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

Research on Few-Shot Classification Algorithms Based on Multivariate Relationship Modeling

A Thesis Submitted to Chongqing University

in partial fulfillment of the requirement

for the degree of

Master of Science

in

Software Engineering

by

XXX

Supervisor: Prof. XXX

June, 2024

摘 要

近年来，深度学习技术已在诸多图像分类任务上取得了显著成就，但这些任务的成功往往依赖于海量的标注数据，在数据匮乏的情况下，很多深度学习模型便无法精确识别物体类别。为了提升深度学习模型在数据匮乏情况下的效果，旨在模拟人类识别物体过程的少样本分类（Few-Shot Classification，简称 FSC）被提出并取得了一定进展。在少样本分类任务中，如何将模型在大量数据上学习到的知识迁移到新的类别是解决问题的关键。虽然基类数据与新类数据的类别不同，但其数据间的多种关系是具有相似性和关联性的。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，从而将在基类上学习到的知识迁移至新类。基于此，本文以数据的多元关系为切入点，对多粒度样本关系和语义-视觉多空间关系进行建模，以推动少样本分类问题的研究进展。本文主要工作如下：

（1）针对少样本分类模型特征提取能力不足的问题，本文开展基于多粒度样本关系建模的少样本分类研究，提出了多粒度样本关系对比学习（Multi-Grained Sample Relation Contrastive Learning，简称 MGSRCL）模型。该模型将样本关系划分为三种类型：样本内关系、类内关系和类间关系，并使用变换一致性学习约束样本内关系，使用类对比学习约束类内关系与类间关系，对多种粒度的样本关系进行充分挖掘与细致建模。在多个基准数据集上的大量实验证明，MGSRCL 方法通过建模多粒度样本关系提升了模型的特征提取能力与泛化能力，有效提高了少样本分类准确率，并为其他两阶段方法提供了一个优质的预训练模型。

（2）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，本文开展基于语义-视觉多空间关系建模的少样本分类研究，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。该模型引入语义信息作为视觉信息的补充，通过语义-视觉多空间映射网络将语义特征映射到视觉空间，并使用简单有效的跨模态分类和跨模态特征对齐策略对语义特征与视觉特征进行建模。大量实验证明，SVMSMA 模型能够有效建立语义信息与视觉信息的联系，丰富了样本特征的信息来源，从而能够利用语义信息提升样本特征的多样性，增强模型对新类别的适应能力和泛化能力，并在 MGSRCL 的基础上进一步提升了少样本分类准确率。

关键词：少样本分类；关系建模；对比学习；语义信息表示

Abstract

In recent years, deep learning technology has made remarkable achievements in image classification tasks. But the success of these tasks often depends on vast amounts of labeled data, and in the absence of data, many deep learning models are unable to accurately identify object categories. In order to improve the effectiveness of deep learning models in the case of data scarcity, Few-Shot Classification (FSC), which aims to simulate the object recognition process of human, has been proposed and made some progress. In FSC, how to transfer the knowledge learned by the model on a large amount of data to new categories is the key to solve the problem. Although the categories of the base classes and the novel classes are different, the relationship between the data is similar and relevant. By modeling these relationships on the base classes, the internal relationship between the data can be better understood and mined, so that the knowledge learned on the base classes can be transferred to the novel classes. Based on this, this paper takes the multivariate relationship of data as the starting point to model the multi-grained sample relation and semantic-visual multi-space relation, so as to promote the research progress of FSC. The main works of this paper are as follows:

(1) Aiming at the lack of feature extraction ability of FSC model, this paper carries out a study of FSC based on multi-grained sample relationship modeling, and proposes a Multi-Grained Sample Relation Contrastive Learning (MGSRCL) model. The model divides the sample relations into three types: intra-sample relation, intra-class relation and inter-class relation, and uses transformation consistency learning to constrain intra-sample relation, uses class contrastive learning to constrain intra-class relation and inter-class relation, so as to fully mine and carefully model the sample relations of various granularity. Extensive experiments on multiple benchmark datasets prove that MGSRCL can improve the feature extraction ability and generalization ability of the model by modeling multi-grained sample relations, effectively improve the accuracy of FSC, and provide an excellent pre-training model for other two-stage methods.

(2) In view of the problem that the category representative features cannot be captured only based on the visual features of a small number of samples, this paper carries out a research on FSC based on semantic-visual multi-space relationship modeling, and proposes a Semantic-Visual Multi-Space Mapping Adapter (SVMSMA) model. The model introduces semantic information as a supplement to visual information, maps semantic

features to visual space through semantic-visual multi-space mapping network, and uses simple and effective cross-modal classification and cross-modal feature alignment strategies to model semantic features and visual features. Extensive experiments have proved that SVMSMA model can effectively establish the connection between semantic information and visual information, enrich the information sources of sample features, and thus improve the diversity of sample features by using semantic information, enhance the model's adaptability and generalization ability to new categories, and further improve the accuracy of FSC on the basis of MGSRCL.

Key words: Few-Shot Classification; Relationship Modeling; Contrastive Learning; Semantic Information Representation

目 录

图目录

表目录

1 绪 论

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 研究背景及意义

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[1–6]，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 国内外研究现状与挑战

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[7–10]，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[7] 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[8] 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[9] 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[11–14] 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[11] 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[13] 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[15] 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[16,17]。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[16] 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[17] 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）[18] 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[19–28]。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[19] 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[20,21] 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[22] 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[23–28] 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[29]、GloVe[30]、BERT[31] 等自然语言模型或者 CLIP[32] 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[33] 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[34] 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[35] 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[36] 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 本文研究内容与创新点

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[12]、tierdImageNet[37]、CIFAR-FS[38] 和 CUB-200-2011[39] 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。

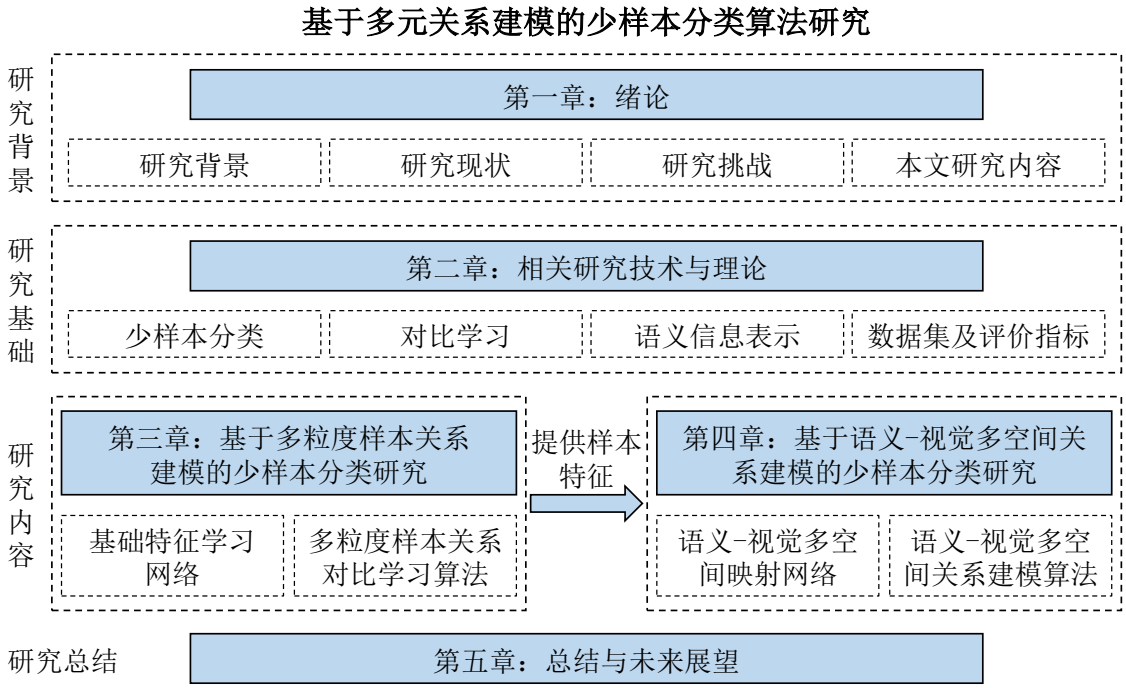


图 1.1 本文组织结构图。

Fig. 1.1 The organizational structure of the paper.

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，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 相关研究技术与理论

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[40] 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[41] 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[42]

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[42]）。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。

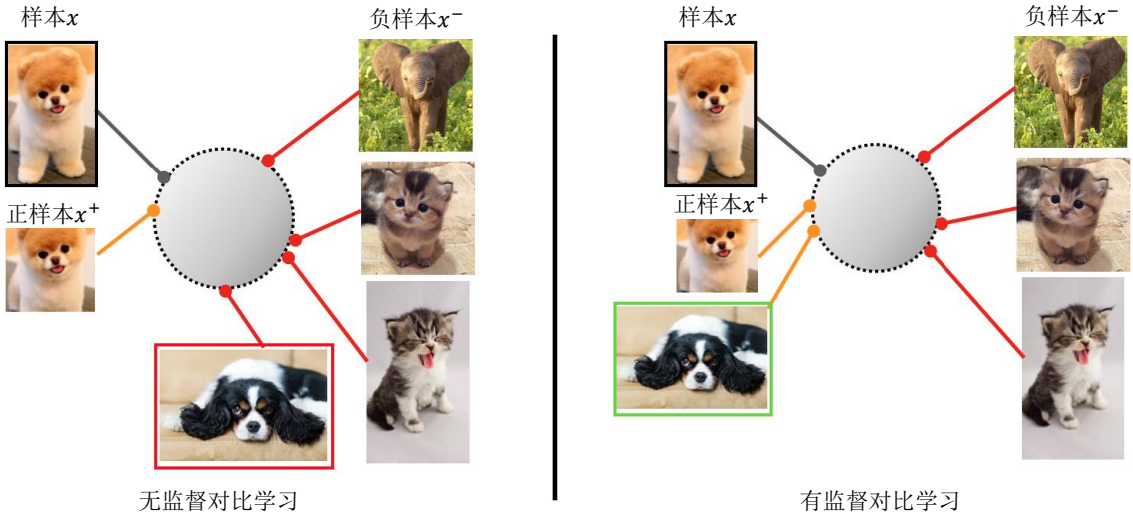


图 2.2 无监督对比学习与有监督对比学习。

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[29,43] 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[29]）。

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。

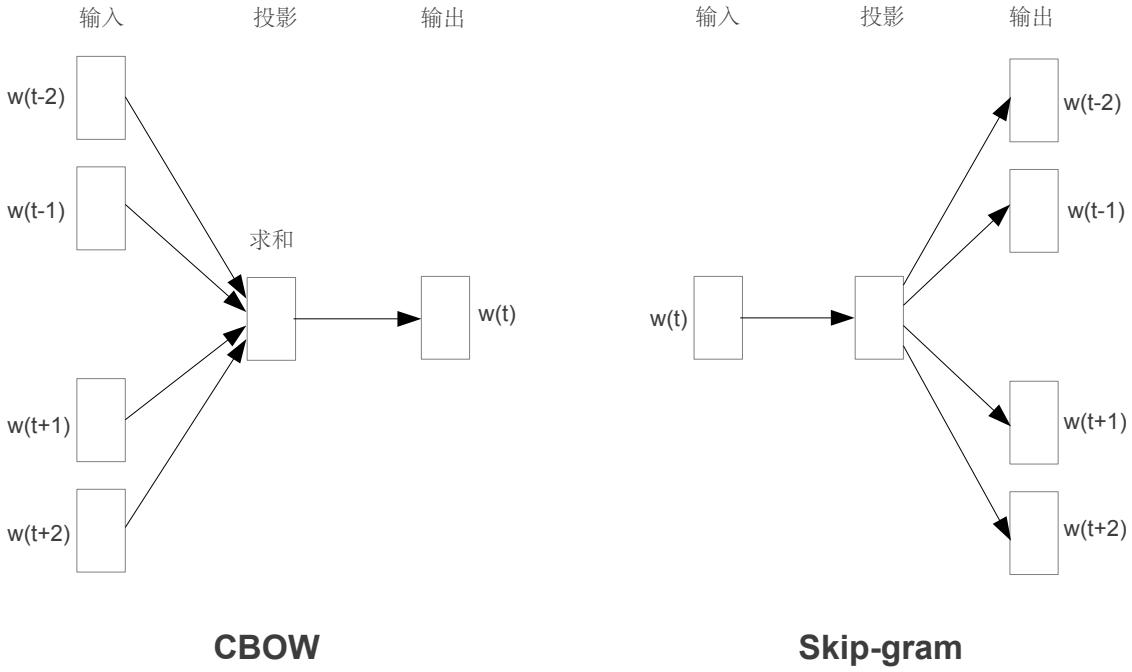


图 2.3 CBOW 模型与 Skip-Gram 模型示意图。

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[30] 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）。

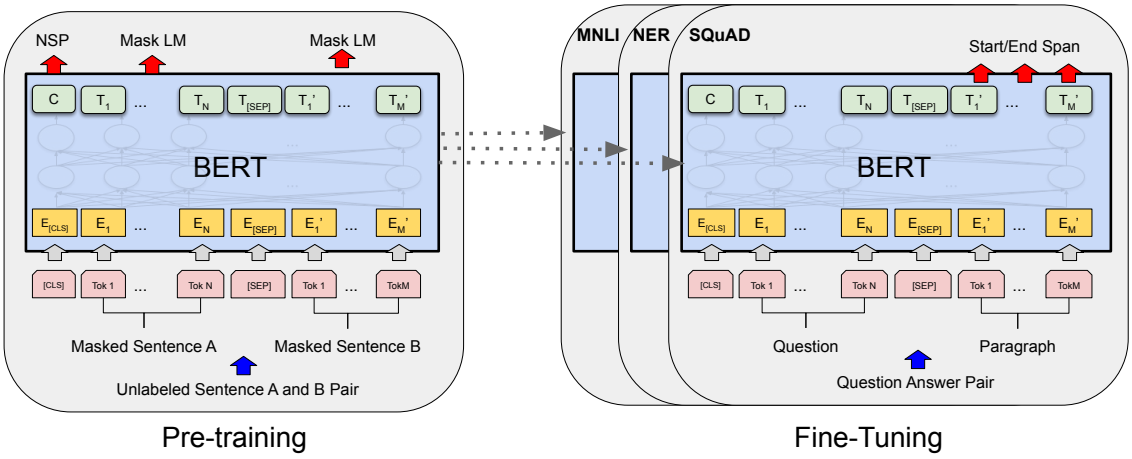


图 2.4 BERT 的整体预训练和微调过程。

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

2.3.3 BERT

Devlin 等人[31] 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[31]），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[44]、RoBERTa[45]、ALBERT[46] 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。

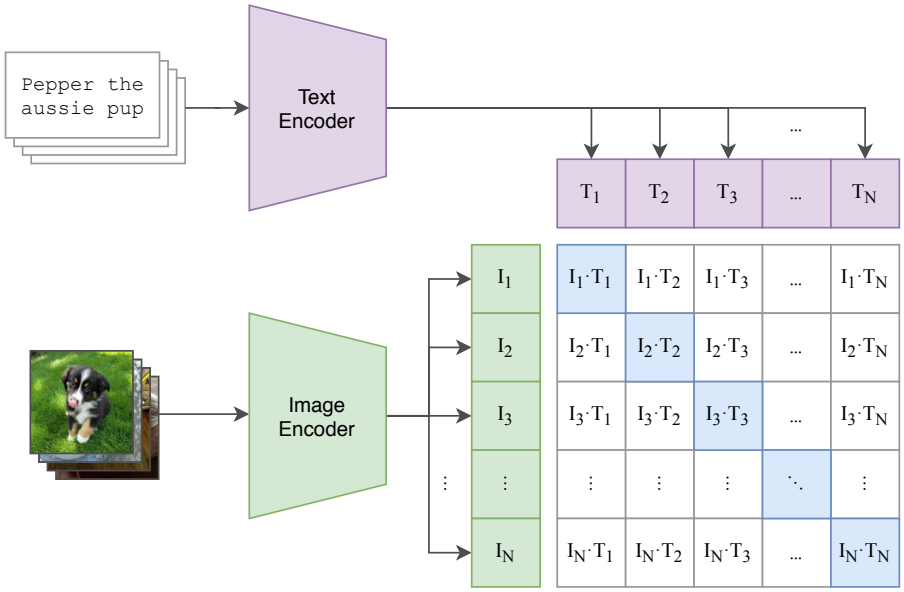


图 2.5 CLIP 模型预训练示意图。

Fig. 2.5 Illustration of pre-training CLIP model.

