第 48 卷　第 3 期
2025 年 3 月

计　算　机　学　报
CHINESE JOURNAL OF COMPUTERS

Vol. 48　No. 3
March 2025

基于自监督学习和二阶表示的小样本图像分类

李兆亮[1)]　贾令尧[1)]　张冰冰[2)]　李培华[1)]

[1)]（大连理工大学信息与通信工程学院　辽宁 大连　116024）

[2)]（大连民族大学计算机科学与工程学院　辽宁 大连　116650）

摘　要　小样本图像分类旨在利用少量的标注样本实现对未见类别的预测。最近的研究表明，预训练策略和图像表示方法在该任务中发挥着关键作用。然而，这些方法的应用仍面临两个主要挑战：第一，自监督学习在小样本分类的预训练阶段尚未得到充分的探索；第二，二阶表示在不同粒度的小样本任务中的作用尚不明确，制约了其在复杂任务中的应用。针对上述问题，本文首先提出了一个多任务协同优化的预训练方法，实现了对比式自监督、生成式自监督和有监督学习的联合训练。该方法旨在促进模型学习具有迁移性的特征，从而提升模型的泛化性能。其次，本文利用紧致的双线性池化对模型进行微调，以获取更具分辨力的二阶表示，从而进一步增强模型的非线性建模能力。最后，本文提出了一种基于类间相似关系的任务难度指标，用于量化小样本任务的分类粒度，并通过线性探测分析系统地研究了二阶表示在粗细粒度不同的小样本任务中的表现。实验表明，多任务协同的预训练有效提高了模型的泛化性能，并且不同的分支任务呈现相互促进的效果；在更加困难的细粒度任务中，二阶表示相对于一阶表示展现出更强的线性可分性，这为一阶和二阶表示在不同场景中的应用提供了有益参考。本文通过广泛的消融实验深入评估了每个关键设计的贡献。与当前最先进的方法相比，本文方法在 miniImageNet 和 CUB 数据集的 1-shot/5-shot 分类任务中分别取得 0.66%/0.53%和 3.12%/0.98%的提升，在 tieredImageNet 数据集的 5-shot 分类任务中取得可比结果（87.19% vs. 87.31%），在跨域数据集 miniImageNet→CUB、miniImageNet→Aircraft 和 miniImageNet→Cars 中分别取得 1.25%、1.96%和 4.34%的提升，验证了本文方法的有效性。

关键词　小样本图像分类；自监督学习；有监督学习；二阶表示；任务难度指标

中图法分类号　TP391　　　**DOI 号**　10.11897/SP.J.1016.2025.00586

Few-Shot Image Classification Based on Self-Supervised Learning and Second-Order Representation

LI Zhao-Liang[1)]　JIA Ling-Yao[1)]　ZHANG Bing-Bing[2)]　LI Pei-Hua[1)]

[1)]（School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning　116024）

[2)]（School of Computer Science and Engineering, Dalian Minzu University, Dalian, Liaoning　116650）

Abstract　Few-shot image classification aims to accurately recognize unseen categories with a few annotated samples. Recent studies have shown that both pre-training strategies and effective image representation methods play crucial roles in this task. However, these methods still face two main challenges in their application. Firstly, self-supervised learning is underexplored in the pre-training stage of few-shot classification task. Secondly, although second-order representation has made significant progress in few-shot image classification tasks, its role in tasks with varying granularities remains unclear, limiting its application in complex tasks. To address the above issues, we first propose a multi-task co-optimization pre-training method. It is based on the contrastive learning framework that encompasses the joint training of contrastive self-supervised

收稿日期：2024-06-29；在线发布日期：2024-12-17。本课题得到国家自然科学基金（62471083）资助。**李兆亮**，硕士，主要研究方向为小样本图像分类和自监督学习。E-mail：upclzl@163.com。**贾令尧**，博士研究生，主要研究方向为图像分类和深度学习。**张冰冰**，博士，讲师，主要研究方向为视频行为识别、图像分类和深度学习。**李培华**（通信作者），博士，教授，博士生导师，中国计算机学会（CCF）高级会员，主要研究领域为图像/视频识别、目标检测和语义分割。E-mail：peihuali@dlut.edu.cn。

learning, generative self-supervised learning and supervised learning. It encourages the model to adapt to diverse training pretext tasks, aiming to promote the model to learn transferable representation, thereby improving the model's generalization performance on downstream tasks. The proposed generative task directly operates on the feature maps of the input without the need for image masks, thus seamlessly facilitating compatibility with contrastive learning pretext tasks. Secondly, we fine-tune the model using compact bilinear pooling (CBP) to obtain more discriminative second-order representations, thereby enhancing the model's nonlinear modeling capability. Finally, we propose a task difficulty metric based on inter-class similarity to measure the granularity of few-shot classification tasks, providing a quantitative assessment for complex tasks. Through linear probing analysis, this paper quantitatively investigates the performance of second-order representations in few-shot tasks with different levels of granularity, and verifies their applicability and advantages in diverse scenarios. Specifically, we freeze the fine-tuned model and directly extract the second-order representations of images from unseen categories. Subsequently, we conduct meta-testing using linear probing to evaluate the performance of these second-order representations in few-shot image classification tasks. The granularity differences in few-shot image classification tasks are reflected in the varying inter-class similarities across tasks, which determine the nonlinearity of the decision boundaries. We employ the task difficulty metric to quantify this variation. By comparing the classification accuracies of first-and second-order representations across tasks of varying difficulty, we demonstrate the advantages of second-order representations. Experimental results show that the multi-task co-optimization pre-training method effectively improves the generalization performance of the model, and different branch tasks exhibit a mutually reinforcing effect. In more difficult fine-grained tasks, second-order representations exhibit stronger linear separability compared to first-order ones, providing valuable insights for the application of first-and second-order representations in different scenarios. We thoroughly evaluate the contribution of each key design by extensive ablation experiments. Compared to the state-of-the-art methods, our method improves performance by 0.66%/0.53% and 3.12%/0.98% on the 1-shot/5-shot classification tasks of miniImageNet and CUB datasets, respectively, and achieves comparable accuracy (87.19% vs. 87.31%) on tieredImageNet. In the cross-domain dataset miniImageNet→CUB, miniImageNet→Aircraft, and miniImageNet→Cars, our method also achieves improvement of 1.25%, 1.96% and 4.34%. The compelling results on the six standard few-shot image benchmarks verify the effectiveness of the proposed method.

