3">***</td></tr>
<tr><td>学科专业</td><td>研究方向</td><td colspan="2">学制</td><td>学位授予年</td></tr>
<tr><td>软件工程</td><td>计算机视觉</td><td colspan="2">3 年</td><td>***</td></tr>
<tr><td>论文提交日期</td><td>***</td><td colspan="2">论文总页数</td><td>31</td></tr>
<tr><td>导师姓名</td><td>***</td><td colspan="2">职称</td><td>教授</td></tr>
<tr><td colspan="2">答辩委员会主席</td><td colspan="3">A</td></tr>
<tr><td colspan="5">电子版论文提交格式
文本（✓）图像 () 视频 () 音频 () 多媒体 () 其他 ()</td></tr>
</table>