2.3.4 CLIP

近年来，Radford 等人[32] 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[32]），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[12]、tieredImageNet[37]、CIFAR-FS[38]，以及一个细粒度数据集：CUB-200-2011（CUB）[39]，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[12] 和 tieredImageNet 数据集[37] 均为 ImageNet[47] 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[48] 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[38] 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[39] 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[49] 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[19,23] 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

少样本分类旨在模拟人类识别物体的过程，这一目标使其受到了广泛关注，并发展出多种方法。其中基于元学习的方法[7,8,48]是主流少样本分类算法之一，这些方法在训练阶段模拟少样本分类任务，并尝试训练一个基础模型，使其能够迅速适应新任务。基于度量的方法[11–13,50]旨在设计度量函数来计算样本之间的距离或相似度，以能够通过比较样本间距离在少量样本情况下也能有效分类。此外，基于增加额外样本以缓解数据匮乏问题的直觉，许多基于数据增强的方法[16,17,33]被提出，通过合成额外样本来增加样本多样性以提高少样本分类性能。然而，这些方法通常涉及复杂的训练阶段，或者在测试阶段需要添加许多额外样本，这带来了较高的计算成本。

近期研究[19–21]显示，在整个基类数据集上对模型使用分类任务进行完全监督形式的预训练，然后在元测试阶段将模型特征提取网络参数冻结并用来提取图像特征，最后使用提取的特征对每个少样本分类任务训练分类器并进行预测，可以实现与上述复杂少样本方法相媲美的性能。这些工作的成功揭示了特征学习在少样本分类中的重要性。为了获得更好的特征提取网络，众多研究者聚焦于少样本分类的特征学习阶段，并提出了一系列令人印象深刻的工作[19,22–25,51–53]。其中，很多方法采用对比学习作为辅助任务取得了很好的结果[23–25]，这是因为对比学习可以缓解网络仅通过交叉熵损失学习基类的最具区分性特征，而忽视获取某些次级区分性特征的问题，通过对样本关系建模提高了网络的特征提取能力以及在新类数据集上的泛化性。例如，IER[23]利用无监督对比损失来约束图像在不同变换下的不变性。PAL[24]使用有监督对比学习[42]对教师模型进行初步训练，随后利用教师模型为学生模型提供软标签。

上述采用对比学习作为辅助任务的方法取得了良好的性能，但它们直接使用无监督或有监督对比学习方法来增强网络的特征提取能力，这可能没有充分挖掘利用样本关系的潜力。无监督对比学习[40,41]将同一样本的不同变换版本（使用不同的数据增强来获得）视为正样本对，将不同样本视为负样本对，不考虑它们的类

别标签。尽管这种方法有效地提升了网络对于不同变换不变性的学习能力，并增加了不同类别样本之间的区分度，但它也无意中将属于同一类别的多个样本在特征空间推得更远，这在一定程度上对特征学习是不利的。有监督对比学习[42]通过确保相同类别的样本比不同类别的样本在特征空间具有更紧密的距离来克服上述问题。但在建模样本关系时，它将样本的变换版本和其他同类样本同等对待，这种策略并不合适，因为样本与其变换版本在语义内容上几乎完全一致，而只与其同类样本共享相似的语义内容。换句话说，在学习的特征空间中，样本应与其变换版本比与同类样本更接近。

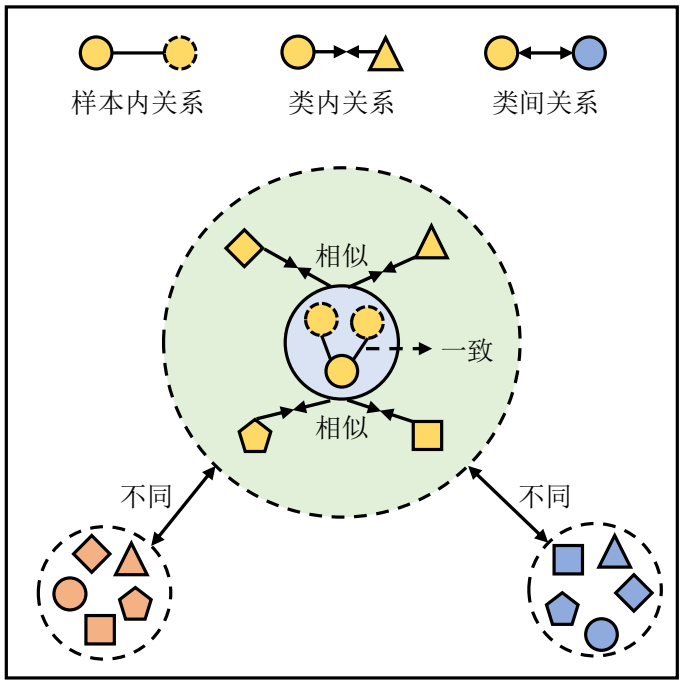


图 3.1 样本关系示意图。在该图中，不同形状与颜色分别代表不同样本与类别。同一样本的不同变换由相同的颜色和形状表示。样本关系包括三种类型: 样本内关系、类内关系和类间关系。本章所提方法约束同一样本不同变换版本在语义内容上保持一致，同类样本保持相似，非同类样本保持不同。

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

为了解决上述问题，本文重新审视了多种样本关系，并提出了一个针对少样本分类的多粒度样本关系对比学习（Multi-Grained Sample Relation Contrastive Learning, 简称 MGSRCL）方法。MGSRCL 将样本关系分为三种不同粒度的类型：同一样

本不同变换版本的样本内关系（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），这三种样本关系揭示了数据在不同层次的内在结构和差异，如图3.1所示。

针对这三种不同粒度的样本关系，MGSRCL 提出了两个模块对其进行约束。首先对于同一样本不同变换间的样本内关系，本文通过变换一致性学习（Transformation Consistency Learning，简称 TCL）策略进行约束。这一策略的灵感来源于这样一个认识：同一样本在不同但轻微的变换下应保持其本质的语义不变性。TCL 通过对齐样本及其变换版本的预测标签分布来确保其在标签输出上的一致性，从而保证样本和其不同变换版本在特征空间保持高度的语义一致性。第二种样本关系是同类样本的类内关系，由于同类样本的语义内容不像同一样本不同变换那样高度一致，忽视同类样本之间的语义差异并将它们映射到特征空间中的同一位置会因为简化了模型的学习过程而导致模型崩塌。因此，本文采用类对比学习（Class Contrastive Learning，简称 CCL）来以一种相对的形式约束这种样本关系以及第三种样本关系，即不同类样本的类间关系。CCL 不追求同类样本在特征空间中的绝对一致性，而着重于以一种相对距离的方式在特征空间中增强同类样本间的内聚程度，同时增大不同类别间的分离程度，从而提高了模型对不同类别样本的区分能力。

以验证本章方法的有效性，本章在四个基准数据集上进行了广泛的实验，包括三个普通少样本分类数据集：miniImageNet[12]、tieredImageNet[37]、CIFAR-FS[38]，以及一个细粒度少样本分类数据集：CUB-200-2011[39]。实验结果表明，本章方法取得了优异的结果，并且可以作为预训练模型提升其他两阶段少样本分类方法的性能。

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。

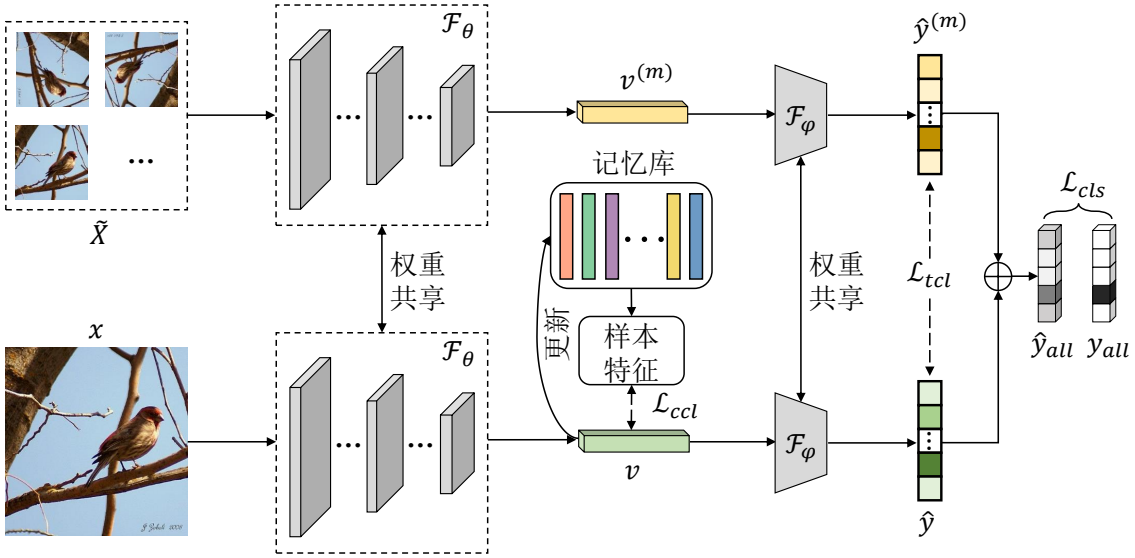


图 3.2 多粒度样本关系对比学习模型（MGSRCL）示意图。它包含一个特征提取网络 $\mathcal{F}_\theta$ 和一个分类器 $\mathcal{F}_\varphi$。在此图中，v 和 $v^{(m)}$ 代表原始图像 x 及其第 m 个变换版本 $x^{(m)}$ 的特征，其中 $x^{(m)} \in \widehat{X}$。$\bigoplus$ 是一个连接操作，用于对原始图像的预测输出 $\widehat{y}$ 与 M 个变换的预测输出 $\widehat{y}^{(1)}, ..., \widehat{y}^{(m)}, ..., \widehat{y}^{(M)}$ 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.

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[23,24,27] 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——亦即其语义内容。

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[54,55] 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[40,42] 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[23] 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。

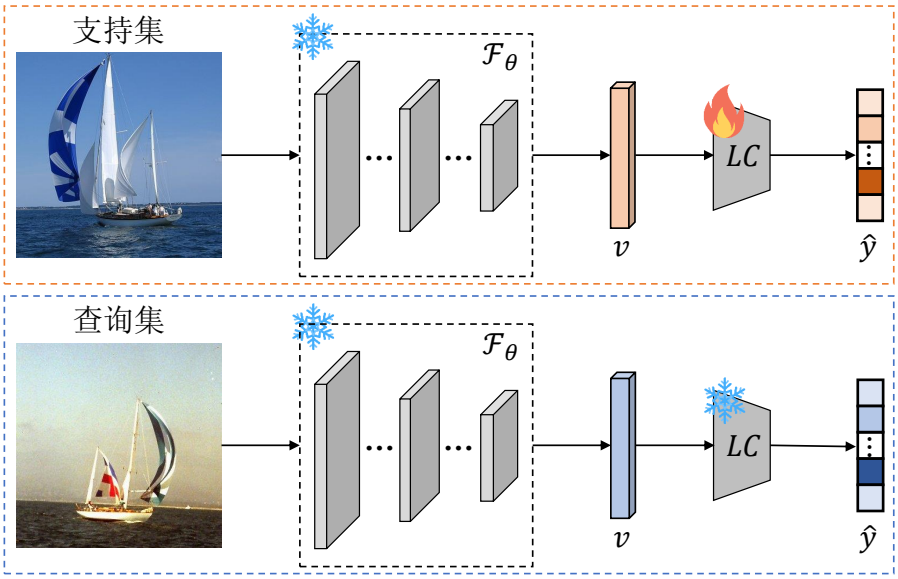


图 3.3 MGSRCL 模型推理过程示意图。推理过程中，使用冻结参数的特征提取网络 $\mathcal{F}_\theta$ 提取支持集与查询集的图像特征。其中，支持集特征被用来训练一个逻辑回归分类器 LC，查询集特征则是用来测试分类器性能。

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.

3.3 实验设置及结果分析

在本节中，首先介绍了本章方法的实验设置，包括数据集设置、网络结构、优化设置、数据增强方式，然后分析了基于多粒度样本关系对比学习的少样本特征学习算法实验结果，接下来对模型的多个模块以及超参数进行了消融实验和分析，最后对模型所提取特征进行了可视化分析。

3.3.1 实验设置

（1）实验数据集介绍

本文在四个常用的少样本分类基准数据集上进行了实验，其中包括三个普通少样本分类数据集：miniImageNet[12]、tieredImageNet[37]、CIFAR-FS[38]，以及一个细粒度数据集：CUB-200-2011（CUB）[39]。对于所有数据集，本文保持了与其他工作[19,23,56]相同的划分准则。在实验中，对于 miniImageNet、tieredImageNet 和 CUB，图像大小为 84×84，而对于 CIFAR-FS，图像大小为 32×32。

（2）网络结构

遵循先前的研究工作[13,19,23]，本文采用了 ResNet-12 作为特征提取网络。变换一致性学习（TCL）在标签输出上进行，而类对比学习（CCL）在经过全局池化层后的特征上进行，这些技术不需要额外的网络层。此外，本文添加了一个自监督学习模块，该模块是一个由两个全连接层、一个批归一化层和一个激活函数组成的多层感知机。

（3）优化设置

在所有实验中，本文使用了具有 0.9 动量和 $5e^{-4}$ 权重衰减的随机梯度下降（Stochastic Gradient Descent，简称 SGD）优化器。初始学习率设定为 0.05，随后在特定轮次以 0.1 的倍数衰减。对于 tieredImageNet，总共训练了 60 个轮次，学习率分别在第 30、40 和 50 轮次后衰减。对于其他数据集，训练了 80 个轮次，学习率在第 60 和 70 轮次后衰减。关于超参数，本文对所有数据集设置为以下值：$\alpha = 1.0$，$\beta = 0.1$，$\tau_1 = 4.0$，$\tau_2 = 0.1$。

（4）数据增强

为了缓解过拟合问题并进行 MGSRCL 模型中的变换一致性学习（TCL），本文在训练特征提取网络时添加了一些数据增强样本。数据增强包括三种长宽比例不同的裁剪缩放变换、三种旋转变换（分别为 90°、180° 和 270°）、一种随机擦除、一种图像灰度化和一种 Sobel 边缘检测。在训练过程中，这些数据增强样本与原始样本一同作为模型的输入。

3.3.2 基准数据集实验结果

为了评估方法的有效性，本文在四个数据集上进行了广泛的实验。表3.1、3.2、3.3和3.4分别展示了 miniImageNet、tieredImageNet、CIFAR-FS 和 CUB 数据集上一些现有少样本分类方法和本文方法的实验结果。此外，本文还将所提模型作为预训练网络结合到三个两阶段少样本分类算法中，并与原方法性能进行对比以进一步验证本文所获得特征提取网络的有效性。

（1）普通少样本分类

如表3.1和3.2所示，本文方法 MGSRCL 在 miniImageNet 和 tieredImageNet 数据集上与其他方法相比取得了出色的性能。具体而言，MGSRCL 在 miniImageNet 上的

1-shot 和 5-shot 任务中分别达到了 69.57% 和 84.41% 的准确率，在 tieredImageNet 上分别达到了 72.98% 和 86.23% 的准确率。特别是在 5-way 1-shot 任务中，MGSRCL 达到了最优实验结果，在两个数据集上分别比次优结果高 0.20% 和 0.30%。在 CIFAR-FS 数据集上，MGSRCL 在 1-shot 和 5-shot 少样本分类任务中分别达到了 78.54% 和 88.64% 的准确率，在 1-shot 任务上比次优的 PAL 方法高 1.44%，如表3.3所示。值得注意的是，与采用知识蒸馏技术的 RFS[19]、PAL[24] 和 SCL[25]，以及使用元学习方法的 DeepEMD[13] 和 ESPT[53] 不同，MGSRCL 无需进行第二次训练或元学习微调阶段。本文方法使得预训练模型能够达到与最先进方法相媲美甚至超越的性能，这证明了所提方法能够增强模型的特征提取能力和泛化能力，从而使得所提取到的新类特征具有更好的判别性。

表 3.1 MGSRCL 在 miniImageNet 数据集上的分类准确率（%）。最优结果用粗体表示，带有“†”标记的方法表示结果是使用作者提供代码所实现。

Table 3.1 Classification accuracy (%) of MGSRCL on miniImageNet. The best results are shown in bold, and methods with the “†” indicate that the result was implemented using author-supplied code.

方法	特征提取网络	5-way 1-shot	5-way 5-shot
MAML[7]	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92
ProtoNet[11]	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66
DeepEMD[13]	ResNet-12	65.91 ± 0.82	82.41 ± 0.56
RFS-distill[19]	ResNet-12	64.82 ± 0.60	82.14 ± 0.43
AssoAlign[57]	ResNet-18	59.88 ± 0.67	80.35 ± 0.73
GIFSL[58]	ResNet-12	65.47 ± 0.63	82.75 ± 0.42
MELR[59]	ResNet-12	67.40 ± 0.43	83.40 ± 0.28
IEPT[52]	ResNet-12	67.05 ± 0.44	82.90 ± 0.30
IER[23]	ResNet-12	66.82 ± 0.80	84.35 ± 0.51
RENet[56]	ResNet-12	67.60 ± 0.44	82.58 ± 0.30
PAL[24]	ResNet-12	69.37 ± 0.64	84.40 ± 0.44
HandCrafted[22]	ResNet-12	67.14 ± 0.76	83.11 ± 0.69
PDA[60]	ResNet-12	65.75 ± 0.43	83.37 ± 0.30
SCL-distill[25]	ResNet-12	67.40 ± 0.76	83.19 ± 0.54
HGNN[61]	ResNet-12	67.02 ± 0.20	83.00 ± 0.13
APP2S[62]	ResNet-18	64.82 ± 0.12	81.31 ± 0.22
DGAP[63]	ResNet-12	61.35 ± 0.62	78.85 ± 0.46
ESPT[53]	ResNet-12	68.36 ± 0.19	84.11 ± 0.12
Meta-HP[64]	ResNet-12	62.49 ± 0.80	77.12 ± 0.62
SAPENet[65]	ResNet-12	66.41 ± 0.20	82.76 ± 0.14
FEAT+DFR[66]	ResNet-12	67.74 ± 0.86	82.49 ± 0.57
DiffKendall[67]	ResNet-12	65.56 ± 0.43	80.79 ± 0.31
FEAT[68]	ResNet-12	66.78 ± 0.20	82.05 ± 0.14

表 3.1 （续）
Table 3.1 (continued)

方法	特征提取网络	5-way 1-shot	5-way 5-shot
MGSRCL + FEAT	ResNet-12	69.27 ± 0.21	83.59 ± 0.13
Meta-Baseline† [69]	ResNet-12	63.38 ± 0.23	79.48 ± 0.16
MGSRCL + Meta-Baseline	ResNet-12	69.01 ± 0.23	83.94 ± 0.15
STVAE [34]	ResNet-12	63.62 ± 0.80	80.68 ± 0.48
MGSRCL + STVAE	ResNet-12	67.29 ± 0.89	82.62 ± 0.58
MGSRCL	ResNet-12	**69.57 ± 0.45**	**84.41 ± 0.30**

表 3.2 MGSRCL 在 tieredImageNet 数据集上的分类准确率（%）。最优结果用粗体表示，带有“†”标记的方法表示结果是使用作者提供代码所实现。

Table 3.2 Classification accuracy (%) of MGSRCL on tieredImageNet. The best results are shown in bold, and methods with the “†” indicate that the result was implemented using author-supplied code.