Keywords few-shot image classification; self-supervised learning; supervised learning; second-order representation; task difficulty metric

1 引 言

得益于大规模的标注数据，深度学习在各种视觉任务上取得了令人印象深刻的结果[1-4]。然而，在实际应用中，获取图像及其标签可能是困难和昂贵的，例如：一些生物在野外可能很少见，病虫害图像需要专家级的标注。深度学习算法在这些标注缺乏的场景下面临泛化性不足的挑战。受到人类能够利用少量样本快速掌握新概念的启发，学者们提出了小样本分类任务以应对上述挑战。该任务的关键在于如何从基类数据中学习可推广到新类、分辨力强的特征表示。为此，研究者们从训练范式和表征设计两个主要方面展开探索。

在训练范式方面，目前主流的小样本学习方法通常包括预训练和元学习两个阶段[5-14]。在基类数据上的预训练被用于提高元学习阶段模型的表达能力。最初，多数工作采用常规的有监督分类任务进

行预训练。但这类方法可能导致模型对基类数据的过拟合，从而降低其对未知类别的泛化性能[12]。自监督学习利用数据自身的结构属性来构造监督信号[15]，从而实现模型的自监督训练。这种预训练范式使得模型能够捕捉到对下游任务更有帮助的语义表示。许多不同的视觉代理任务被提出用于学习这种可迁移的图像表示。其中，对比学习[15-18]和掩码图像建模[19-22]展现了优越的学习能力和可扩展性。然而，基于自监督学习的预训练范式在小样本分类领域尚未得到充分的探索。具体而言，PSST 等方法[23-25]虽然在元学习阶段设计了角度预测、图像区域顺序预测等自监督任务来辅助训练，但它们并没有探讨预训练机制对模型性能的影响。相比之下，CVET[12]设计了多种对比损失进行联合训练，在一定程度上改善了预训练模型的泛化效果。HCT[11]使用对比式自监督和词元监督任务对视觉变换器(Vision Transformer，ViT[3])进行预训练，有效缓解了模型的过拟合问题。FewTURE[13]和 CPEA[14]采用掩码图像建模和对比学习对 ViT 进行预训练，取得了良好的泛化效果。尽管如此，如何在预训练阶段更有效地利用自监督学习提高模型表征的泛化性，仍是一个需要深入探索的问题。

在表征设计方面，当前方法聚焦于获取具有分辨力的图像表示，并结合特性不同的测度提高小样本分类的鲁棒性。通常的方法是使用神经网络提取图像的一阶特征，包括图像的特征图和均值池化向量，并结合欧式距离、余弦距离、推土距离等测度[5,7-9,26]进行度量学习。相比之下，为了充分利用特征中丰富的统计信息，CovNet[6]、ADM[27]和 Deep-BDC[10]等方法提取图像的二阶表示用于相似度判别，并取得了领先的性能。然而，在不同的小样本分类任务中，由于图像的类间相似关系各异，带来的分类粒度和识别难度互不相同。二阶表示在不同粒度的小样本分类任务中的具体优势仍需进一步探究和展示。

针对上述两个问题，本文首先提出了一个多任务协同优化的预训练方法。它基于对比学习框架，在卷积网络中实现了对比式自监督、生成式自监督和有监督分类任务的联合训练，用以增强模型对未见类别的泛化性能。具体而言，本文通过数据增广获得一张图像的全局视角和局部视角，并通过不同的网络分支进行编码从而实现多任务学习。对比式自监督任务在不同增广视角的表达向量之间计算对比损失，旨在促使模型捕获不同增广视角之间的联系，用于学习图像的全局上下文信息。对于每个全局视角的特征图，生成式自监督任务利用其余增广视角的特征图进行对齐和重建，从而使模型拥有利用有限的像素信息预测图像局部细节的能力。有监督学习对所有增广视角进行类别判别，计算交叉熵损失以辅助模型的训练。相关研究[11-12]表明，在自监督学习中利用标签信息可以引导模型学习任务相关的特征，从而提高模型的分类表现。

其次，本文量化分析了二阶表示在粗细粒度不同的小样本任务中的表现。具体而言，为了获取图像的二阶表示以评估其在小样本分类中的作用，本文利用紧致的双线性池化(Compact Bilinear Pooling，CBP)[28]代替全局平均池化(Global Average Pooling，GAP)对预训练模型进行微调。二阶微调用于进一步提升模型的非线性建模和细节感知能力。本文利用微调后的模型聚合图像的二阶表示，并采用线性探测(Linear Probing)的方案在未见类别上进行小样本测试和详细的分析。在分类粒度不同的小样本任务中，图像的类间相似关系存在差异，这决定了判别边界的非线性程度。本文通过特征可视化技术来展现这种分类粒度的变化，并定义了一种任务难度指标来量化不同任务之间分类粒度的差异。通过比较一阶表示和二阶表示在不同难度任务上的分类表现，明确了二阶表示的优势场景。本文的主要贡献归纳如下：

(1) 本文提出了一种多任务协同优化的预训练方法用于增强模型对下游任务的泛化性能。该方法在对比学习框架中实现了对比式自监督、生成式自监督和有监督学习的联合训练，并可以自然地与其他两阶段的方法相结合。

(2) 本文基于图像的类间相似关系定义了一种任务难度指标，用于量化不同粒度的小样本任务的分类难度。利用 CBP 对预训练模型进行微调，以获取图像的二阶表示。基于线性探测的小样本分类范式，本文详细分析了一阶和二阶表示在不同粒度的小样本任务中的具体表现。实验表明，二阶表示在高难度任务上具有更强的线性可分性。

(3) 本文进行了广泛的消融实验以深入理解所提出方法的工作原理，并在 5 个挑战性的数据集上实现了当前最佳的结果。

2 相关工作

2.1 小样本图像分类

小样本图像分类在现实世界中拥有广阔的应用

前景。现有的方法可以概括地分为以下两类。基于优化的方法在基类上设计了一种良好的模型初始化策略[29-30]。随后，只需对少量的标注数据进行梯度优化，模型即可快速适应新的类别。基于度量的方法旨在学习一个泛化性的表示空间，并设计相似度比较函数实现分类。这类方法展现出巨大的前景，并实现了最先进的性能。现有的工作考虑了不同的度量函数，例如：欧氏距离[26]、余弦距离[5,11]、推土距离[8]、KL 散度[27]以及参数化距离[5,7,10-11,13-14, 31-32]等。通常的方法使用图像的一阶表示作为度量函数的输入。一些工作利用协方差池化[6,10,27]等计算图像更具分辨力的二阶表示，有效地提高了度量算法的鲁棒性。最近的研究均在基类数据上进行预训练，用以提升特征提取器在元学习阶段的表征效果[5-14]。与仅在基类上进行元训练的方法相比，预训练的引入显著地提高了相关方法的分类性能。这些进展表明，精心设计的预训练范式和图像表示能够有效地提升度量方法的性能。