方法	特征提取网络	5-way 1-shot	5-way 5-shot
MAML [7]	32-32-32-32	51.67 ± 1.81	70.30 ± 1.75
ProtoNet [11]	64-64-64-64	53.31 ± 0.89	72.69 ± 0.74
DeepEMD [13]	ResNet-12	71.16 ± 0.87	86.03 ± 0.58
RFS-distill [19]	ResNet-12	71.52 ± 0.69	86.03 ± 0.49
AssoAlign [57]	ResNet-18	69.29 ± 0.56	85.97 ± 0.49
GIFSL [58]	ResNet-12	72.39 ± 0.66	86.91 ± 0.44
MELR [59]	ResNet-12	72.14 ± 0.51	87.01 ± 0.35
IEPT [52]	ResNet-12	72.24 ± 0.50	86.73 ± 0.34
IER [23]	ResNet-12	71.87 ± 0.89	86.82 ± 0.58
RENet [56]	ResNet-12	71.16 ± 0.51	85.28 ± 0.35
PAL [24]	ResNet-12	72.25 ± 0.72	86.95 ± 0.47
PDA [60]	ResNet-12	72.28 ± 0.49	86.70 ± 0.33
SCL-distill [25]	ResNet-12	71.98 ± 0.91	86.19 ± 0.59
HGNN [61]	ResNet-12	72.05 ± 0.23	86.49 ± 0.15
APP2S [62]	ResNet-18	70.83 ± 0.15	84.15 ± 0.29
DGAP [63]	ResNet-12	70.10 ± 0.67	84.99 ± 0.46
ESPT [53]	ResNet-12	72.68 ± 0.22	**87.49 ± 0.14**
Meta-HP [64]	ResNet-12	68.26 ± 0.72	82.91 ± 0.36
SAPENet [65]	ResNet-12	68.63 ± 0.23	84.30 ± 0.16
FEAT+DFR [66]	ResNet-12	71.31 ± 0.93	85.12 ± 0.64
DiffKendall [67]	ResNet-12	70.76 ± 0.43	85.31 ± 0.34
FEAT [68]	ResNet-12	70.80 ± 0.23	84.79 ± 0.16
MGSRCL + FEAT	ResNet-12	72.02 ± 0.23	86.19 ± 0.15

表 3.2 （续）
Table 3.2 (continued)

方法	特征提取网络	5-way 1-shot	5-way 5-shot
Meta-Baseline†[69]	ResNet-12	68.74 ± 0.26	83.45 ± 0.18
MGSRCL + Meta-Baseline	ResNet-12	69.79 ± 0.26	83.55 ± 0.18
STVAE†[34]	ResNet-12	68.32 ± 0.94	83.79 ± 0.66
MGSRCL + STVAE	ResNet-12	72.03 ± 0.89	84.49 ± 0.66
MGSRCL	ResNet-12	**72.98 ± 0.51**	86.23 ± 0.34

表 3.3 MGSRCL 在 CIFAR-FS 数据集上的分类准确率（%）。最优结果用粗体表示，带有“†”标记的方法表示结果是使用作者提供代码所实现。

Table 3.3 Classification accuracy (%) of MGSRCL on CIFAR-FS. The best results are shown in bold, and methods with the “†” indicate that the result was implemented using author-supplied code.

方法	特征提取网络	5-way 1-shot	5-way 5-shot
MAML[7]	32-32-32-32	58.90 ± 1.90	71.50 ± 1.00
ProtoNet[11]	64-64-64-64	55.50 ± 0.70	72.00 ± 0.60
RFS-distill[19]	ResNet-12	73.90 ± 0.80	86.90 ± 0.50
GIFSL[58]	ResNet-12	74.58 ± 0.38	87.68 ± 0.23
IER[23]	ResNet-12	76.83 ± 0.82	89.26 ± 0.58
RENet[56]	ResNet-12	74.51 ± 0.46	86.60 ± 0.32
PAL[24]	ResNet-12	77.10 ± 0.70	88.00 ± 0.50
HandCrafted[22]	ResNet-12	76.68 ± 0.59	87.49 ± 0.73
SCL-distill[25]	ResNet-12	76.50 ± 0.90	88.00 ± 0.60
ConstellationNet[70]	ResNet-12	75.40 ± 0.20	86.80 ± 0.20
APP2S[62]	ResNet-18	73.12 ± 0.22	85.69 ± 0.16
Meta-HP[64]	ResNet-12	73.74 ± 0.57	86.37 ± 0.32
FEAT†[68]	ResNet-12	75.97 ± 0.21	87.34 ± 0.14
MGSRCL + FEAT	ResNet-12	79.91 ± 0.21	**90.18 ± 0.14**
Meta-Baseline†[69]	ResNet-12	74.56 ± 0.39	86.24 ± 0.27
MGSRCL + Meta-Baseline	ResNet-12	78.51 ± 0.24	88.60 ± 0.16
STVAE[34]	ResNet-12	76.30 ± 0.60	87.00 ± 0.40
MGSRCL + STVAE	ResNet-12	**80.92 ± 0.72**	86.38 ± 0.60
MGSRCL	ResNet-12	78.54 ± 0.47	88.64 ± 0.32

（2）细粒度少样本分类

为了进一步验证所提方法的泛化能力，本文还在一个细粒度少样本分类数据集（CUB）上对所提的 MGSRCL 方法进行了实验，实验结果如表3.4所示。本文方

法在 1-shot 和 5-shot 任务中都取得了最优结果，分别达到了 86.14% 和 94.75% 的分类准确率，比次优结果高出 0.69% 和 0.02%。这些结果表明，在不同类别之间差异微小的细粒度数据集上，本文方法通过探索不同粒度的样本关系并对它们进行细致建模，能够更好地区分细粒度类别，进一步证明了方法的有效性。

表 3.4 MGSRCL 在 CUB 数据集上的分类准确率（%）。最优结果用粗体表示，带有“†”标记的方法表示结果是使用作者提供代码所实现。

Table 3.4 Classification accuracy (%) of MGSRCL on CUB. The best results are shown in bold, and methods with the “†” indicate that the result was implemented using author-supplied code.

方法	特征提取网络	5-way 1-shot	5-way 5-shot
FEAT[68]	64-64-64-64	68.87 ± 0.22	82.90 ± 0.15
DeepEMD[13]	ResNet-12	75.65 ± 0.83	88.69 ± 0.50
AssoAlign[57]	ResNet-18	74.22 ± 1.09	88.65 ± 0.55
MELR[59]	64-64-64-64	70.26 ± 0.50	85.01 ± 0.32
IEPT[52]	64-64-64-64	69.97 ± 0.49	84.33 ± 0.33
RENet[56]	ResNet-12	79.49 ± 0.44	91.11 ± 0.24
HGNN[61]	ResNet-12	78.58 ± 0.20	90.02 ± 0.12
APP2S[62]	ResNet-12	77.64 ± 0.19	90.43 ± 0.18
ESPT[53]	ResNet-12	85.45 ± 0.18	94.02 ± 0.09
SAPENet[65]	64-64-64-64	70.38 ± 0.23	84.47 ± 0.14
FEAT+DFR[66]	ResNet-12	77.14 ± 0.21	88.97 ± 0.13
Bi-FRN[71]	ResNet-12	85.44 ± 0.18	94.73 ± 0.09
FEAT†[68]	ResNet-12	77.60 ± 0.45	89.20 ± 0.28
MGSRCL + FEAT	ResNet-12	84.23 ± 0.19	92.67 ± 0.10
Meta-Baseline†[69]	ResNet-12	75.04 ± 0.24	87.57 ± 0.14
MGSRCL + Meta-Baseline	ResNet-12	**88.37 ± 0.18**	**95.52 ± 0.09**
STVAE[34]	ResNet-12	77.32 ± 0.00	86.84 ± 0.00
MGSRCL + STVAE	ResNet-12	84.35 ± 0.76	93.69 ± 0.39
MGSRCL	ResNet-12	86.14 ± 0.38	94.75 ± 0.19

（3）与其他方法结合

此外，作为一种基于特征学习的方法，本文方法可以为两阶段元学习方法和一些生成方法提供一个良好的预训练模型，帮助它们实现更好的性能。为了证明这一点，本文选择了两种元学习方法（FEAT[68]、Meta-Baseline[69]）和一种生成方法（STVAE[34]），在四个数据集上进行了实验。因为这些方法的特征提取网络与本文存在一些差异，或者它们没有在相应数据集上进行实验，本文对一些方法根据原作者提供代码进行了重新实现，实验结果以“†”进行标记。如表3.1和3.2所示，当使用本文的预训练模型时，FEAT、Meta-Baseline 和 STVAE 在 miniImageNet 数据

集的 1-shot 任务中的分类准确率分别提高了 2.49%、5.63% 和 3.67%，在 5-shot 任务中分别提高了 1.54%、4.46% 和 1.94%。FEAT、Meta-Baseline 和 STVAE 在使用本文模型作为预训练模型时，在 tieredImageNet 数据集上的性能也同样有所提升。在 CIFAR-FS 数据集上，使用本文预训练模型的 STVAE 和 FEAT 在 1-shot 和 5-shot 任务中分别取得了最优结果，准确率分别为 80.92% 和 90.18%，如表3.3所示。进一步地，本文还在 CUB 数据集上对这些方法进行了实验，其中 Meta-Baseline 在 1-shot 和 5-shot 少样本分类任务中表现突出，分别达到了 88.37% 和 95.52% 的准确率，如表3.4所示。这些实验结果表明，本文方法可以为这些两阶段元学习方法和生成方法提供一个优质的预训练模型，以提高它们的性能，也再一次验证了特征提取网络对少样本分类问题的重要性。

表 3.5 MGSRCL 在 miniImageNet、CIFAR-FS 和 CUB 数据集上的模块消融实验。最优结果用粗体表示。

Table 3.5 Module ablation experiments of MGSRCL on miniImageNet, CIFAR-FS and CUB. The best results are shown in bold.

数据集	方法	5-way 1-shot	5-way 5-shot
miniImageNet	Baseline	66.78 ± 0.43	83.82 ± 0.29
	Baseline w/ SS	67.76 ± 0.44	84.31 ± 0.28
	Baseline w/ TCL	68.45 ± 0.44	84.37 ± 0.29
	Baseline w/ CCL	68.61 ± 0.44	84.13 ± 0.29
	Baseline w/ TCL & CCL	69.21 ± 0.45	84.37 ± 0.30
	Baseline w/ all	**69.57 ± 0.45**	**84.41 ± 0.30**
CIFAR-FS	Baseline	74.39 ± 0.46	88.10 ± 0.33
	Baseline w/ SS	76.42 ± 0.45	88.62 ± 0.32
	Baseline w/ TCL	77.65 ± 0.47	88.51 ± 0.32
	Baseline w/ CCL	77.61 ± 0.47	88.59 ± 0.32
	Baseline w/ TCL & CCL	78.01 ± 0.48	88.27 ± 0.33
	Baseline w/ all	**78.54 ± 0.47**	**88.64 ± 0.32**
CUB	Baseline	82.18 ± 0.40	93.70 ± 0.20
	Baseline w/ SS	83.46 ± 0.39	94.18 ± 0.20
	Baseline w/ TCL	83.16 ± 0.39	93.74 ± 0.20
	Baseline w/ CCL	85.30 ± 0.38	94.50 ± 0.19
	Baseline w/ TCL & CCL	85.53 ± 0.38	94.30 ± 0.19
	Baseline w/ all	**86.14 ± 0.38**	**94.75 ± 0.19**

3.3.3 消融实验

（1）讨论不同模块对模型性能的影响

为了研究MGSRCL模型中每个模块对模型性能的影响，本文在miniImageNet、CIFAR-FS和CUB三个数据集上进行了全面的消融实验。本文的基准模型（Baseline）与RFS[19]相同，但为了缓解过拟合问题并实施变换一致性学习（TCL），本文添加了一些数据增强样本。

如表3.5所示，本文的基准模型在miniImageNet、CIFAR-FS和CUB的5-way 1-shot少样本分类任务中分别达到了66.78%、74.39%和82.18%的准确率。当添加自监督模块以预测执行了哪种图像变换时，在三个数据集上相较于基准模型分别获得了0.98%、2.03%和1.28%的效果提升。通过约束同一样本在不同变换下的样本内关系（在基准模型上添加TCL模块），在三个数据集上分别获得了1.67%、3.26%和0.98%的效果提升。通过约束同类样本的类内关系和不同类样本的类间关系（在基准模型上添加CCL模块），在三个数据集上相较于基准模型分别获得了1.83%、3.22%和3.12%的效果提升。此外，对于5-way 5-shot少样本分类任务，添加不同模块也使得模型取得了优于或与基准模型相当的结果。在CUB数据集上，添加CCL模块的结果明显优于添加其他模块，这是因为CUB是一个细粒度数据集，不同类别之间的差异相对较小，将不同类别在特征空间进行推远在CUB数据集上比在miniImageNet和CIFAR-FS数据集上更加有效。

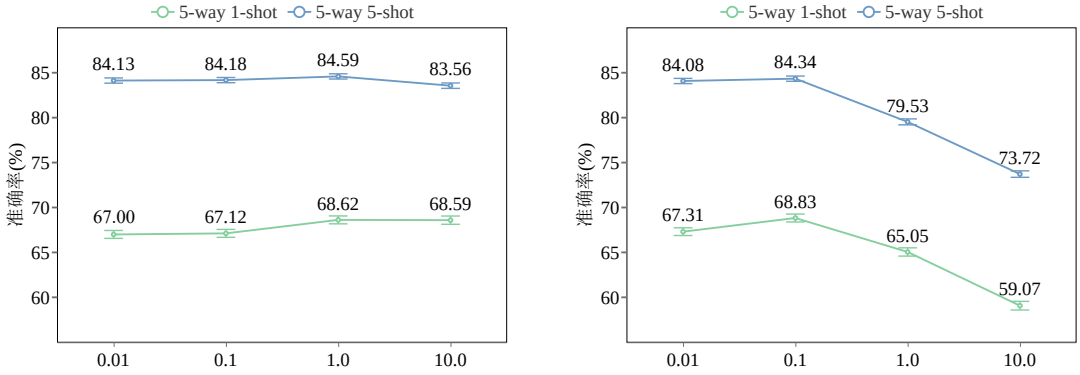
此外，联合使用TCL和CCL模块时，本文模型在5-way 1-shot任务中的结果优于仅使用其中一个模块，分类准确率在miniImageNet、CIFAR-FS和CUB数据集上分别达到了69.21%、78.01%和85.53%的准确率。当使用所有三个模块（TCL、CCL和SS）时，本文模型在三个数据集上都取得了最优性能，分别为69.57%、78.54%和86.14%的1-shot任务准确率，以及84.41%、88.64%和94.75%的5-shot任务准确率。综上所述，这些实验结果证明了本文方法每个模块的作用，以及所提出的TCL模块和CCL模块对挖掘不同粒度样本关系的有效性。

（2）讨论超参数 α 和 β 对模型性能的影响

α 和 β 是用来调整不同损失权重的超参数。本文通过将 α 和 β 设置为不同数值来评估模型在miniImageNet、CIFAR-FS和CUB数据集上的性能，从而确定其最优值。

首先，为了观察模型分类结果随参数变化的趋势，在讨论一个超参数的影响时，本文将另一个超参数设置为0。如图3.4、3.5和3.6所示，对于超参数 α，模型性能的变化较为平缓，整体呈现先上升再下降的趋势。当 $\alpha = 1.0$ 时，模型性能在三个数据集上都达到最优，分别达到68.62%、78.12%和83.54%的1-shot任务准确率以及84.59%、88.63%和94.03%的5-shot准确率。而对于超参数 β，随着 β 的增加，模型性能最初有所提高，然后呈现迅速下降的趋势，当 $\beta = 0.1$ 时，模型性能达到最优，分别达到68.83%、78.29%和86.01%的1-shot任务准确率以及84.34%、88.62%和94.75%的5-shot任务准确率。

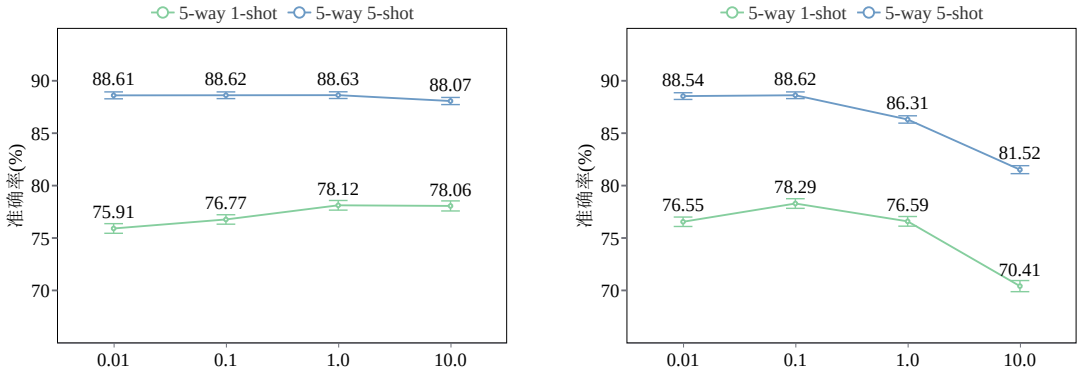
另外，为了全面准确地确定超参数 α 和 β 的最优值，本文采用网格搜索调参



(a) α　　(b) β

图 3.4　MGSRCL 在 miniImageNet 数据集上的超参数 α 和 β 消融实验。

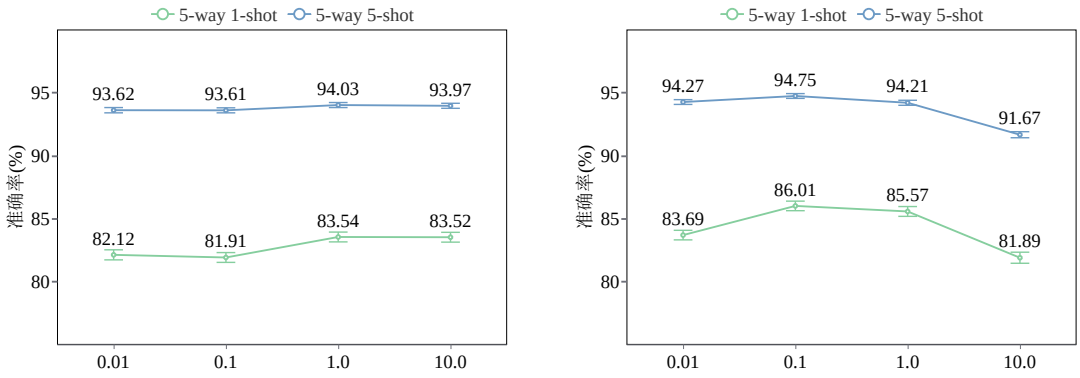
Fig. 3.4　Hyperparameters α and β ablation experiments of MGSRCL on miniImageNet.



(a) α　　(b) β

图 3.5　MGSRCL 在 CIFAR-FS 数据集上的超参数 α 和 β 消融实验。

Fig. 3.5　Hyperparameters α and β ablation experiments of MGSRCL on CIFAR-FS.



(a) α　　(b) β

图 3.6　MGSRCL 在 CUB 数据集上的超参数 α 和 β 消融实验。

Fig. 3.6　Hyperparameters α and β ablation experiments of MGSRCL on CUB.

表 3.6 MGSRCL 在 miniImageNet 数据集上的超参数 α 和 β 消融实验。最优结果用粗体表示。

Table 3.6 Hyperparameters α and β ablation experiments of MGSRCL on miniImageNet. The best results are shown in bold.

α \ β	0.01	0.1	1.0	10.0
0.01	67.57 ± 0.43	68.76 ± 0.44	65.71 ± 0.46	58.77 ± 0.47
0.1	67.07 ± 0.43	68.82 ± 0.44	65.88 ± 0.46	59.53 ± 0.48
1.0	68.95 ± 0.44	**69.57 ± 0.45**	66.68 ± 0.48	59.40 ± 0.48
10.0	68.37 ± 0.46	69.22 ± 0.47	67.16 ± 0.48	59.25 ± 0.48

表 3.7 MGSRCL 在 CIFAR-FS 数据集上的超参数 α 和 β 消融实验。最优结果用粗体表示。

Table 3.7 Hyperparameters α and β ablation experiments of MGSRCL on CIFAR-FS. The best results are shown in bold.

α \ β	0.01	0.1	1.0	10.0
0.01	76.87 ± 0.46	78.06 ± 0.46	76.04 ± 0.49	71.23 ± 0.52
0.1	77.12 ± 0.45	78.41 ± 0.46	76.67 ± 0.48	70.91 ± 0.53
1.0	78.01 ± 0.47	**78.54 ± 0.47**	76.60 ± 0.50	70.45 ± 0.52
10.0	78.02 ± 0.48	77.87 ± 0.49	76.46 ± 0.50	71.24 ± 0.52

表 3.8 MGSRCL 在 CUB 数据集上的超参数 α 和 β 消融实验。最优结果用粗体表示。

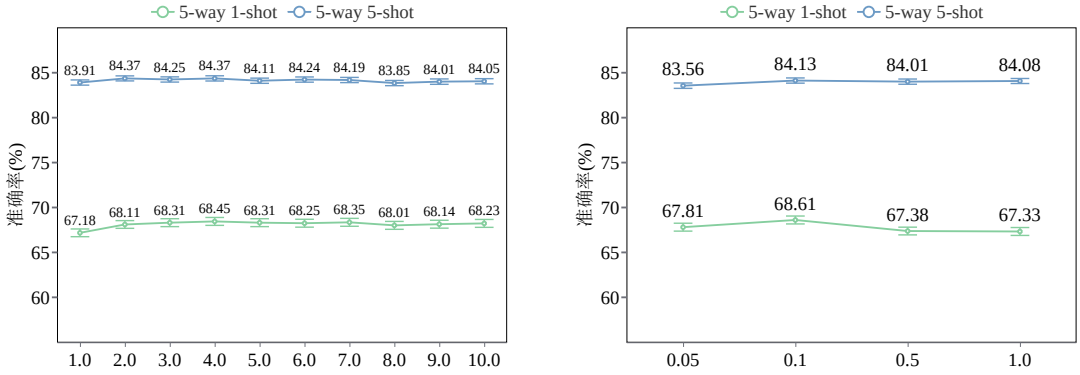
Table 3.8 Hyperparameters α and β ablation experiments of MGSRCL on CUB. The best results are shown in bold.

α \ β	0.01	0.1	1.0	10.0
0.01	82.79 ± 0.39	86.00 ± 0.38	85.42 ± 0.38	82.24 ± 0.43
0.1	82.89 ± 0.39	85.86 ± 0.37	86.07 ± 0.37	81.83 ± 0.44
1.0	83.82 ± 0.39	**86.14 ± 0.38**	86.09 ± 0.38	81.87 ± 0.44
10.0	85.12 ± 0.39	85.81 ± 0.38	86.07 ± 0.39	80.63 ± 0.46

法来进行实验，在 miniImageNet、CIFAR-FS 和 CUB 数据集上 5-way 1-shot 少样本分类任务的实验结果分别如表3.6、3.7和3.8所示。实验结果表明，当 α 和 β 分别设置为 1.0 和 0.1 时，模型在三个数据集上同时达到最优性能，1-shot 任务准确率分别为 69.57%、78.54% 和 86.14%。因此，在最终模型中，本文将 α 设置为 1.0，β 设置为 0.1。

（3）讨论超参数 τ_1 和 τ_2 对模型性能的影响

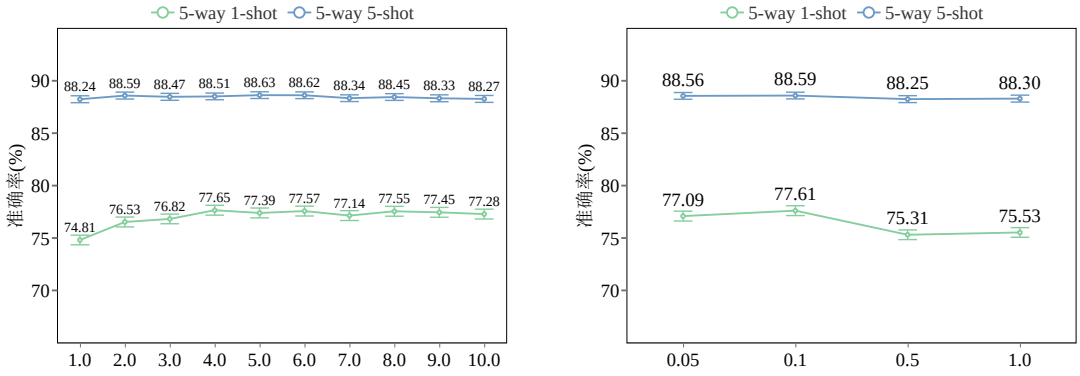
此外，本文还讨论了温度参数 τ_1 和 τ_2 对实验结果的影响。为了更清晰地观察



(a) τ_1　　(b) τ_2

图 3.7　MGSRCL 在 miniImageNet 数据集上的超参数 τ_1 和 τ_2 消融实验。

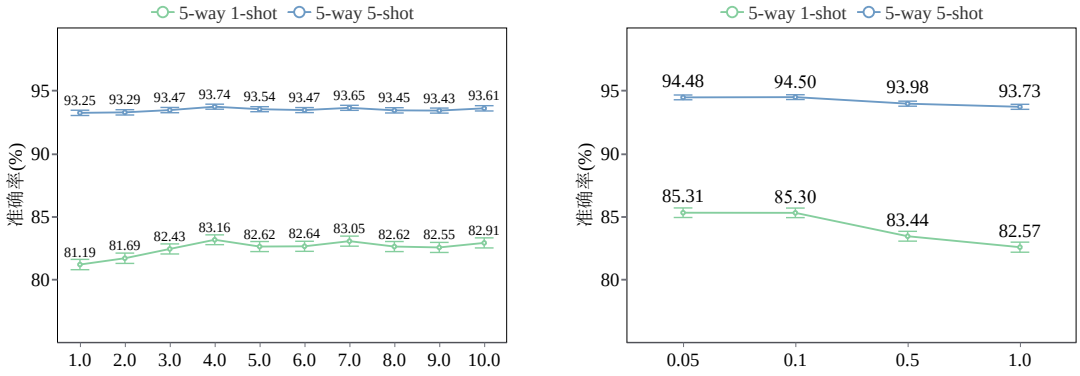
Fig. 3.7　Hyperparameters τ_1 and τ_2 ablation experiments of MGSRCL on miniImageNet.



(a) τ_1　　(b) τ_2

图 3.8　MGSRCL 在 CIFAR-FS 数据集上的超参数 τ_1 和 τ_2 消融实验。

Fig. 3.8　Hyperparameters τ_1 and τ_2 ablation experiments of MGSRCL on CIFAR-FS.



(a) τ_1　　(b) τ_2

图 3.9　MGSRCL 在 CUB 数据集上的超参数 τ_1 和 τ_2 消融实验。

Fig. 3.9　Hyperparameters τ_1 and τ_2 ablation experiments of MGSRCL on CUB.

温度参数在对应模块中起到的作用，在讨论温度参数时，本文只保留了相关模块，去除了其他模块。首先，τ_1 用于平滑预测输出以提供更多概率分布差异信息，本文令其从 1.0 到 10.0 变化，在 miniImageNet、CIFAR-FS 和 CUB 三个数据集上评估模型分类性能。如图3.7、3.8和3.9所示，当 τ_1 从 1.0 变化到 10.0 时，实验结果并未发生显著变化。当执行 5-way 1-shot 分类任务时，三个数据集都在 τ_1 设置为 4.0 时取得了最优结果，执行 5-way 5-shot 分类任务时，miniImageNet 和 CUB 数据集的也都在 τ_1 设置为 4.0 时取得了最优结果，CIFAR-FS 则是在 τ_1 设置为 5.0 时取得了最优结果。综合考虑，最终将 τ_1 设置为 4.0。τ_2 是 CCL 组件中使用的温度参数。本文评估了当 τ_2 设置为 0.05、0.1、0.5 和 1.0 时的模型性能。除了在 CUB 数据集的 5-way 1-shot 分类任务上将 τ_2 设置为 0.05 时取得了最优结果，其余数据集以及 5-way 5-shot 任务均是当 $\tau_2 = 0.1$ 时模型达到了最优结果。因此，在最终模型中，MGSRCL 将 τ_2 设置为 0.1。

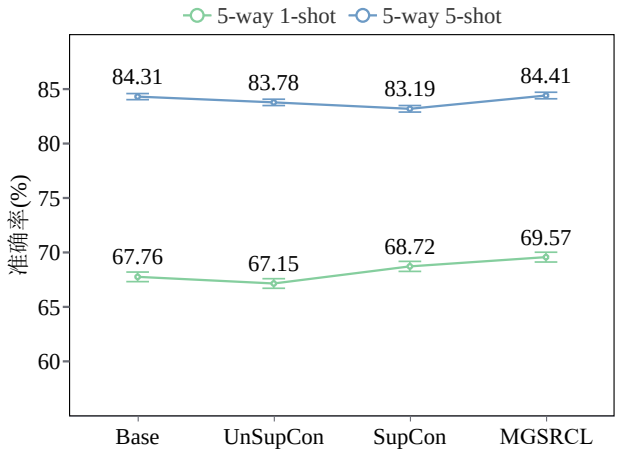


图 3.10 在 miniImageNet 数据集上的不同样本关系挖掘策略对比实验。

Fig. 3.10 Experimental comparison of different sample relationship mining strategies on miniImageNet.

（4）样本关系挖掘策略探讨

近年来，一些少样本分类方法利用对比学习来挖掘样本关系。然而，这些方法通常直接使用无监督对比学习（Unsupervised Contrastive Learning，简称 UnSupCon）[40] 或有监督对比学习（Supervised Contrastive Learning，简称 SupCon）[42] 作为辅助损失，这使得它们未能充分挖掘样本间的关系。为了表明本文方法相较于这些方法的优越性，本文基于所提出的基础特征学习网络进行了实验①。在实施过程中，UnSupCon 和 SupCon 损失作为辅助损失被直接添加到原基础特征学习网络中，与本文方法相同。如图3.10、3.11和3.12所示，在 miniImageNet、CIFAR-FS 和 CUB 数

① 此处，本文使用了 SupCon 提供的代码来实现无监督对比学习方法（SimCLR）和有监督对比学习方法（SupCon）。源代码可在https://github.com/HobbitLong/SupContrast获取。

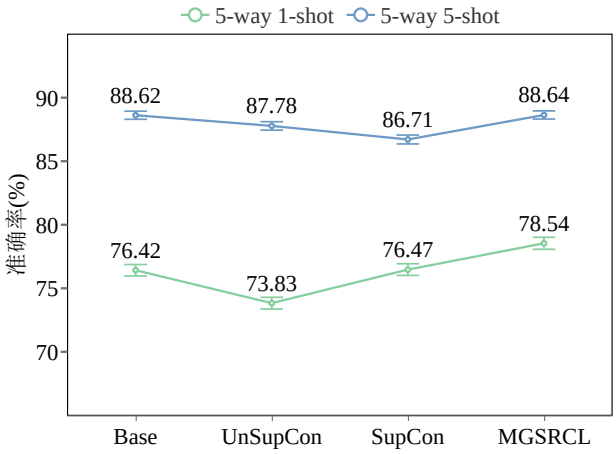


图 3.11　在 CIFAR-FS 数据集上的不同样本关系挖掘策略对比实验。

Fig. 3.11　Experimental comparison of different sample relationship mining strategies on CIFAR-FS.

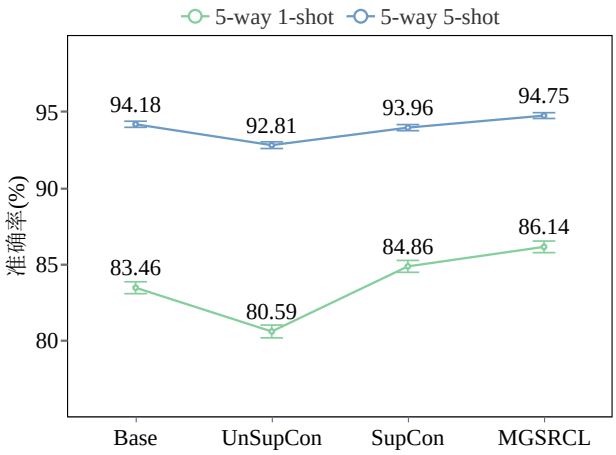
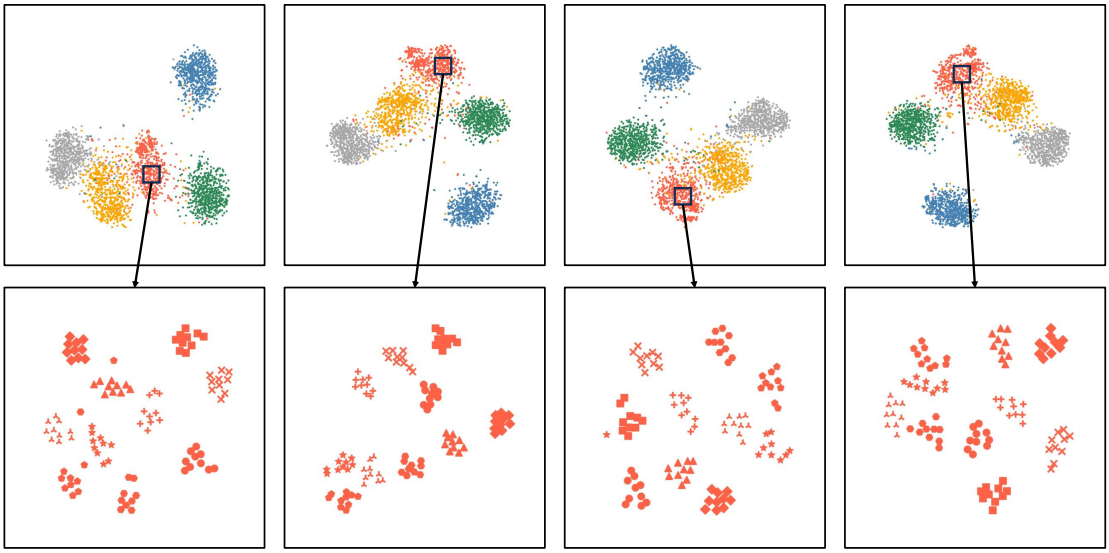


图 3.12　在 CUB 数据集上的不同样本关系挖掘策略对比实验。

Fig. 3.12　Experimental comparison of different sample relationship mining strategies on CUB.