与前人的工作不同，本文将基类上的训练分为两个阶段。首先，本文设计了一种多任务协同优化的预训练策略，用于学习可迁移的图像表示。随后，本文利用 CBP 对预训练模型进行微调，进一步提升模型的细节感知能力。最后，本文利用微调后的模型提取图像的二阶表示直接在新类上进行元测试。

2.2　小样本图像分类与自监督学习

自监督学习使用数据的内在结构构建监督信号用于监督模型训练[15]。因此，许多工作借助自监督学习来提升模型在新类上的泛化性能。最初，旋转预测、图像区域顺序预测等[23-25]自监督代理任务被作为一个辅助分支参与元训练。CTX[33]成功地将对比学习集成到元学习中。CVET[12]设计了多种对比损失与分类损失进行预训练，并在元学习阶段引入对比学习，提高了小样本任务的基准。HCT[11]使用对比学习和词元监督任务对视觉变换器进行预训练，有效地改善了模型的泛化性能。FewTURE 等[13-14]证明掩码图像建模和对比学习能够在基类数据上独立地训练 Transformer，用以学习可迁移的图像表示，并取得了领先的结果。

与之不同，本文从自监督学习和有监督学习任务协同优化的角度出发，设计了一种新颖的预训练方法。该方法基于对比学习的机制，实现了对比式自监督任务、生成式自监督任务和有监督任务的联合训练，旨在提升模型的泛化性能。

2.3　细粒度图像分类中的高阶表示

以双线性/协方差池化为代表的高阶表征学习[28,34-36]表明，捕获图像的局部细节特征对细粒度图像分类是至关重要的。这类方法对特征图中每个空间位置的特征向量建模通道相关性，随后聚合所有的相关性矩阵以获得图像的高阶表示。这些工作显式地给出了 k 阶多项式特征映射($k\geq 2$)、协方差池化等计算形式，为获取更具分辨力的高阶表示提供了选择。本文利用无参的 CBP 算法聚合图像的二阶表示，进而分析二阶表示在小样本分类任务中的具体优势。

3　方　法

3.1　问题定义

小样本图像分类旨在将模型在基类 C_{base} 中学习的知识迁移到新类 C_{novel} 上，基类和新类数据集的类别互不相同，$C_{\text{base}}\cap C_{\text{novel}}=\varnothing$。一个小样本分类任务被称为一个“$N$-way K-shot”问题。一个任务包含 N 个类别，每个类有 K 张支撑图像和 Q 张查询图像。所有的支撑图像构成该任务的支撑集，记为 $D^s=\{x_i,y_i\}_{i=1}^{NK}$；所有的查询图像构成该任务的查询集，记为：$D^q=\{x_i,y_i\}_{i=1}^{NQ}$；$y_i$ 是图像 x_i 的类别标签。模型利用少量的支撑图像学习类别概念，并在查询集上评估模型对该任务的分类性能。在基类上构造的任务可用于元训练，而在新类上构造的任务被用于元测试。小样本分类的目标是在基类上学习一个模型，使其能在若干随机的新类任务上实现良好的预测。

3.2　方法概述

本文基于迁移学习的范式，骨干模型的训练分为预训练和二阶微调两个阶段，随后直接进行元测试，整体方法如图 1 所示。预训练基于 DINO[18]的对比学习框架，学生网络分支以下标 s 标记，教师网络分支以下标 t 标记。对图像 x 进行多视角数据增广，以获得其增广集合 $V_1=\{x_1^g,\cdots,x_n^g,x_1^l,\cdots,x_m^l\}$ 和 $V_2=\{x_1^g,\cdots,x_n^g\}$。裁剪原始图像较大区域的增广图像称为全局视角，用上标 g 标记，共有 n 张；较小区域的增广图像称为局部视角，用上标 l 标记，共有 m 张。本文分别利用视觉编码器 f_s 和 f_t 提取 V_1 和 V_2 中图像的特征图，其高、宽和通道数分别为 H、W 和 C。然后通过不同的网络分支进行变换和编码，从而实现多任务学习，用以获取可迁移的图像表示。