据集上，添加 UnSupCon 后模型性能无论是 1-shot 分类任务还是 5-shot 分类任务，结果与基础模型相比都有所下降。这可以归因于 UnSupCon 将图像的变换视为正样本，而将其他图像视为负样本，会导致其在特征空间将同类样本推远，从而造成性能下降。另一方面，将 SupCon 整合到模型中并未受到此问题的影响，虽然在 5-way 5-shot 少样本分类任务上结果略微降低，但其在 1-shot 分类任务上结果都有所提升，且在 miniImageNet 与 CUB 数据集上提升明显。然而，SupCon 将一个样本的变换及其同类样本视为相同的关系，这是不合适的，因为一个样本的不同变换应具有一致的语义内容，而同类样本的语义内容应仅相似而不是完全一致。相比之下，本文提出的方法充分考虑了不同粒度的样本关系，并对它们进行了细致地建模，从而在三个数据集上都取得了最优结果，这表明了本文方法更充分地挖掘了样本关系并对其进行了有效建模。



(a) Base　(b) Base + TCL　(c) Base + CCL　(d) MGSRCL

图 3.13 miniImageNet 数据集上不同模型所提取特征的 t-SNE 可视化结果。不同颜色代表不同类别（第一行），同一形状代表同一样本的不同变换版本（第二行）。

Fig. 3.13 The t-SNE visualization results of features extracted by different models on miniImageNet. Different colors represent different categories (first row), and the same shape represents different transformed versions of the same sample (second row).

3.3.4 可视化分析

为了更好地展示 MGSRCL 模型的有效性，本文在 miniImageNet 上随机选取了 5 个新类，并使用 t-SNE 对不同模型提取的特征进行了可视化实验，如图3.13所示，其中 Base 表示本文的基础特征学习网络，Base + TCL 和 Base + CCL 分别代表基础特征学习网络整合了 TCL 模块以及 CCL 模块，而 MGSRCL 则代表本文的最终模型。在此图里，第一行展示了类内样本关系和类间样本关系，第二行则是展示了样本内关系，其中第一行图用不同颜色代表不同的类，第二行图则是将同一样本的不同变换版本用同一个形状表示。为便于检验第一行分类边界的质量，本文提供了如下数值数据：d_1/d_2，其中 d_1 表示五个类别样本与其样本中心间平均距离的均值（即类内样本内聚度），d_2 则表征各类别中心之间距离的平均值（即不同类别中心间的离散度）。具体数值如下：(a) Base：1.01，(b) Base + TCL：0.99，(c) Base + CCL：0.91，(d) MGSRCL：0.90。这些数据直观地反映了不同模型在分类边界质量上的表现差异。首先，在第一行图中，可以观察到基础特征学习网络模型 Base 已经具有明确的分类边界，如图3.13(a)所示，这证明使用全部基类数据训练一个特征提取网络便可取得一个较好的效果。与 Base 模型和 Base + TCL 模型相比，在 Base + CCL 模型和最终的 MGSRCL 模型中，同一类别内的样本表现出更好的内聚性，不同类样本之间边界更加明显，这证明了约束同类样本的类内关系和不同类样本的类间关系的有效性。此外，通过约束同一样本在不同变换下的样本

内关系，同一样本不同变换版本可以在特征空间保持更好的一致性，如图3.13(b)和图3.13(d)中的第二行图所示。而在其他模型中，一些变换的特征与其他变换的特征相距甚远，这展示了 TCL 模块的作用。综上所述，通过可视化实验证明，本文所提方法通过约束多种粒度的样本关系，对不同的样本关系进行了细致有效的建模，使得模型在新类上所提取的特征具有更好的判别性，从而达到了更高的分类准确率。

3.4 本章小结

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，针对少样本特征学习模型由于基类数据和新类数据类别不同而面临的模型特征提取能力不足的问题，充分挖掘了多种粒度的样本关系，并通过对比学习对多粒度样本关系进行建模，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，以帮助模型提取到更具判别性的样本特征。MGSRCL 方法使用一致性学习（TCL）模块通过标签分布对齐约束同一样本的不同变换具有一致的语义内容从而对样本内关系进行建模，以及类对比学习（CCL）模块通过拉近同类样本特征，同时推远不同类样本特征从而对类内样本关系以及类间样本关系进行建模。此外，还利用自监督学习技术令网络学习图像进行了何种变换，来增强基础特征学习网络的特征学习能力。在 miniImageNet、tieredImageNet、CIFAR-FS 和 CUB-200-2011 数据集的大量实验表明，MGSRCL 在各个少样本基准数据集都取得了优异的性能表现，并且还可以作为预训练模型整合到其他两阶段少样本分类方法中提升它们的分类性能。

综上所述，本章提出的基于多粒度样本关系建模的少样本特征学习算法通过对多种样本关系进行建模，充分挖掘了不同样本间的关系信息，提高了网络所提取特征的质量，并可以为其他少样本分类算法提供优质的预训练网络和高质量的样本特征。

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

4.1.1 研究动机

基于视觉的少样本算法，包括基于元学习、度量、数据增强、以及特征学习的方法已经取得了显著进展。然而，这些方法仍面临一些局限性，影响了模型性能的提升。特别是，这些方法往往仅关注样本的视觉信息，忽视了语义信息的潜在价值，使得模型难以从少量样本中学习到具有足够泛化能力的特征表示，这一点直接制约了模型对新类别的识别能力，限制了模型的性能上限。

图 4.1　人类认识新类别的过程示意。以斑马为例进行说明。

Fig. 4.1　Human process for recognizing new categories. Illustrated by the example of a zebra.

与深度学习模型不同，人类在认识新类别时，依赖的不仅是视觉形象，更多的是对于该形象背后语义的理解和推理。例如，告知某人“斑马”是一种有黑白条纹的“马”，即使在没有看到斑马的情况下，该人也能凭借对“马”的认知和对“黑白条纹”的描述，快速理解并识别“斑马”，如图4.1所示。这种通过语义信息桥接的学习过程，是人类认识新事物的一大优势，也是机器学习中尚待深入挖掘的宝贵

资源。受到人类认识新物体的启发，研究者认为语义信息在计算机视觉任务中也可发挥同样的作用，尤其是在零样本分类[72–78] 以及少样本分类[34–36,79–84] 任务中。为了模拟人类认识新类别的过程从而更好地建立新类与基类之间的联系，很多研究者开始使用自然语言模型[29–31] 或多模态模型[32] 的文本编码器提取类别名称的语义特征作为语义信息，并将其用到少样本分类方法中，如1.2.1中所述。

然而，这些工作虽在少样本分类任务中引入了语义信息并进行有效利用，但仍存在一些不足。部分基于特征生成的方法，如 STVAE[34] 使用语义信息作为条件训练一个特征生成模型。在训练好特征生成模型后，为了提高少样本测试任务中的样本多样性，需要向支持集中添加大量样本，这一做法使得后续的分类器训练时间急剧增长。并且部分生成模型训练过程较为复杂，例如生成对抗网络（Generative Adversarial Networks，简称 GAN）[18] 在训练时需要交替训练生成器和判别器；以及扩散模型（Diffusion Model）[85] 在训练时则是需要逐步加噪和去噪过程。基于语义修正的方法[35,36] 虽然不需要在进行少样本测试任务时对支持集样本进行扩充，但这些方法往往需要设计精细且复杂的信息融合模块，并且这些模块的设计还可能会对自然语言模型或多模态模型所提取的语义信息产生不利影响。具体来说，自然语言模型或多模态模型在大规模语料库上进行了训练，因此提取的语义特征具有较强的泛化性，而过于复杂的信息融合模块可能使得模型在基类数据集产生过拟合问题，削弱这种泛化性，进而降低模型的整体表现。

4.1.2 方法概述

为解决上述问题并充分利用语义信息以补充视觉信息，进而提升少样本分类任务的准确率，本章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，即语义-视觉多空间映射适配（Semantic-Visual Multi-Space Mapping Adapter，简称 SVMSMA）模型。SVMSMA 模型以一种简单的特征适配方式利用语义信息，不需要对语义特征进行复杂信息融合操作，从而避免了降低语义特征泛化性的问题，并能够有效丰富样本特征的信息来源，弥补仅使用视觉信息的不足。

由于在少样本分类测试任务中，查询集只能够获得视觉信息（已知语义信息则已知类别），为了执行测试任务以及使用跨模态分类任务（将在后续介绍）对网络进行优化，本文采用将语义信息映射到视觉空间的方案对两种信息之间的关系进行建模。因此，本文首先提出了一个语义-视觉多空间映射网络（Semantic-Visual Multi-Space Mapping Network，简称 SVMSMN），采用两种可独立使用的模式将语义信息映射到视觉空间：1）单模态映射，2）多模态映射。单模态映射是指使用类似零样本学习的思想，仅将语义信息映射到视觉空间，在执行测试任务时仅使用语义信息来获取语义映射特征，并不会使用到支持集样本的视觉特征。多模态映射则是在语义信息的基础上使用视觉信息对其进行补充，将两个模态信息融合后再映射到视觉空间，以对支持集样本的视觉信息加以利用。

另外，为了建模语义信息与视觉信息的关系，从而能够使用语义映射特征对

支持集的视觉特征进行补充，提升少样本分类任务性能，本文提出了两个模块以对语义-视觉多空间映射网络进行优化使得语义映射特征适配视觉空间，分别是 1）跨模态分类（Cross-Modal Classification，简称 CMC）模块，2）跨模态特征对齐（Cross-Modal Feature Alignment，简称 CMFA）模块。其中，跨模态分类模块使用预训练的视觉特征提取网络的分类器对上述语义映射特征进行分类，以对映射网络进行优化使得每个类别的语义信息和对应的视觉特征建立联系。跨模态特征对齐模块则是将语义映射特征与通过视觉特征提取网络得到的视觉特征原型进行对齐，以这种方式对映射后的特征进行修正，使其更接近类别原型，从而能够取得更好的分类结果。

本章方法同样在四个少样本分类数据集进行了实验，包括 miniImageNet[12]、tieredImageNet[37]、CIFAR-FS[38]，以及 CUB-200-2011[39]。实验结果表明，本章方法可有效建模语义-视觉空间关系，利用语义信息对视觉信息进行补充，从而取得优异的少样本分类结果。

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

在本节中，首先对使用语义信息的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语义-视觉多空间关系建模的特征适配模型进行简要介绍；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。

4.2.1 符号定义

在本章中，由于引入了语义信息，各种符号定义与第三章有所不同。基类数据集与新类数据集分别表示为：

$$\begin{aligned}\mathcal{D}_{base} &= \{(x,y,s)|x \in X^{base}, y \in Y^{base}, s \in S^{base}\},\\ \mathcal{D}_{novel} &= \{(x,y,s)|x \in X^{novel}, y \in Y^{novel}, s \in S^{novel}\}.\end{aligned} \tag{4.1}$$

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、s 分别表示样本图像、样本标签、以及样本语义特征；X^{base}、Y^{base}、S^{base} 分别表示基类样本图像集合、标签集合、语义特征集合；X^{novel}、Y^{novel}、S^{novel} 则分别表示新类样本图像集合、标签集合、语义特征集合。在本章中，语义特征是通过将类别名称/提示文本 + 类别名称输入自然语言处理模型或者多模态模型的文本编码器得到的。

与上一章相同，本章也通过在 $\mathcal{D}_{novel}$ 中采样大量少样本分类任务并计算平均准确率来评估模型性能。不同的是，本章所采样少样本分类任务的支持集 $\mathcal{S}_{\mathcal{T}}$ 除了包含样本图像 x_i 以及样本标签 y_i 外，还包含了样本语义特征 s_i。因此，支持集表

示为：

$$\mathcal{S}_{\mathcal{T}} = \{(x_i, y_i, s_i) | x_i \in X^{\mathcal{T}}, y_i \in Y^{\mathcal{T}}, s_i \in S^{\mathcal{T}}\}_{i=1}^{N \times K}, \tag{4.2}$$

其中，$X^{\mathcal{T}}$、$Y^{\mathcal{T}}$、$S^{\mathcal{T}}$ 分别表示所采样任务的样本图像集合、标签集合、语义特征集合；N 和 K 则是表示类别数目以及每个类别样本数目。查询集的样本类别未知，因此无法获取其语义特征，表示为：

$$\mathcal{Q}_{\mathcal{T}} = \{(x_i, y_i) | x_i \in X^{\mathcal{T}}, y_i \in Y^{\mathcal{T}}\}_{i=1}^{N \times Q}, \tag{4.3}$$

Q 表示每个类别用作测试的样本数目。

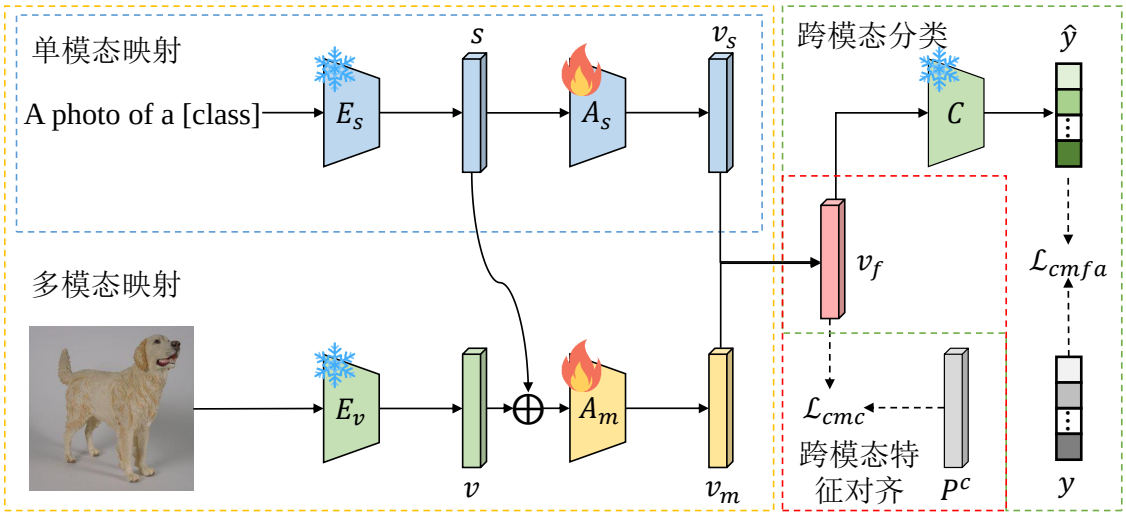


图 4.2　语义-视觉多空间映射适配模型示意图。该模型包含一个视觉特征提取网络 E_v、一个语义特征提取网络 E_s，两个语义-视觉多空间映射网络 A_s 与 A_m，以及一个分类器 C。在此图中，v 与 s 分别表示视觉特征与语义特征，v_s 与 v_m 则是通过 A_s 与 A_m 获得的映射特征，统称为 v_f。$\widehat{y}$、y、P^c 分别表示预测输出、真实标签和对应类原型特征。$\mathcal{L}_{cmc}$ 与 $\mathcal{L}_{cmfa}$ 表示跨模态分类损失和跨模态特征对齐损失。另外，模型训练过程中，E_v、E_s 与 C 的参数都被冻结。

Fig. 4.2　Illustration of Semantic-Visual Multi-Space Mapping Adapter. The model includes a visual feature extractor E_v, a semantic feature extractor E_s, two semantic-visual multi-space mapping networks A_s and A_m, and a classifier C. In this figure, v and s represent the visual and semantic features, respectively, while v_s and v_m are the mapping features obtained through A_s and A_m, collectively referred to as v_f. $\widehat{y}$, y, and P^c denote the predicted output, true label, and corresponding class prototype features, respectively. $\mathcal{L}_{cmc}$ and $\mathcal{L}_{cmfa}$ represent the cross-modal classification loss and cross-modal feature alignment loss, respectively. Additionally, during the model training process, the parameters of E_v, E_s, and C are all frozen.

4.2.2 整体框架

本章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，即语义-视觉多空间映射适配模型（Semantic-Visual Multi-Space Mapping Adapter，简称

SVMSMA），用以建模样本的语义-视觉多空间关系。该算法旨在利用语义信息作为视觉信息的补充，丰富模型所提取样本特征的信息来源，提升模型在新类上的泛化能力。如图4.2所示，SVMSMA 模型主要包含三个部分：语义-视觉多空间映射网络（Semantic-Visual Multi-Space Mapping Network，简称 SVMSMN），跨模态分类（Cross-Modal Classification，简称 CMC）模块，以及跨模态特征对齐（Cross-Modal Feature Alignment，简称 CMFA）模块。具体来说，语义-视觉多空间映射网络（SVMSMN）用以将语义特征映射到视觉空间，以方便后续与视觉特征进行建模；跨模态分类（CMC）模块通过对映射后的语义特征执行分类任务，以优化映射网络从而使得语义特征与视觉特征建立联系；跨模态特征对齐（CMFA）模块则是将映射后的语义特征与视觉特征原型进行对齐，以对映射后的特征进行修正从而获得更接近类别原型的特征。接下来，本节将会对 SVMSMA 模型的每个部分进行详细介绍。

4.2.3 语义-视觉多空间映射网络

以方便后续对语义信息与视觉信息进行建模，需要将语义特征和视觉特征投影到相同空间。为了充分利用自然语言处理模型或多模态模型在大规模语料库上对类别名称建立的联系，以及后续建模算法（包括跨模态分类以及跨模态特征对齐）的进行，本文采用将语义特征映射到视觉空间的方式，提出了语义-视觉多空间映射网络（Semantic-Visual Multi-Space Mapping Network，简称 SVMSMN），以使得两种信息的特征向量维度一致。本章提出的语义-视觉多空间映射网络可分为两种模式：1）单模态映射，2）多模态映射。这两种模式的核心区别在于它们所处理特征信息的模态种类不同，以下分别对其进行介绍。

（1）单模态映射

单模态映射模式专注于语义信息的处理，其核心思想借鉴了零样本分类领域的思想。该模式的目标是建立一个能够将纯语义信息（例如，文本描述或类别名称）映射至视觉特征空间的高效网络。在实践中，这一过程首先通过语义特征提取网络 E_s，将类名或提示文本 $text$ 转化为语义特征 s，

$$s = E_s(text). \tag{4.4}$$

随后，单模态映射网络 A_s 负责将这些语义特征映射至视觉空间，生成对应的单模态映射特征 v_s。该过程可表示为以下公式，

$$v_s = A_s(s). \tag{4.5}$$

值得注意的是，在本章少样本分类的测试任务中，单模态映射不依赖于任何视觉信息，即不需要支持集的视觉特征参与，这一点使得其更类似于零样本分类

的设置。

（1）多模态映射

与单模态映射相比，多模态映射模式采用了一种更为综合和动态的策略，它不仅处理语义信息，同时也将视觉信息作为语义信息的补充。这种模式首先利用预训练的视觉特征提取网络 E_v，从给定样本图像 x 中提取出丰富的视觉特征 v，

$$v = E_v(x), \tag{4.6}$$

另外通过语义特征提取网络 E_s 获取相应的语义特征 s，如公式4.4所示。接着，需要对视觉特征和语义特征进行融合，并使用多模态映射网络 A_m 将融合后的多模态特征映射到视觉空间，得到对应的多模态映射特征 v_m，本章所采用的融合方式为最简单的 concat 操作。该过程可表达为以下公式，

$$v_m = A_m(\text{concat}(v, s)), \tag{4.7}$$

在少样本分类任务中，这种多模态映射模式能够充分利用支持集的视觉特征，以及待分类样本的语义特征，使得所提取特征的信息来源更加丰富，从而更好地适应少样本分类的挑战。

训练过程中，无论是单模态映射模式还是多模态映射模式，SVMSMN 都以适配器（Adapter）的方式被插入视觉特征提取网络 E_v 与语义特征提取网络 E_s 之后，视觉特征分类器 C 之前，以在不牺牲特征提取网络泛化能力的前提下，适应本文后续提出的跨模态分类任务和跨模态特征对齐任务。

4.2.4 语义-视觉多空间关系建模算法

为了建模语义信息和视觉信息的关系，本章提出了两个模块对上一节介绍的语义-视觉多空间映射网络进行优化：1）跨模态分类模块，2）跨模态特征对齐模块。以下将分别对两个模块及其损失进行介绍。

（1）跨模态分类

跨模态分类（Cross-Modal Classification，简称 CMC）模块旨在通过分类任务对语义-视觉多空间映射网络进行优化，强化语义信息与视觉信息之间的联系，使得映射后的语义特征能够与实际视觉特征具有一致性。该模块利用预训练视觉特征提取网络时得到的分类器，对通过映射网络得到的单模态或多模态映射特征进行分类，从而促使模型学习到每个类别的语义信息和对应视觉信息之间的紧密对应关系。

具体而言，跨模态分类模块接收映射网络输出的单模态映射特征 v_s 或多模态映射特征 v_m 作为输入，利用分类器 C 计算每个样本的概率分布。该过程可以表

达为以下公式，

$$\widehat{y} = \text{Softmax}(C(v_f)), \tag{4.8}$$

其中 v_f 表示输入的映射特征，可以为 v_s 或 v_m，$\widehat{y}$ 表示样本的预测概率分布。跨模态分类的损失函数采用交叉熵损失，表示为以下公式，

$$\mathcal{L}_{cmc} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \widehat{y}_i, \tag{4.9}$$

其中 N 表示样本数目，y_i 表示样本类别标签。通过最小化 $\mathcal{L}_{cmc}$，模型能够学习到如何将语义信息有效映射到视觉空间，并确保映射后的语义特征与实际类别之间具有高度的一致性。

（2）跨模态特征对齐

跨模态特征对齐（Cross-Modal Feature Alignment，简称 CMFA）模块的目的是通过对齐映射特征 v_f 和视觉特征提取网络得到的视觉特征原型 P^c，以对 v_f 进行修正，使其更接近类别原型，从而提高映射特征 v_f 对每个类别的代表性，进一步提升少样本分类的准确性。

该模块首先计算每个类别的视觉特征原型 P^c，即该类别所有样本视觉特征的平均值。P^c 可通过以下公式计算，

$$P^c = \frac{1}{|X^c|}\sum_{i=1}^{|X^c|} v_i, \tag{4.10}$$

其中 X^c 是类别 c 中所有样本的集合，v_i 是样本 x_i 的视觉特征。接着，CMFA 模块计算映射后的视觉特征和对应类别原型之间的距离，并通过最小化该距离来实现特征对齐，该损失函数表示为以下公式，

$$\mathcal{L}_{cmfa} = \frac{1}{N}\sum_{i=1}^{N} ||v_{fi} - P^{y_i}||_2^2, \tag{4.11}$$

其中 $||\cdot||_2$ 表示 L2 范数，用于衡量映射特征 v_{fi} 与对应类别原型 P^{y_i} 之间的欧式距离。通过优化损失函数 $\mathcal{L}_{cmfa}$，模型能够引导语义映射特征更贴近于类别的视觉中心，从而获得更具代表性的样本特征，在少样本分类的测试阶段实现更高的准确率。

4.2.5 模型优化

结合公式4.9和4.11，本文提出的 SVMSMA 模型总体损失函数可以表示为：

$$\mathcal{L}_{total} = \mathcal{L}_{cmc} + \alpha \cdot \mathcal{L}_{cmfa}, \tag{4.12}$$

其中，α 是用来衡量不同损失权重的超参数。

SVMSMA 模型在整个基类数据集进行训练，没有采用元学习的方式，并通过最小化损失函数 $\mathcal{L}_{total}$ 对模型参数进行联合优化，通过引入语义信息并对语义-视觉多空间关系进行建模，从而能够丰富模型所获得的信息，更好地迁移在基类数据集上学习到的知识，提升模型的泛化能力。另外，本文中所使用的语义特征提取网络 E_s 基于现有的自然语言处理模型或多模态模型的文本编码器，视觉特征提取网络 E_v 基于使用基类数据集的图像样本预训练的特征提取网络，分类器 C 基于训练视觉特征提取网络时所使用的分类器，E_s、E_v 和 C 的网络参数在本章模型训练时都是冻结的，换句话说，只有单模态映射网络 A_s 和多模态映射网络 A_m 参与了网络参数更新，并且这两个网络是单独进行训练的。

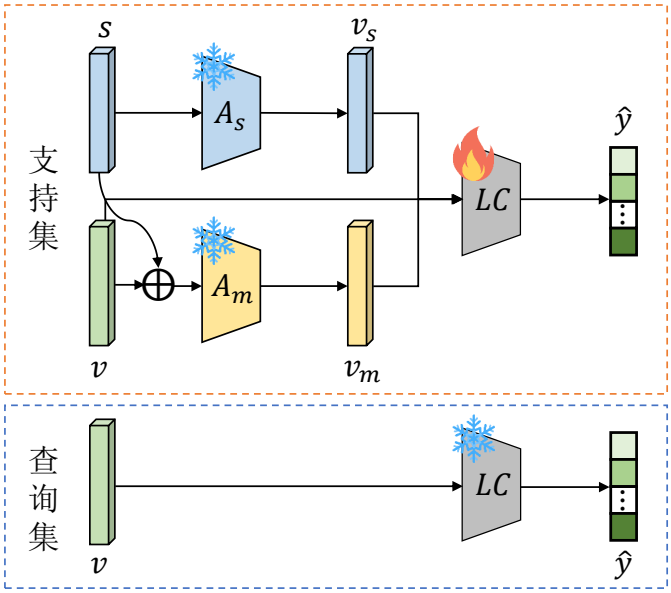


图 4.3 SVMSMA 模型推理过程示意图。推理过程中，使用支持集视觉特征 v、单模态映射特征 v_s 和多模态映射特征 v_m 训练一个逻辑回归分类器 LC，并使用查询集视觉特征测试分类器性能。

Fig. 4.3 Illustration of SVMSMA model inference process. During the inference process, a logistic regression classifier LC is trained using the visual features v, single-modal mapping features v_s, and multi-modal mapping features v_m of the support set, and the classifier's performance is tested using the visual features of the query set.

4.2.6 模型推理

与第三章相同，本章模型在训练完成之后，在测试阶段同样冻结所有模型参数并通过使用逻辑回归分类器 LC 执行多个少样本分类任务，将平均准确率作为模型的评价指标。不同的是，每个任务的推理过程中，本章会使用单模态映射网络 A_s 以及多模态映射网络 A_m 得到的语义映射特征 v_s 和 v_m 对原本支持集 $\mathcal{S}_\mathcal{T}$ 的视觉特征 v 进行扩充，以丰富支持集样本特征的信息来源及其多样性，从而使分

类器学习到更好的分类边界，达到更好的分类性能。与生成方法经常扩充支持集样本至原来的几十倍甚至上百倍不同，本文方法仅将支持集样本特征扩充至原来的 3 倍。对于查询集，由于无法获得其类别名称（获得类别名称相当于已知其类别），则是将其视觉特征输入训练完成的分类器 LC 获得其预测类别。本章模型推理过程如图4.3所示。

4.3 实验设置及结果分析

在本节中，首先介绍了本章方法的实验设置，包括实验使用数据集、网络结构、以及优化设置，然后分析了基于语义-视觉多空间关系建模的少样本特征适配算法实验结果，接下来对模型的各个模块以及超参数进行了消融实验和分析，最后对模型所提取特征进行了可视化分析。

4.3.1 实验设置

（1）实验数据集介绍

与第三章相同，本章方法同样在四个少样本分类基准数据集进行了实验，包括三个普通少样本分类数据集：miniImageNet[12]、tieredImageNet[37]、CIFAR-FS[38]，以及一个细粒度少样本分类数据集：CUB-200-2011（CUB）[39]。不同的是，第三章方法仅使用了数据集中每个样本的图像以及标签，而本章方法还使用了数据集中包含的样本类别名称来提供语义信息。

（2）网络结构

在没有特殊说明的情况下，本文所使用视觉特征提取网络为第三章 MGSRCL 模型的特征提取网络，语义特征提取网络与 SP-CLIP[36] 相同，为使用 ViT-B/16 架构作为图像编码器的 CLIP 模型的文本编码器，其输出语义特征维度为 512。并且在使用 CLIP 模型获取语义特征时，本章方法使用了基于提示的方法，将提示文本“A photo of a”与类别名称“[class]”拼接后作为输入，如图4.2所示。语义-视觉多空间映射网络则是由两层全连接，和处于两层全连接之间的 ReLU 激活函数组成的多层感知机。输入层维度为语义特征维度（单模态映射）或视觉特征维度和语义特征维度之和（多模态映射），隐藏层维度为 4096，输出层维度为视觉特征维度。

（3）优化设置

对于所有实验，本文采用具有 $5e^{-4}$ 权重衰减的自适应矩估计（Adaptive Moment Estimation，简称 Adam）优化器对模型进行优化，对于 tieredImageNet 数据集，学习率设置为 $1e^{-4}$，其余数据集均设置为 $1e^{-5}$，对于所有数据集均训练 100 个轮次。在训练过程中，视觉特征提取网络 E_v、语义特征提取网络 E_s、分类器 C 参数均被冻结，只有单模态映射网络 A_s 或多模态映射网络 A_m 参与参数更新。对于超参数 α，本文在 CIFAR-FS 数据集将其设置为 10.0，其他数据集均设置为 1.0，超参数讨论将在4.3.3部分进行。

4.3.2 基准数据集实验结果

为了评估 SVMSMA 模型的有效性，本文在四个数据集上进行了大量实验。表4.1、4.2、4.3和4.4分别展示了本章方法在 miniImageNet、tieredImageNet、CIFAR-FS 和 CUB 数据集上的实验结果以及一些其他现有少样本分类方法的结果。

（1）普通少样本分类

与其他少样本方法相比较，本章方法 SVMSMA 在 miniImageNet、CIFAR-FS 数据集达到了最优结果，在 tieredImageNet 数据集达到了第二优结果，如表4.1、4.2和4.3所示。具体而言，本章提出的 SVMSMA 模型在 miniImageNet 数据集上的 5-way 1-shot 和 5-way 5-shot 任务分别达到了 79.58% 和 85.19% 的准确率，与第二优方法相比分别高出了 7.27% 和 0.79%。在 CIFAR-FS 数据集，SVMSMA 模型在 1-shot 和 5-shot 任务则是分别达到了 86.35% 和 89.42% 的准确率，比分别在 1-shot 和 5-shot 任务上达到次优结果的 SP-CLIP 和 IER 方法高出了 4.17% 和 0.16%。在 tieredImageNet 数据集，则是 SP-CLIP 方法达到了最优结果，本章方法达到了第二优结果，1-shot 与 5-shot 任务准确率分别是 75.35% 和 86.35%。这些实验结果证明了本章方法的有效性。

表 4.1 SVMSMA 在 miniImageNet 数据集上的分类准确率（%）。最优结果用粗体表示，“Visual”与“Semantic”分别表示没有使用语义信息的方法与基于语义的方法。

Table 4.1 Classification accuracy (%) of SVMSMA on miniImageNet. The best results are shown in bold, “Visual” and “Semantic” respectively refer to methods without the use of semantic information and methods based on semantic information.

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	MAML[7]	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92
	ProtoNet[11]	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66
	DeepEMD[13]	ResNet-12	65.91 ± 0.82	82.41 ± 0.56
	RFS-distill[19]	ResNet-12	64.82 ± 0.60	82.14 ± 0.43
	AssoAlign[57]	ResNet-18	59.88 ± 0.67	80.35 ± 0.73
	FEAT[68]	ResNet-12	66.78 ± 0.20	82.05 ± 0.14
	GIFSL[58]	ResNet-12	65.47 ± 0.63	82.75 ± 0.42
	MELR[59]	ResNet-12	67.40 ± 0.43	83.40 ± 0.28
	IEPT[52]	ResNet-12	67.05 ± 0.44	82.90 ± 0.30
	IER[23]	ResNet-12	66.82 ± 0.80	84.35 ± 0.51
	RENet[56]	ResNet-12	67.60 ± 0.44	82.58 ± 0.30
	PAL[24]	ResNet-12	69.37 ± 0.64	84.40 ± 0.44
	HandCrafted[22]	ResNet-12	67.14 ± 0.76	83.11 ± 0.69
	SCL-distill[25]	ResNet-12	67.40 ± 0.76	83.19 ± 0.54
	HGNN[61]	ResNet-12	67.02 ± 0.20	83.00 ± 0.13
	APP2S[62]	ResNet-18	64.82 ± 0.12	81.31 ± 0.22

表 4.1 （续）
Table 4.1 (continued)

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	ESPT[53]	ResNet-12	68.36 ± 0.19	84.11 ± 0.12
	Meta-HP[64]	ResNet-12	62.49 ± 0.80	77.12 ± 0.62
	SAPENet[65]	ResNet-12	66.41 ± 0.20	82.76 ± 0.14
	FEAT+DFR[66]	ResNet-12	67.74 ± 0.86	82.49 ± 0.57
	DiffKendall[67]	ResNet-12	65.56 ± 0.43	80.79 ± 0.31
	MetaDiff[86]	ResNet-12	64.99 ± 0.77	81.21 ± 0.56
Semantic	KTN[79]	64-64-128-128	64.42 ± 0.72	74.16 ± 0.56
	DualTriNet[33]	ResNet-18	58.12 ± 1.37	76.92 ± 0.69
	AM3[35]	ResNet-12	65.30 ± 0.49	78.10 ± 0.36
	TRAML[80]	ResNet-12	67.10 ± 0.52	79.54 ± 0.60
	AM3-BERT[81]	ResNet-12	68.42 ± 0.51	81.29 ± 0.59
	CMGNN-DPGN[82]	ResNet-12	71.38 ± 0.51	82.60 ± 0.47
	STVAE[34]	ResNet-12	63.62 ± 0.80	80.68 ± 0.48
	SP-CLIP[36]	Visformer-T	72.31 ± 0.40	83.42 ± 0.30
	SVMSMA	ResNet-12	**79.58 ± 0.35**	**85.19 ± 0.29**

表 4.2 SVMSMA 在 tieredImageNet 数据集上的分类准确率（%）。最优结果用粗体表示，带有“†”标记的方法表示结果是使用作者提供代码所实现，“Visual”与“Semantic”分别表示没有使用语义信息的方法与基于语义的方法。

Table 4.2 Classification accuracy (%) of SVMSMA on tieredImageNet. The best results are shown in bold, and methods with the “†” indicate that the result was implemented using author-supplied code, “Visual” and “Semantic” respectively refer to methods without the use of semantic information and methods based on semantic information.