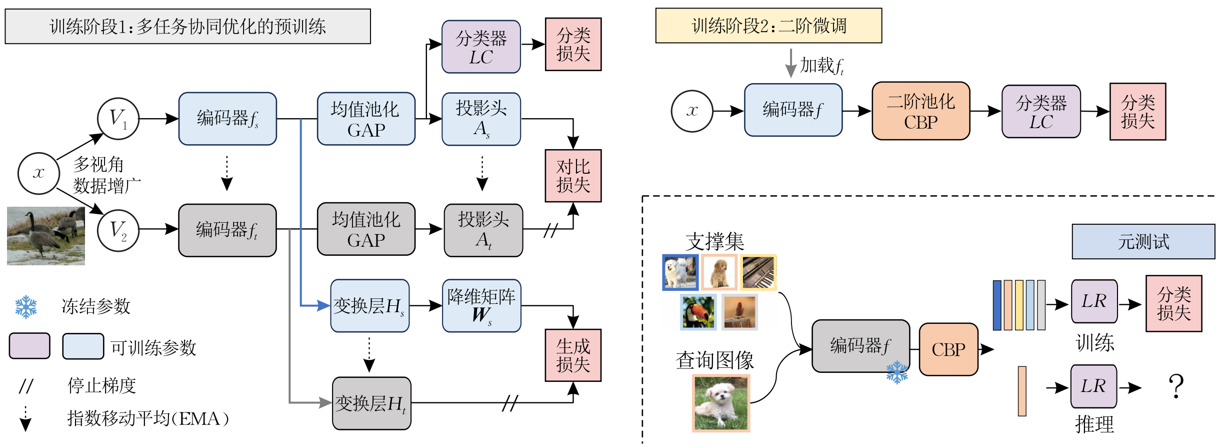


图 1 本文方法概览

在预训练框架中，A_s、A_t、H_s 和 H_t 是多层感知机，其设计思想与文献[17-18]相同，用于提取对数据增广不变的特征，进而计算自监督损失。其中 A_s 和 A_t 由两部分组成，先是一个隐藏层维度为 1024、输出维度为16 384的多层感知机(Multi-Layer Perceptron, MLP)；随后由 $L2$ 二范数归一化和输出维度为 256 的线性层组成。H_s 和 H_t 是一个 2 层的 MLP，对每个空间位置的特征向量进行变换。它的隐藏层维度为 256，输入和输出维度相同，最后对输出特征进行 $L2$ 二范数归一化。LC 是一个线性层，参与分类损失的计算。教师网络参数 θ_t 使用学生网络参数 θ_s 的指数移动平均(Exponential Moving Average, EMA)进行更新：$\theta_t=\lambda\theta_t+(1-\lambda)\theta_s$。特别地，本文设计的生成式自监督任务可以自然地融入到基于卷积网络的对比学习框架中，进而使模型具备局部信息预测的能力。

其次，本文利用 CBP 聚合图像的二阶表示，对预训练模型进行端到端的微调。二阶微调用于进一步捕获图像中的细节信息以增强图像表示的分辨力，从而提升模型的线性判别能力。

最后，在元测试阶段，本文冻结编码器并提取图像的二阶表示，在支撑集上训练任务专属的逻辑回归分类器，简记为 LR，并预测查询样本。为了衡量每个小样本任务的分类粒度，本文定义了一种任务难度指标 D_{score}。基于线性探测的元测试方案，本文对一阶和二阶表示在不同粒度的小样本任务中的分类表现进行了详细探究。

3.3 预训练

多任务协同优化的预训练旨在获得更一般的图像表示，其主要包含三个训练分支。对比式自监督任务通过“实例判别”理解图像的全局上下文信息。生成式自监督任务通过“特征重建”增强模型对未知场景的预测能力。分类任务学习图像中的类别概念，用于加速自监督任务的收敛。

(1) 对比损失

对比学习损失旨在缩小同一张图像不同增广视角(正样本对)之间的距离。通过全局视角和局部视角之间的信息交互，提升模型对图像整体的理解能力。本文利用两个网络分支 $g^{ctr}_{\theta_s}=A_s\circ\mathrm{GAP}\circ f_s$ 和 $g^{ctr}_{\theta_t}=A_t\circ\mathrm{GAP}\circ f_t$，分别提取图像 x 每个增广视角 $\tilde{x}$ 的 E 维的表示向量 $\boldsymbol{g}^{ctr}_{\theta_s}(\tilde{x})$ 和 $\boldsymbol{g}^{ctr}_{\theta_t}(\tilde{x})$，“$\circ$”代表复合函数。随后，对每个表示向量使用带温度和中心化的 softmax[18] 进行归一化，获得对应增广视角 E 维的概率分布向量 $\boldsymbol{p}_s(\tilde{x})$ 和 $\boldsymbol{p}_t(\tilde{x})$。上述过程的定义如下：

$$\boldsymbol{p}_s(\tilde{x})=\mathrm{softmax}(\boldsymbol{g}^{ctr}_{\theta_s}(\tilde{x}),\tau_s),\ \tilde{x}\in V_1 \quad (1)$$

$$\boldsymbol{p}_t(\tilde{x})=\mathrm{softmax}(\boldsymbol{g}^{ctr}_{\theta_t}(\tilde{x})-\boldsymbol{c},\tau_t),\ \tilde{x}\in V_2 \quad (2)$$

其中，τ_s 和 τ_t 是温度系数；$\boldsymbol{c}\in R^E$ 是中心化向量，在每个批次中通过 EMA 机制进行更新。这些操作用于获取差异化的概率分布向量 $\boldsymbol{p}_s(\tilde{x})$ 和 $\boldsymbol{p}_t(\tilde{x})$，以避免监督崩溃。

本文利用交叉熵损失来对齐样本 x 不同增广视角之间的概率分布，从而促进模型学习图像的全局上下文信息。对比损失的定义如下：

$$L_{ctr}(x)=-\frac{1}{R_1}\sum_{x'\in V_1}\sum_{\substack{x''\in V_2\\ x''\neq x'}}\boldsymbol{p}_t(x'')\log\boldsymbol{p}_s(x') \quad (3)$$

其中，$R_1=n(n+m-1)$。

(2) 生成损失

基于 ViT 的小样本图像分类方法利用掩码图像建模和对比式自监督[20]进行预训练，实现了领先的性能。然而，在卷积架构下，现有的对比式自监督[15-18]和掩码图像建模[19-22]方法却不能很好地兼容。为此，本文基于对比学习框架设计了一种生成式任务，用以实现两种自监督任务的优势互补。所提出

的生成式任务分支直接在特征图层面进行操作，而无需进行图像掩码，这能够增加本文方法的适配性。

本文提取图像中某个区域 x_i^g 的特征图作为重建目标 $\boldsymbol{F}_i^T$。随后，在图像中提取若干其他采样区域的特征图，并组成一个元素池，记作 $\boldsymbol{F}_i^G$。$\boldsymbol{F}_i^G$ 中可能描述了目标的背部、尾巴、足部、生活环境等信息。利用可学习矩阵 $\boldsymbol{W}_s$ 对元素池中的特征数量进行降维，从而聚合出若干典型的图像特征用于生成未知图像区域的特征 $\boldsymbol{F}_i^T$，具体过程如图 2 所示。这种生成式的任务旨在增强模型对图像局部的理解。即使仅接收到有限的图像信息，模型也能利用学习到的生成能力来预测和丰富目标的语义表示。生成式自监督任务在特征空间中计算生成特征与重建目标之间均方误差作为训练损失，定义如下：

$$L_{gen}(x)=\frac{1}{nr^g}\sum_{i=1}^{n}\|\boldsymbol{F}_i^T-\boldsymbol{W}_s\boldsymbol{F}_i^G\|_2^2 \tag{4}$$