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	MAML[7]	32-32-32-32	51.67 ± 1.81	70.30 ± 1.75
	ProtoNet[11]	64-64-64-64	53.31 ± 0.89	72.69 ± 0.74
	DeepEMD[13]	ResNet-12	71.16 ± 0.87	86.03 ± 0.58
	RFS-distill[19]	ResNet-12	71.52 ± 0.69	86.03 ± 0.49
	AssoAlign[57]	ResNet-18	69.29 ± 0.56	85.97 ± 0.49
	FEAT[68]	ResNet-12	70.80 ± 0.23	84.79 ± 0.16
	GIFSL[58]	ResNet-12	72.39 ± 0.66	86.91 ± 0.44
	MELR[59]	ResNet-12	72.14 ± 0.51	87.01 ± 0.35
	IEPT[52]	ResNet-12	72.24 ± 0.50	86.73 ± 0.34
	IER[23]	ResNet-12	71.87 ± 0.89	86.82 ± 0.58
	RENet[56]	ResNet-12	71.16 ± 0.51	85.28 ± 0.35
	PAL[24]	ResNet-12	72.25 ± 0.72	86.95 ± 0.47

表 4.2 （续）
Table 4.2 (continued)

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	SCL-distill[25]	ResNet-12	71.98 ± 0.91	86.19 ± 0.59
	HGNN[61]	ResNet-12	72.05 ± 0.23	86.49 ± 0.15
	APP2S[62]	ResNet-18	70.83 ± 0.15	84.15 ± 0.29
	ESPT[53]	ResNet-12	72.68 ± 0.22	87.49 ± 0.14
	Meta-HP[64]	ResNet-12	68.26 ± 0.72	82.91 ± 0.36
	SAPENet[65]	ResNet-12	68.63 ± 0.23	84.30 ± 0.16
	FEAT+DFR[66]	ResNet-12	71.31 ± 0.93	85.12 ± 0.64
	DiffKendall[67]	ResNet-12	70.76 ± 0.43	85.31 ± 0.34
	MetaDiff[86]	ResNet-12	72.33 ± 0.92	86.31 ± 0.62
Semantic	AM3[35]	ResNet-12	69.08 ± 0.47	82.58 ± 0.31
	AM3-BERT[81]	ResNet-12	73.76 ± 0.72	87.51 ± 0.75
	CMGNN-DPGN[82]	ResNet-12	72.89 ± 0.49	84.92 ± 0.48
	STVAE†[34]	ResNet-12	68.32 ± 0.94	83.79 ± 0.66
	SP-CLIP[36]	Visformer-T	**78.03 ± 0.46**	**88.55 ± 0.32**
	SVMSMA	ResNet-12	75.35 ± 0.47	86.35 ± 0.34

表 4.3 SVMSMA 在 CIFAR-FS 数据集上的分类准确率（%）。最优结果用粗体表示，“Visual”与“Semantic”分别表示没有使用语义信息的方法与基于语义的方法。

Table 4.3 Classification accuracy (%) of SVMSMA on CIFAR-FS. The best results are shown in bold, “Visual” and “Semantic” respectively refer to methods without the use of semantic information and methods based on semantic information.

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	MAML[7]	32-32-32-32	58.90 ± 1.90	71.50 ± 1.00
	ProtoNet[11]	64-64-64-64	55.50 ± 0.70	72.00 ± 0.60
	RFS-distill[19]	ResNet-12	73.90 ± 0.80	86.90 ± 0.50
	GIFSL[58]	ResNet-12	74.58 ± 0.38	87.68 ± 0.23
	IER[23]	ResNet-12	76.83 ± 0.82	89.26 ± 0.58
	RENet[56]	ResNet-12	74.51 ± 0.46	86.60 ± 0.32
	PAL[24]	ResNet-12	77.10 ± 0.70	88.00 ± 0.50
	HandCrafted[22]	ResNet-12	76.68 ± 0.59	87.49 ± 0.73
	SCL-distill[25]	ResNet-12	76.50 ± 0.90	88.00 ± 0.60
	ConstellationNet[70]	ResNet-12	75.40 ± 0.20	86.80 ± 0.20
	APP2S[62]	ResNet-18	73.12 ± 0.22	85.69 ± 0.16
	Meta-HP[64]	ResNet-12	73.74 ± 0.57	86.37 ± 0.32
Semantic	DualTriNet[33]	ResNet-18	63.41 ± 0.64	78.43 ± 0.62
	STVAE[34]	ResNet-12	76.30 ± 0.60	87.00 ± 0.40

表 4.3（续）
Table 4.3 (continued)

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Semantic	SP-CLIP[36]	Visformer-T	82.18 ± 0.40	88.24 ± 0.32
	SVMSMA	ResNet-12	**86.35 ± 0.35**	**89.42 ± 0.30**

（2）细粒度少样本分类

同样的，本章方法也在细粒度少样本分类数据集 CUB 上进行了实验，结果如表4.4所示。在 CUB 数据集，SVMSMA 模型同样取得了最优结果，在 1-shot 和 5-shot 任务上分别达到了 91.17% 和 94.91% 的平均准确率，比第二优的方法分别高出了 5.72% 和 0.18%。这一实验证明了在类别差异较小的细粒度数据集，语义信息同样能够为视觉信息提供帮助，获得进一步性能提升。

表 4.4 SVMSMA 在 CUB 数据集上的分类准确率（%）。最优结果用粗体表示，“Visual”与“Semantic”分别表示没有使用语义信息的方法与基于语义的方法。

Table 4.4 Classification accuracy (%) of SVMSMA on CUB. The best results are shown in bold, “Visual” and “Semantic” respectively refer to methods without the use of semantic information and methods based on semantic information.

	方法	特征提取网络	5-way 1-shot	5-way 5-shot
Visual	FEAT[68]	64-64-64-64	68.87 ± 0.22	82.90 ± 0.15
	DeepEMD[13]	ResNet-12	75.65 ± 0.83	88.69 ± 0.50
	AssoAlign[57]	ResNet-18	74.22 ± 1.09	88.65 ± 0.55
	MELR[59]	64-64-64-64	70.26 ± 0.50	85.01 ± 0.32
	IEPT[52]	64-64-64-64	69.97 ± 0.49	84.33 ± 0.33
	RENet[56]	ResNet-12	79.49 ± 0.44	91.11 ± 0.24
	HGNN[61]	ResNet-12	78.58 ± 0.20	90.02 ± 0.12
	APP2S[62]	ResNet-12	77.64 ± 0.19	90.43 ± 0.18
	ESPT[53]	ResNet-12	85.45 ± 0.18	94.02 ± 0.09
	SAPENet[65]	64-64-64-64	70.38 ± 0.23	84.47 ± 0.14
	FEAT+DFR[66]	ResNet-12	77.14 ± 0.21	88.97 ± 0.13
	Bi-FRN[71]	ResNet-12	85.44 ± 0.18	94.73 ± 0.09
Semantic	DualTriNet[33]	ResNet-18	69.61 ± 0.46	84.10 ± 0.35
	AM3-BERT[81]	ResNet-12	77.03 ± 0.85	87.20 ± 0.70
	STVAE[34]	ResNet-12	77.32 ± 0.00	86.84 ± 0.00
	SVMSMA	ResNet-12	**91.17 ± 0.28**	**94.91 ± 0.19**

综上所述，本章提出的 SVMSMA 方法在 miniImageNet、CIFAR-FS 和 CUB 三个数据集达到了最优结果，在 tieredImageNet 数据集也达到了第二优结果，这证明

了 SVMSMA 模型通过引入语义信息对视觉信息进行补充，使得样本特征信息来源更加丰富，从而获得更具代表性和多样性的特征，提升模型的泛化能力。另外，相比于 5-way 5-shot 任务，SVMSMA 模型在 5-way 1-shot 任务取得了更为明显的效果提升，这是因为 5-shot 任务样本多样性已较为充足，能够使得分类器从中提取到样本关键特征进行分类，而 1-shot 任务样本数量较少，不足以训练一个具有良好分类边界的分类器，使用语义特征可对视觉特征进行补充，提升特征多样性，优化分类边界。

（3）与 SP-CLIP 方法对比

针对在 tieredImageNet 数据集上本文方法 SVMSMA 较 SP-CLIP 方法差的现象，本文猜测可能是由于 SP-CLIP 方法预训练网络基于 Transformer 架构，并且图像分辨率为 224×244，而 SVMSMA 使用的预训练模型基于卷积网络架构，且图像分辨率为 84×84。由于 Transformer 模型中的自注意力机制在数据集规模较大时能够更好地捕获全局信息，使其达到较卷积网络更好的结果，导致 SP-CLIP 在 tieredImageNet 数据集上能够达到更好的效果。为了验证此观点，本文使用 SP-CLIP 的预训练网络作为视觉特征提取网络在 miniImageNet、tieredImageNet、CIFAR-FS 数据集上进行了进一步实验，以在相同视觉特征提取网络与语义特征提取网络的条件下公平对比本文方法与 SP-CLIP 方法。

表 4.5 SVMSMA 与 SP-CLIP 的对比实验。最优结果用粗体表示。

Table 4.5 Comparison of SVMSMA and SP-CLIP. The best results are shown in bold.

数据集	方法	特征提取网络	5-way 1-shot	5-way 5-shot
miniImageNet	SP-CLIP	Visformer-T	72.31 ± 0.40	83.42 ± 0.30
	SVMSMA	Visformer-T	73.51 ± 0.38	83.29 ± 0.32
	SVMSMA	ResNet-12	**79.58 ± 0.35**	**85.19 ± 0.29**
tieredImageNet	SP-CLIP	Visformer-T	78.03 ± 0.46	88.55 ± 0.32
	SVMSMA	Visformer-T	**79.48 ± 0.44**	**88.59 ± 0.32**
	SVMSMA	ResNet-12	75.35 ± 0.47	86.35 ± 0.34
CIFAR-FS	SP-CLIP	Visformer-T	82.18 ± 0.40	88.24 ± 0.32
	SVMSMA	Visformer-T	82.44 ± 0.38	88.32 ± 0.33
	SVMSMA	ResNet-12	**86.35 ± 0.35**	**89.42 ± 0.30**

此部分实验比较共包含三个模型，分别是 SP-CLIP、使用 SP-CLIP 预训练网络作为视觉特征提取网络的 SVMSMA（Visformer-T）、以及使用上章方法 MGSRCL 作为视觉特征提取网络的 SVMSMA（ResNet-12），如表4.5所示。实验结果显示，在 5-way 1-shot 任务上，SVMSMA（Visformer-T）在三个数据集上都达到了较 SP-CLIP 方法高的分类准确率，在 miniImageNet、tieredImageNet、CIFAR-FS 数据集上

分别取得了 1.20%、1.45% 和 0.26% 的性能提升。在 5-way 5-shot 任务也达到了和 SP-CLIP 方法相当的性能表现，仅在 miniImageNet 数据集准确率略微下降。这些实验结果证明了本文方法通过简单的映射网络以及损失设计便可有效利用语义信息，提升少样本分类准确率。另外，可以观察到，使用 MGSRCL 作为视觉特征提取网络的模型在 miniImageNet、CIFAR-FS 两个数据集取得了最优结果，这一现象进一步说明了第三章提到的特征学习阶段对于少样本分类的重要性。

4.3.3 消融实验

（1）讨论不同模块对模型性能的影响

为了研究 SVMSMA 模型中每个模块对模型的影响，本文在 miniImageNet、CIFAR-FS 和 CUB 三个数据集上进行了消融实验，此部分实验分为四个模型，分别为基准模型（Baseline），添加跨模态分类（CMC）模块的基准模型（Baseline w/ CMC），添加跨模态特征对齐（CMFA）模块的基准模型（Baseline w/ CMFA），以及最终模型 SVMSMA。此处的基准模型（Baseline）为上一章仅使用视觉信息的多粒度样本关系对比学习（MGSRCL）模型。

表 4.6 SVMSMA 在 miniImageNet、CIFAR-FS 和 CUB 数据集上的模块消融实验。最优结果用粗体表示。

Table 4.6 Module ablation experiments of SVMSMA on miniImageNet, CIFAR-FS and CUB. The best results are shown in bold.

数据集	方法	5-way 1-shot	5-way 5-shot
miniImageNet	Baseline	69.57 ± 0.45	84.41 ± 0.30
	Baseline w/ CMC	76.41 ± 0.39	84.74 ± 0.29
	Baseline w/ CMFA	78.82 ± 0.36	84.82 ± 0.29
	SVMSMA	**79.58 ± 0.35**	**85.19 ± 0.29**
CIFAR-FS	Baseline	78.54 ± 0.47	88.64 ± 0.32
	Baseline w/ CMC	84.18 ± 0.39	89.12 ± 0.31
	Baseline w/ CMFA	85.31 ± 0.38	89.18 ± 0.31
	SVMSMA	**86.35 ± 0.35**	**89.42 ± 0.30**
CUB	Baseline	86.14 ± 0.38	94.75 ± 0.19
	Baseline w/ CMC	90.20 ± 0.30	94.65 ± 0.19
	Baseline w/ CMFA	91.02 ± 0.29	94.65 ± 0.19
	SVMSMA	**91.17 ± 0.28**	**94.91 ± 0.19**

在三个数据集上的模块消融实验如表4.6所示。首先，分别添加 CMC 模块和 CMFA 模块后，模型性能较基准模型来说均有较大程度的提升，尤其是在 5-way 1-shot 少样本分类任务上。具体而言，添加 CMC 模块时，在 miniImageNet、CIFAR-FS

和 CUB 数据集 1-shot 任务准确率分别提升了 6.84%、5.64% 和 4.06%，添加 CMFA 模块时准确率分别提升了 9.25%、6.77% 和 4.88%，并且在 5-shot 任务上也都有小幅提升。这证明了 CMC 模块能够使模型将语义特征映射到视觉空间，并通过分类损失使映射后的特征与实际视觉特征具有一致性；以及 CMFA 模块可以使得映射后特征与类别视觉原型接近，使得样本特征更具有代表性，从而提高了分类准确率。

此外，当同时使用两个模块时，模型在三个数据集都达到了最优结果，5-way 1-shot 任务准确率分别为 79.58%、86.35% 和 91.17%，5-way 5-shot 任务上也同样有小幅度提升。这一现象表明了 CMC 模块和 CMFA 模块之间存在着互补性：CMC 模块提供了一致性的基础，确保了映射后特征在视觉空间保持准确的语义理解；而 CMFA 模块进一步细化了这种理解，通过原型特征对齐使得特征更具代表性，同时优化了类内样本紧密程度。因此，通过同时使用两个模块，可以增强模型对不同类别间细微差异的识别能力，进而提高分类准确率。

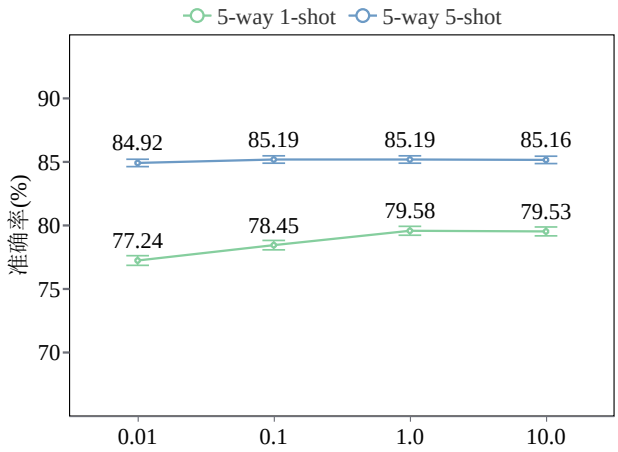


图 4.4 SVMSMA 在 miniImageNet 数据集上的超参数 α 消融实验。

Fig. 4.4 Hyperparameter α ablation experiments of SVMSMA on miniImageNet.

（2）讨论超参数对模型性能的影响

α 是调整不同损失权重占比的超参数，本文通过将其设置为不同数值来讨论 α 对模型的影响，同样在 miniImageNet、CIFAR-FS、CUB 三个数据集进行了实验分析。

如图4.4、4.5和4.6所示，对于 miniImageNet 和 CUB 数据集，无论是 5-way 1-shot 还是 5-way 5-shot 任务，模型均在将 α 设置为 1.0 时达到最优结果，因此对于这两个数据集，α 被设置为 1.0。而对于 CIFAR-FS 数据集，5-way 1-shot 任务在 $\alpha = 10.0$ 时取得最优结果，5-way 5-shot 任务则是在 $\alpha = 1.0$ 取得最优结果，但 5-shot 任务准确率差别很微小，因此在 CIFAR-FS 数据集上，超参数 α 被设置为 10.0。此外，超参数 α 被设置为 0.1、1.0 与 10.0 时，模型结果变化较小，这证明了所提方法对

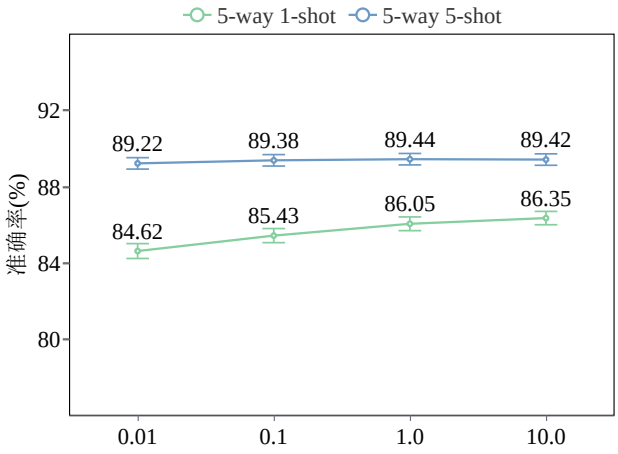


图 4.5 SVMSMA 在 CIFAR-FS 数据集上的超参数 α 消融实验。

Fig. 4.5 Hyperparameter α ablation experiments of SVMSMA on CIFAR-FS.