其中，$r^g=H^gW^g$ 和 $r^l=H^lW^l$ 分别是图像全局视角和局部视角特征图的分辨率。$\boldsymbol{W}_s$ 是一个线性层，用于对 $\boldsymbol{F}_i^G$ 的特征数量降维，实现信息的重组和生成。它的输入维度是 $I=(n-1)r^g+mr^l$，输出维度是 r^g。$\boldsymbol{F}_i^T\in R^{r^g\times C}$ 和 $\boldsymbol{F}_i^G\in R^{I\times C}$ 的定义如下：

$$\boldsymbol{F}_i^T=\mathrm{reshape}(g_{\theta_t}^T(x_i^g)) \tag{5}$$

$$\boldsymbol{F}_i^G=\mathrm{cat}(\boldsymbol{F}_{i,1}^G,\cdots,\boldsymbol{F}_{i,j}^G,\cdots,\boldsymbol{F}_{i,n+m-1}^G) \tag{6}$$

$$\boldsymbol{F}_{i,j}^G=\mathrm{reshape}(g_{\theta_s}^G(\tilde{x}_j)),\ \tilde{x}_j\in V_1/x_i^g \tag{7}$$

其中，$g_{\theta_t}^T=H_t\circ f_t$，$g_{\theta_s}^G=H_s\circ f_s$。reshape(·)在特征图的空间维度上进行操作，cat(·)是拼接操作。

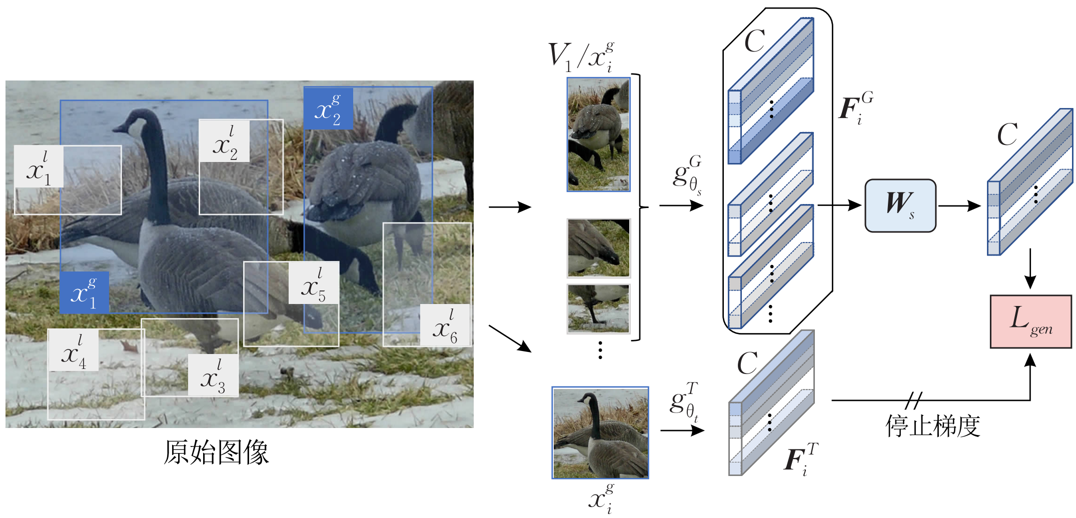


图 2　生成式代理任务的流程图

(1) 分类损失

在小样本图像分类中，基类的数据规模往往不大。为了加快自监督任务的收敛速度，同时增强模型的类别判别能力，本文采用有监督分类任务辅助训练。其损失函数定义如下：

$$L_{CE}(x)=-\frac{1}{R_2}\sum_{\tilde{x}\in V_1}\log\frac{\exp(g_{\theta_s}^{cls}(\tilde{x})^{(y(x))})}{\sum_{i=1}^{M}\exp(g_{\theta_s}^{cls}(\tilde{x})^{(i)})} \tag{8}$$

其中，$y(x)$ 是图像 x 的类别标签，M 是基类的类别总数，$g_{\theta_s}^{cls}=LC\circ\mathrm{GAP}\circ f_s$，$R_2=n+m$。

(2) 训练目标

损失权重 α、β、γ 用于平衡不同损失的影响。在实践中，本文调节损失权重以使不同的损失处在同一数量级。预训练目标定义如下：

$$L(x)=\alpha L_{ctr}(x)+\beta L_{gen}(x)+\gamma L_{CE}(x) \tag{9}$$

其中，预训练的流程如算法 1 所示。为了直观地介绍算法 1 和算法 2，前向和反向过程仅以一个图像样本举例，在实现上仍然以批次的形式进行训练。

算法 1.　预训练算法流程

输入：训练图像 x，模型参数 f_s、f_t，投影头参数 A_s、A_t，重建分支参数 H_s、H_t、$\boldsymbol{W}_s$，分类器参数 LC

输出：优化后的 f_t

FOR x in DataLoader：

　$V_1,V_2=aug(x)$　//多视角增广

　$feat_s=[f_s(\tilde{x}_1),\cdots,f_s(\tilde{x}_i),\cdots,f_s(\tilde{x}_{n+m})]$，$\forall\tilde{x}_i\in V_1$

　$feat_t=[f_t(\tilde{x}_1),\cdots,f_t(\tilde{x}_i),\cdots,f_t(\tilde{x}_n)]$，$\forall\tilde{x}_i\in V_2$

　$L_{ctr}(x)=Ctr\ Loss(A_t\circ\mathrm{GAP}(feat_t),A_s\circ\mathrm{GAP}(feat_s))$
　　//对比损失，式(1)～(3)

　$L_{gen}(x)=Gen\ Loss(feat_t,feat_s)$
　　//生成损失，式(4)～(7)

　$L_{CE}(x)=CE\ Loss(\mathrm{GAP}(feat_s),label(x))$
　　//有监督损失，式(8)

　$L(x)=\alpha L_{ctr}(x)+\beta L_{gen}(x)+\gamma L_{CE}(x)$　//式(9)

　梯度反向传播，参数更新：f_s、A_s、H_s、$\boldsymbol{W}_s$、LC

　指数移动平均，更新教师网络参数
　　$\theta_t=\lambda\theta_t+(1-\lambda)\theta_s$，$\theta\in\{f,A,H\}$

END FOR

3.4 图像的二阶表示

本文利用 CBP 对图像每个空间位置的特征向量计算近似的二次多项式映射，聚合所有的二阶向量得到图像更具分辨力的二阶表示。二阶微调用于进一步提升模型非线性建模的能力。

(1) 向量的二阶映射

张量草图(Tensor Sketch, TS)[37]算法用于计算向量的近似的二次多项式映射 $\phi_{\mathrm{TS}}(\cdot)$。TS 算法能够仅以 $O(D\log D)$ 的计算复杂度得到向量 $\boldsymbol{X}_{ij}\in R^{C}, i=1,\cdots,H; j=1,\cdots,W$ 近似的二次多项式映射向量，且无需可学习的参数。因此，基于 TS 算法的 CBP 具备良好的迁移性，其具体步骤如下：

步骤 1. 哈希映射

哈希函数 $h_1, h_2:[C]\rightarrow[D]$ 将 C 维索引向量的每个分量以均匀分布映射到 D 维索引向量的每个分量中。$s_1, s_2:[C]\rightarrow\{+1,-1\}$ 将 C 维的符号向量的每个分量随机映射为 $+1$ 或 -1。哈希映射将特征向量 $\boldsymbol{X}_{ij}$ 由 C 维映射到 D 维：

$$\boldsymbol{P}_{ij}=\mathrm{Hash}(\boldsymbol{X}_{ij}, h_1, s_1)=[\bar{p}_1,\cdots,\bar{p}_D]^{\mathrm{T}} \tag{10}$$

$$\bar{p}_d=\sum_{c: h_1(c)==d} s_1(c)\boldsymbol{X}_{ij}(c) \tag{11}$$

其中，$c=1,\cdots,C; d=1,\cdots,D; D>C$。同理可得 $\boldsymbol{Q}_{ij}=\mathrm{Hash}(\boldsymbol{X}_{ij}, h_2, s_2)$。

步骤 2. 近似二次多项式映射

$$\phi_{\mathrm{TS}}(\boldsymbol{X}_{ij})=\mathrm{FFT}^{-1}(\mathrm{FFT}(\boldsymbol{P}_{ij})\otimes\mathrm{FFT}(\boldsymbol{Q}_{ij})) \tag{12}$$

其中，$\mathrm{FFT}(\cdot)$ 和 $\mathrm{FFT}^{-1}(\cdot)$ 是快速傅里叶变换及其逆变换，$\otimes$ 代表哈达玛积。$\phi_{\mathrm{TS}}(\boldsymbol{X}_{ij})$ 是二次多项式映射 $\boldsymbol{X}_{ij}^{(2)}=\mathrm{Vec}(\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^{\mathrm{T}})$ 的一种近似。当 $D\rightarrow C^2$ 时，$\phi_{\mathrm{TS}}(\boldsymbol{X}_{ij})$ 的表达能力增强，$E[\langle\phi_{\mathrm{TS}}(\boldsymbol{X}_{ij}),\phi_{\mathrm{TS}}(\boldsymbol{Y}_{ij})\rangle]\approx\langle\boldsymbol{X}_{ij}^{(2)},\boldsymbol{Y}_{ij}^{(2)}\rangle$。$E[\cdot]$ 代表数学期望，$\langle\cdot,\cdot\rangle$ 是向量内积。

(2) 图像的二阶表示

利用 CBP 聚合图像特征图每个空间位置的二阶向量，得到图像的二阶表示 $\bar{z}$：

$$z=\sum_{i=1}^{H}\sum_{j=1}^{W}\phi_{\mathrm{TS}}(\boldsymbol{X}_{ij}) \tag{13}$$

$$\bar{z}=L2(\mathrm{sign}(z)\sqrt{|z|}) \tag{14}$$

其中，平方根 $\sqrt{\cdot}$，绝对值 $|\cdot|$ 和取符号运算 $\mathrm{sign}(\cdot)$ 是元素级的操作。$L2(\cdot)$ 是二范数归一化。$\bar{z}$ 作为线性分类器的输入，进行二阶微调或元测试。

3.5 元测试方法

在元测试阶段，本文冻结二阶微调后的模型以提取图像的二阶表示，使用支撑集的二阶表示训练任务专属的线性分类器并预测查询样本的类别，以此来评估二阶表示在每个小样本任务中的作用。而在核方法的视角下，这种测试范式的意义得到了新的扩展。视觉编码器提取两张图像的特征图 $\boldsymbol{X},\boldsymbol{Y}\in R^{H\times W\times C}$。使用 CBP 和 $\phi_{\mathrm{TS}}(\cdot)$ 聚合图像的二阶表示等价于利用非线性的二次多项式核函数进行分类，能够增强分类器的判别能力。其解释如下：

$$\begin{aligned}\sum_{ij}\kappa(\boldsymbol{X}_{ij},\boldsymbol{Y}_{ij})&=\sum_{ij}\langle\phi(\boldsymbol{X}_{ij}),\phi(\boldsymbol{Y}_{ij})\rangle\\&=\left\langle\sum_{ij}\phi(\boldsymbol{X}_{ij}),\sum_{ij}\phi(\boldsymbol{Y}_{ij})\right\rangle\end{aligned} \tag{15}$$

在实践中，采用 GAP 聚合图像的一阶表示代表 $\phi(\cdot)$ 是恒等映射，此时核函数 $\kappa(\cdot,\cdot)$ 仅是线性核的建模能力。二阶微调和元测试阶段的算法流程如算法 2 和算法 3 所示。

算法 2. 二阶微调算法流程

输入：训练图像 x，预训练网络 f_t，分类器 LC_{finetune}

输出：优化后的骨干网络 f_{CBP} 或分类器 LC_{finetune}

1. f=load params(f_t) //加载预训练参数
2. IF LC_{finetune} is not None: //加载微调的 LC 参数
3. LC=load params(LC_{finetune})
4. END IF
5. FOR x in DataLoader:
6.   IF LC_{finetune} is None: //冻结骨干网络微调 LC
7.   Freeze(f)
8.   END IF
9. $feat_{\mathrm{2rd}}$=CBP∘$f(x)$ //提取二阶向量，式(10)～(14)
10. $L_{CE}(x)=CE\ Loss(LC(feat_{\mathrm{2rd}}), label(x))$ //计算监督损失
11.   梯度反向传播，更新可训练的参数
12. END FOR