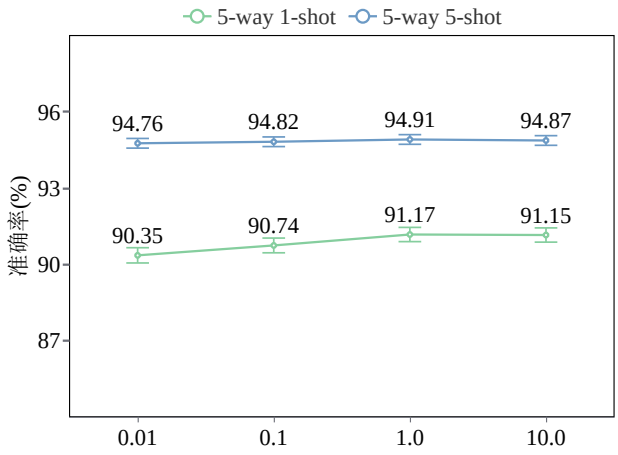


图 4.6 SVMSMA 在 CUB 数据集上的超参数 α 消融实验。

Fig. 4.6 Hyperparameter α ablation experiments of SVMSMA on CUB.

超参数 α 的稳定性。

（3）讨论不同映射模式特征的分类准确率

如4.2.3所述，根据映射网络输入特征模态不同，可将其分为两种模式：单模态映射和多模态映射。为讨论不同映射模式特征对少样本分类结果的影响，本文进行了实验分析。

如表4.7所示，对于 5-way 1-shot 任务，无论是单模态映射特征，还是多模态映射特征，均取得了较视觉特征高的准确率。这证明了利用语义特征可以更好地建立新类与基类之间联系，从而迁移在基类上学习到的知识，也证明了语义信息对少样本分类的重要性。另外，在将三个特征共同作为分类器的训练数据时，模型能够取得进一步的效果提升。对于 5-way 5-shot 任务，相对于视觉特征，单模态映射特征和多模态映射特征的准确率均有所降低，尤其是单模态映射特征，降低幅

表 4.7 SVMSMA 在 miniImageNet、CIFAR-FS 和 CUB 数据集上的不同特征消融实验。最优结果用粗体表示。

Table 4.7 Different features ablation experiments of SVMSMA on miniImageNet, CIFAR-FS and CUB. The best results are shown in bold.

数据集	特征	5-way 1-shot	5-way 5-shot
miniImageNet	视觉特征	69.57 ± 0.45	84.41 ± 0.30
	单模态映射特征	73.73 ± 0.39	73.83 ± 0.39
	多模态映射特征	78.34 ± 0.37	81.29 ± 0.34
	SVMSMA	**79.58 ± 0.35**	**85.19 ± 0.29**
CIFAR-FS	视觉特征	78.54 ± 0.47	88.64 ± 0.32
	单模态映射特征	83.44 ± 0.37	83.60 ± 0.38
	多模态映射特征	86.00 ± 0.36	87.40 ± 0.33
	SVMSMA	**86.35 ± 0.35**	**89.42 ± 0.30**
CUB	视觉特征	86.14 ± 0.38	94.75 ± 0.19
	单模态映射特征	83.58 ± 0.43	83.53 ± 0.43
	多模态映射特征	89.90 ± 0.31	93.60 ± 0.22
	SVMSMA	**91.17 ± 0.28**	**94.91 ± 0.19**

度尤为明显。导致这种现象出现的原因是因为每个类别的语义特征以及视觉原型特征都是固定的，将其映射到视觉空间并执行分类与特征对齐任务会使其得模型学习到一个投影相对固定的映射网络，即映射后特征的多样性会较差。因此即使其更接近类别中心，但由于多样性差会导致其不如使用多样性较好的视觉特征训练出来的分类边界更加鲁棒，导致取得较差的 5-shot 任务结果。虽然多模态映射模式中将视觉特征与语义特征共同输入网络缓解了这种情况，但仍比视觉特征的分类准确率低。本文通过使用三种特征一起训练分类器，利用视觉特征的多样性解决了此问题，取得了较使用单一特征更好的结果。

（4）讨论不同语义特征提取网络对模型性能的影响

为了进一步证明本文方法的可推广性，本文也使用其他模型作为语义特征提取网络：GloVe 与 SBERT，对于这两个模型，本文使用类别名称作为模型的输入，因此表示为 SVMSMA-GloVe (name) 与 SVMSMA-SBERT (name)。

如表4.8所示，在 1-shot 任务上，SVMSMA-GloVe 与 SVMSMA-SBERT 都通过提高样本特征多样性的手段取得了优于基准模型的表现，从而进一步证明了语义特征对于少样本分类问题的重要性以及本文方法的可迁移性。而在 5-shot 任务上，模型性能均有所降低，尤其是在 miniImageNet 数据集上。这是因为这两个模型提取的语义特征迁移性不够好，语义映射特征并不像使用 CLIP 模型时具有代表性，原本较多数量的视觉特征便可以使分类器学习到良好分类边界，而增加了这些样

表 4.8 SVMSMA 在 miniImageNet、CIFAR-FS 和 CUB 数据集上的不同语义特征消融实验。最优结果用粗体表示。

Table 4.8 Different semantic features ablation experiments of SVMSMA on miniImageNet, CIFAR-FS and CUB. The best results are shown in bold.

数据集	方法	5-way 1-shot	5-way 5-shot
miniImageNet	Baseline	69.57 ± 0.45	84.41 ± 0.30
	SVMSMA-GloVe (name)	71.53 ± 0.46	81.49 ± 0.34
	SVMSMA-SBERT (name)	73.29 ± 0.43	82.94 ± 0.32
	SVMSMA-CLIP (name)	78.16 ± 0.37	84.93 ± 0.28
	SVMSMA-CLIP (text)	**79.58 ± 0.35**	**85.19 ± 0.29**
CIFAR-FS	Baseline	78.54 ± 0.47	88.64 ± 0.32
	SVMSMA-GloVe (name)	83.19 ± 0.40	88.25 ± 0.32
	SVMSMA-SBERT (name)	82.24 ± 0.41	88.15 ± 0.31
	SVMSMA-CLIP (name)	85.34 ± 0.37	89.24 ± 0.30
	SVMSMA-CLIP (text)	**86.35 ± 0.35**	**89.42 ± 0.30**
CUB	Baseline	86.14 ± 0.38	94.75 ± 0.19
	SVMSMA-GloVe (name)	87.67 ± 0.35	93.83 ± 0.21
	SVMSMA-SBERT (name)	87.40 ± 0.35	94.20 ± 0.20
	SVMSMA-CLIP (name)	91.06 ± 0.28	**94.91 ± 0.19**
	SVMSMA-CLIP (text)	**91.17 ± 0.28**	**94.91 ± 0.19**

本会对样本特征空间造成一定的干扰，使其偏离原本分类边界，且为向远离真实分类边界的方向偏离，造成性能下降。虽说通过调整各种特征的比例能够缓解或避免这种现象，但为了方法的简易性，对此部分内容本文不再进行描述。

此外，本文也讨论了是否使用提示文本对模型性能的影响，使用类别名称作为 CLIP 文本编码器输入进行了实验。使用类别名称作为输入时，模型表示为 SVMSMA-CLIP (name)，添加提示文本时则表示为 SVMSMA-CLIP (text)，如表4.8所示。可以观察到，虽然 SVMSMA-CLIP (name) 模型也取得了优异的效果，但仍不如添加提示文本后模型表现，在 miniImageNet、CIFAR-FS 和 CUB 三个数据集 1-shot 任务上分别降低了 1.42%、1.01% 和 0.11% 的准确率，在 5-shot 任务也有略微降低。这一实验证明了提示文本对模型表现具有一定影响，也为将来通过设计提示文本或将其参数化进一步提升模型性能提供了依据。

4.3.4 可视化分析

为了更好地展示 SVMSMA 方法能够通过引入语义特征对视觉特征进行补充，本文在 miniImageNet 数据集上随机选取了 5 个新类并使用 t-SNE 对每个类别所有样本的视觉特征，随机选取一个 5-way 1-shot 任务支持集样本的视觉特征（△）、单模态映射特征（□）、多模态映射特征（☆）进行了可视化实验，如图4.7所示。在此

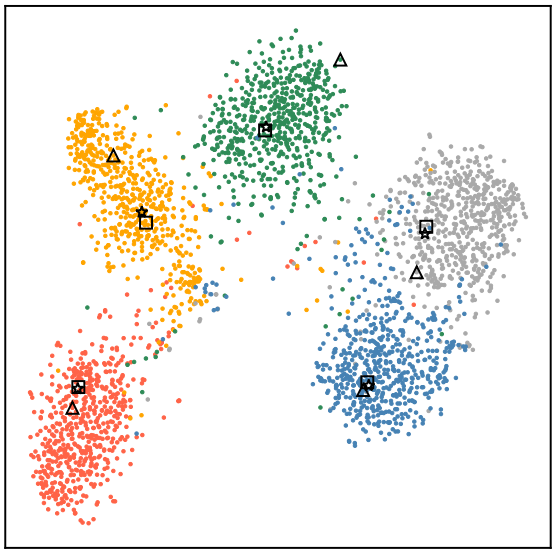


图 4.7 miniImageNet 数据集上不同特征的 t-SNE 可视化结果。不同颜色代表不同类别。视觉特征、单模态映射特征、多模态映射特征分别用 △、□ 和 ☆ 表示。

Fig. 4.7 The t-SNE visualization results of different features on miniImageNet. Different colors represent different categories. visual features, single-modal mapping features, and multi-modal mapping features are represented by △, □, and ☆, respectively.

图里，可以看到无论是单模态特征还是多模态特征都能较好地落在样本簇中，这使得将它们加入逻辑回归器的训练数据可以提升支持集样本特征的多样性，从而在查询集上达到更好的分类效果。另外，在某些较为极端的情况下，如图中绿色类别和灰色类别所示，支持集样本视觉特征处于类别边缘或者分类边界。对于这种情况，如果单独使用视觉特征一般会取得较差的分类结果。而使用本文提出方法所获取的单模态映射特征和多模态映射特征则处于样本簇较为中央的位置，从而能够校正视觉特征造成的偏差，优化分类器学习到的分类边界，达到更好的准确率。综上所述，SVMSMA 方法能够使用语义信息对视觉信息进行补充，通过增加支持集样本的方式丰富支持集样本的多样性，同时对视觉特征处于样本簇边缘时造成的分类边界偏差进行修正，从而提高少样本分类任务的准确率。

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。

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]. Communications of the ACM, 2017, 60(6): 84-90.

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE 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 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] 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.

[7] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//International conference on machine learning. PMLR, 2017: 1126-1135.

[8] 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.

[9] Rusu A A, Rao D, Sygnowski J, et al. Meta-learning with latent embedding optimization[A]. 2018.

[10] 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.

[11] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning[J]. Advances in neural information processing systems, 2017, 30.

[12] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29.

[13] 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.

[14] 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.

[15] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1199-1208.

[16] 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.

[17] 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.

[18] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[19] Tian Y, Wang Y, Krishnan D, et al. Rethinking few-shot image classification: a good embedding is all you need?[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer, 2020: 266-282.

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

[21] Dhillon G S, Chaudhari P, Ravichandran A, et al. A baseline for few-shot image classification [A]. 2019.

[22] 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.

[23] 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.

[24] 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.

[25] Ouali Y, Hudelot C, Tami M. Spatial contrastive learning for few-shot classification[C]//Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21. Springer, 2021: 671-686.

[26] Yang Z, Wang J, Zhu Y. Few-shot classification with contrastive learning[C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX. Springer, 2022: 293-309.

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

[28] 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.

[29] 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.

[30] 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

(EMNLP). 2014: 1532-1543.

[31] 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.

[32] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[33] 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.

[34] 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.

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

[36] 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.

[37] Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification [A]. 2018.

[38] Bertinetto L, Henriques J F, Torr P H, et al. Meta-learning with differentiable closed-form solvers [A]. 2018.

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

[40] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[41] 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.

[42] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 18661-18673.

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

[44] 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 (EMNLP-IJCNLP). 2019: 3982-3992.

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

[46] 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.

[47] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009

IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[48] Ravi S, Larochelle H. Optimization as a model for few-shot learning[C]//International conference on learning representations. 2017.

[49] Triantafillou E, Zemel R, Urtasun R. Few-shot learning through an information retrieval lens[J]. Advances in neural information processing systems, 2017, 30.

[50] Zhu H, Zhao R, Gao Z, et al. Light transformer learning embedding for few-shot classification with task-based enhancement[J]. Applied Intelligence, 2023, 53(7): 7970-7987.

[51] Lee H, Hwang S J, Shin J. Self-supervised label augmentation via input transformations[C]// International Conference on Machine Learning. PMLR, 2020: 5714-5724.

[52] Zhang M, Zhang J, Lu Z, et al. Iept: Instance-level and episode-level pretext tasks for few-shot learning[C]//International Conference on Learning Representations. 2021.

[53] Rong Y, Lu X, Sun Z, et al. Espt: A self-supervised episodic spatial pretext task for improving few-shot learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023.

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

[55] Fuglede B, Topsoe F. Jensen-shannon divergence and hilbert space embedding[C]//International symposium onInformation theory, 2004. ISIT 2004. Proceedings. IEEE, 2004: 31.

[56] Kang D, Kwon H, Min J, et al. Relational embedding for few-shot classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8822-8833.

[57] Afrasiyabi A, Lalonde J F, Gagné C. Associative alignment for few-shot image classification[C]// Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, 2020: 18-35.

[58] Mazumder P, Singh P, Namboodiri V P. Gifsl-grafting based improved few-shot learning[J]. Image and Vision Computing, 2020, 104: 104006.

[59] Fei N, Lu Z, Xiang T, et al. Melr: Meta-learning via modeling episode-level relationships for few-shot learning[C]//International Conference on Learning Representations. 2021.

[60] Liu G, Zhao L, Fang X. Pda: Proxy-based domain adaptation for few-shot image recognition[J]. Image and Vision Computing, 2021, 110: 104164.

[61] Yu T, He S, Song Y Z, et al. Hybrid graph neural networks for few-shot learning[C]//Proceedings of the AAAI conference on artificial intelligence: Vol. 36. 2022: 3179-3187.

[62] Ma R, Fang P, Drummond T, et al. Adaptive poincaré point to set distance for few-shot classification[C]//Proceedings of the AAAI conference on artificial intelligence: Vol. 36. 2022: 1926-1934.

[63] Cui Z, Lu N, Wang W, et al. Dual global-aware propagation for few-shot learning[J]. Image and Vision Computing, 2022, 128: 104574.

[64] Zhang L, Zhou F, Wei W, et al. Meta-hallucinating prototype for few-shot learning promotion [J]. Pattern Recognition, 2023, 136: 109235.

[65] Huang X, Choi S H. Sapenet: Self-attention based prototype enhancement network for few-shot

learning[J]. Pattern Recognition, 2023, 135: 109170.

[66] Cheng H, Wang Y, Li H, et al. Disentangled feature representation for few-shot image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[67] Zheng K, Zhang H, Huang W. Diffkendall: A novel approach for few-shot learning with differentiable kendall's rank correlation[J]. Advances in Neural Information Processing Systems, 2023, 36: 49403-49415.

[68] Ye H J, Hu H, Zhan D C, et al. Few-shot learning via embedding adaptation with set-to-set functions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8808-8817.

[69] Chen Y, Liu Z, Xu H, et al. Meta-baseline: Exploring simple meta-learning for few-shot learning [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9062-9071.

[70] Xu W, Xu Y, Wang H, et al. Attentional constellation nets for few-shot learning[C]//International Conference on Learning Representations. 2021.

[71] Wu J, Chang D, Sain A, et al. Bi-directional feature reconstruction network for fine-grained few-shot image classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 2821-2829.

[72] Akata Z, Perronnin F, Harchaoui Z, et al. Label-embedding for attribute-based classification[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 819-826.

[73] Niu L, Cai J, Veeraraghavan A, et al. Zero-shot learning via category-specific visual-semantic mapping and label refinement[J]. IEEE Transactions on Image Processing, 2018, 28(2): 965-979.

[74] Song J, Shen C, Yang Y, et al. Transductive unbiased embedding for zero-shot learning[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1024-1033.

[75] Xian Y, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5542-5551.

[76] Vyas M R, Venkateswara H, Panchanathan S. Leveraging seen and unseen semantic relationships for generative zero-shot learning[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. Springer, 2020: 70-86.

[77] Zhao X, Shen Y, Wang S, et al. Boosting generative zero-shot learning by synthesizing diverse features with attribute augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 36. 2022: 3454-3462.

[78] 张鲁宁, 左信, 刘建伟. 零样本学习研究进展[J/OL]. 自动化学报, 2020, 46: 1-23. DOI: 10.16383/j.aas.c180429.

[79] Peng Z, Li Z, Zhang J, et al. Few-shot image recognition with knowledge transfer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 441-449.

[80] Li A, Huang W, Lan X, et al. Boosting few-shot learning with adaptive margin loss[C]//

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 12576-12584.

[81] Yan K, Bouraoui Z, Wang P, et al. Aligning visual prototypes with bert embeddings for few-shot learning[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. 2021: 367-375.

[82] Liu S, Xie Y, Yuan W, et al. Cross-modality graph neural network for few-shot learning[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[83] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2020, 32(2): 349-369.

[84] 葛轶洲, 刘恒, 王言, 等. 小样本困境下的深度学习图像识别综述[J/OL]. 软件学报, 2022, 33: 193-210. DOI: 10.13328/j.cnki.jos.006342.

[85] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[86] Zhang B, Luo C, Yu D, et al. Metadiff: Meta-learning with conditional diffusion for few-shot learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024.

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