算法 3. 元测试算法流程

输入：新类中若干小样本任务 tasks，骨干网络 f_{CBP}

输出：测试准确度 *acc* 和 95% 置信度 *confindence*

1. f=load params(f_{CBP}) //加载模型参数
2. Freeze(f) //冻结骨干网络
3. *accs*=[] //数据列表，存储每一个任务的准确度
4. FOR task in tasks:
5. D^s, D^q=task //获取支撑集、查询集图像
6. $feat_s$=[CBP∘$f(x_1)$,…,CBP∘$f(x_{NK})$] $x_i\in D^s$, $i=1,\cdots,NK$ //提取支撑集二阶表示
7. $feat_q$=[CBP∘$f(x_1)$,…,CBP∘$f(x_{NQ})$] $x_i\in D^q$, $i=1,\cdots,NQ$ //提取查询集二阶表示
8. $LC=LR(feat_s, feat_q)$ //训练任务专属的逻辑回归分类器，并获取参数
9. $logits=LC(feat_q)$
10.   accs.append(compute_acc(*logits*, $label_q$)) //计算小样本分类任务的准确度
11. END FOR
12. *acc*, *confidence*=mean(*accs*), conf(*accs*) //计算所有任务的准确度和 95% 的置信度

3.6　元测试中的任务难度指标

为了定量评估二阶表示在分类粒度不同的小样本任务中的性能表现，本节提出了一种任务难度指标。在现实世界中，组成每个小样本分类任务的类别和图像具有一定的随机性。每个任务的类间相似关系各不相同，因而带来的分类粒度和识别难度各不相同。例如，对于水母、钢琴、城堡、比目鱼和帆船这 5 个粗粒度类别，它们的类间相似度较低，这反映出不同类别的图像表示相互远离，因而该任务的分类边界容易拟合。相反，对于纽芬兰犬、迷你犬、家雀、知更鸟和巨嘴鸟这 5 个类别，相似类别的表示空间可能出现重叠，这将导致类间边界出现显著的非线性，从而增加了分类器的拟合难度。

基于这种观察，本文利用类间的平均相似度来定量描述一个小样本任务的分类粒度。对于一个“N-way K-shot”问题，提取支撑集 N 个类别的原型向量$\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_N]\in R^{C\times N}$，类原型的定义如下：

$$\boldsymbol{u}_i=\sum_{k=1}^{K}\mathrm{GAP}\circ f(x_{ik}),\ i=1,2,\cdots,N \tag{16}$$

其中，x_{ik} 是支撑集中第 i 类的第 k 个样本。

任务难度指标应该明确反应类间边界的拟合难度，主要取决于相似类别。不相似类别的图像表示在特征空间中距离较远，对拟合难度没有实质的贡献。因此，本文选取类间相似度矩阵中最高的 M 个相似度用于计算任务的难度指标：

$$D_{\mathrm{score}}=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{sim}_i \tag{17}$$

$$\boldsymbol{sim}=\mathrm{descend}\left(\mathrm{triu}\left(\frac{\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}}{\|\boldsymbol{U}\|_2^{\mathrm{T}}\|\boldsymbol{U}\|_2}\right)\right) \tag{18}$$

其中，triu(·)是取上三角矩阵并向量化，descend(·)是向量的降序排列，$\boldsymbol{sim}\in R^{N(N-1)/2}$。当 $N=5$ 时，$1\leq M\leq 10$。如果 $M\to 1$，则 D_{score}关注最相似的类别；如果 $M\to 10$，则 D_{score}关注所有的类间相似关系。在全文中，M 默认设置为 5。

4　实　验

本节首先介绍所用的数据集和详细的实验设置，随后与当前领先的方法进行对比，最后对本文方法进行广泛的消融实验和量化分析。

4.1　数据集

本文在四个主流的基准任务上进行了详细的实验评估。miniImageNet[38]（简记为 mini.）包含 100 个类，每类 600 张图像，其中训练集 64 类，验证集 16 类，测试集 20 类。tieredImageNet[39]包含 608 个类，共 779 165 张图像，其中训练集 351 类，验证集 97 类，测试集 160 类。CUB[40]是一个细粒度的鸟类数据集，共 11 788 张图像，包含 200 个类，其中训练集 100 类，验证集 50 类，测试集 50 类。Aircraft[41]是一个细粒度的飞机数据集，共 10 000 张图像，其中训练集 50，验证集 25 类，测试集 25 类。Cars[42]是一个细粒度的汽车数据集，共 16 185 张图像，其中训练集 130，验证集 17 类，测试集 49 类。跨域任务 miniImageNet→CUB 代表在 miniImageNet 所有类上进行预训练和微调，然后直接在 CUB 的测试集上进行元验证和元测试。miniImageNet→Aircraft 和 miniImageNet→Cars 同样遵循上述设置。

4.2　实验设置

本文在单卡 RTX3090 GPU 和 Intel i5-13600KF CPU 上进行实验。算法实现基于 PyTorch 1.9 框架[43]。本文采用 ResNet12[4,44]和 ResNet18[4]作为骨干网络。ResNet12 输入图像的分辨率为 84×84，ResNet18 输入图像的分辨率为 224×224。与 Deep-BDC[10]、CTX[33]等类似，本文移除了网络末端的降采样以获取更多的卷积特征。下面依次介绍预训练、二阶微调和元测试阶段详细的实验设置。

在预训练阶段，本文采用 SGD 优化器，动量为 0.9，权重衰减为 1e-4。起始学习率为 0.7/256 批次，终止学习率为 1e-6，在预热 10 个轮次后采用余弦策略调整。采用混合精度训练，实际的批次为 180，总共训练 200 个轮次，其中在 tieredImageNet 上仅训练 100 个轮次。教师网络参数采用 EMA 更新，λ 的起始动量为 0.996，终止动量为 1.0，使用余弦策略调整。在式(2)中，c 的 EMA 动量为 0.9。在式(9)中，损失权重 $\alpha=1.0$、$\beta=5.0$、$\gamma=1.0$。数据增广包含随机缩放裁剪、颜色抖动、随机灰度变换、随机水平翻转、高斯模糊和像素反转。多视角增广策略遵循 DINO[18]的设计。

在二阶微调阶段，本文采用 SGD 优化器，动量为 0.9，权重衰减为 5e-4，批次为 64。首先，冻结骨干网络以提取图像的二阶表示，并训练一个线性分类器。此时，学习率为 0.1，训练 30 个轮次，每 10 个轮次学习率乘 0.1。随后，本文以 1e-4 的学习率端到端微调 100 个轮次。数据增广采用标准的随机缩放裁剪、随机水平翻转和颜色抖动。

在元测试阶段，本文采用逻辑回归器学习线性分类边界，正则项参数为 5.0。数据增广采用标准的短边缩放和中心裁剪。本文报告 10 000 个小样本任务的平均测试准确度和 95%的置信区间。

4.3　与同类方法的比较和分析

本文在 4 个主流的小样本分类基准上进行了性

能评估，并报告了在 5-way 1-shot 和 5-way 5-shot 设置下的分类准确度，如表 1～表 3 所示。与最先进的方法相比，本文方法在 3 个数据集上取得了具有竞争力的结果，验证了本文方法的有效性。

通用目标数据集的识别结果如表 1 所示。对于 miniImageNet 数据集，本文方法在 1-shot 设置下相较于最先进的 CVET[12] 提升 0.66%，在 5-shot 设置下则比 FGFL[48] 提升 0.53%。对于 tieredImageNet 数据集，本文方法在 1-shot 设置下取得与最先进的 CORL[49] 可比的结果(73.50±0.23 vs. 73.82±0.58)，在 5-shot 设置下与 FGFL[48] 的结果相当(87.19±0.15 vs. 87.21±0.61)，且本文结果的波动更小。

表 1 在通用目标数据集上的准确度(排名第一和第二的结果分别用加粗和加粗斜体标记(下同))

方法	网络架构	图像尺寸	miniImageNet		tieredImageNet	
			1-shot/%	5-shot/%	1-shot/%	5-shot/%
ProtoNet[26,10]	ResNet12	84×84	62.11±0.44	80.77±0.30	68.31±0.51	83.85±0.36
CovNet[6,10]	ResNet12	84×84	64.59±0.44	82.02±0.29	69.75±0.52	84.21±0.26
GoodEmbed[7]	ResNet12	84×84	64.82±0.44	82.14±0.30	71.52±0.69	86.03±0.58
DeepEMD[8]	ResNet12	84×84	65.91±0.82	82.41±0.56	71.16±0.87	86.03±0.58
FEAT[45]	ResNet12	84×84	66.78±0.20	82.05±0.14	70.80±0.23	84.79±0.16
FRN[9]	ResNet12	84×84	66.45±0.19	82.83±0.13	71.16±0.22	86.01±0.15
MCL[46]	ResNet12	84×84	67.36±0.19	83.63±0.13	71.76±0.22	86.01±0.15
TPMN[47]	ResNet12	84×84	67.64±0.63	83.44±0.43	72.24±0.70	86.55±0.63
Meta DeepBDC[10]	ResNet12	84×84	67.34±0.43	84.46±0.28	72.34±0.49	**87.31±0.32**
PSST[23]	WRN-28-10	84×84	64.16±0.44	80.64±0.32	—	—
CC+rot[24]	WRN-28-10	84×84	64.03±0.45	80.68±0.33	70.53±0.51	84.98±0.36
FewTURE[13]	ViT-S	224×224	68.02±0.88	84.51±0.53	72.96±0.92	86.43±0.67
CVET[12]	ResNet12	84×84	***70.19±0.46***	84.66±0.29	72.62±0.51	86.62±0.33
FGFL[48]	ResNet12	84×84	69.14±0.80	***86.01±0.62***	73.21±0.88	***87.21±0.61***
CORL[49]	ResNet12	84×84	65.74±0.53	83.03±0.33	**73.82±0.58**	86.76±0.52
本文方法	ResNet12	84×84	**70.85±0.19**	**86.54±0.13**	***73.50±0.23***	87.19±0.15

细粒度数据集 CUB 的识别结果如表 2 所示。本文方法显著优于基于一阶特征的 GoodEmbed[7]、DeepEMD[8] 和 FRN[9] 方法，分别提升了 4.12%～9.53%(1-shot)和 1.82%～5.82%(5-shot)，这突显了本文方法的优势。同时，本文方法也领先于同类的基于二阶表示的方法 CovNet[6] 和 Meta DeepBDC[10]。与仅使用有监督学习进行预训练的 Meta DeepBDC[10] 相比，本文方法分别取得 3.12%(1-shot)和 0.98%(5-shot)的性能提升。这一结果表明，本文提出的自监督学习和有监督学习协同优化的预训练方法能够有效地提高模型在未见类别上的泛化性能。

表 2 在细粒度数据集 CUB 上的准确度

方法	模型	CUB	
		1-shot/%	5-shot/%
DeepEMD[8]	ResNet12	77.14±0.29	88.98±0.49
FRN[9]	ResNet12	83.55±0.19	92.92±0.10
FGFL[48]	ResNet12	80.77±0.90	92.01±0.71
LaplacianShot[50]	ResNet18	80.96	88.68
Baseline++[5]	ResNet18	67.02±0.90	83.58±0.54
CovNet[6,10]	ResNet18	80.76±0.42	92.05±0.20
GoodEmbed[7,10]	ResNet18	77.92±0.46	89.94±0.26
FRN[9,10]	ResNet18	82.55±0.19	92.98±0.10
Meta DeepBDC[10]	ResNet18	***83.55±0.40***	***93.82±0.17***
LP-FT-FB[51]	WRN-28-10	82.16	91.88
本文方法	ResNet18	**86.67±0.17**	**94.80±0.08**

跨域任务 miniImageNet→CUB 的识别结果如表 3 所示。与 Meta DeepBDC[10] 相比，本文方法取得了 4.73%的性能提升，与当前最先进的 FGFL[48] 相比，本文方法进一步带来 1.25%的性能提升。此外如 4.5 节表10所示，本文在 miniImageNet→Aircraft 和 miniImageNet→Cars 上均取得当前领先的结果，在 5-shot 设置下分别超过当前最先进Meta DeepBDC[10] 有 1.96%和 4.34%。这些结果表明，本文的模型具有更强的知识迁移能力。

表 3 在跨域任务 miniImageNet→CUB 上的准确度

方法	模型	5-shot/%
ProtoNet[10,26]	ResNet12	67.19±0.38
Baseline[5]	ResNet18	65.57±0.70
ADM[10,27]	ResNet12	70.55±0.43
FEAT[45,48]	ResNet12	62.28
DeepEMD[8,48]	ResNet12	77.34
BML[52]	ResNet12	72.42±0.54
CovNet[6,10]	ResNet12	76.77±0.34
GoodEmbed[7,10]	ResNet12	67.43±0.44
FRN[9]	ResNet12	77.09±0.15
Meta DeepBDC[10]	ResNet12	77.87±0.33
FGFL[48]	ResNet12	***81.35***
本文方法	ResNet12	**82.60±0.14**

4.4 消融实验

本文方法有两个关键的设计需要着重评估，包

括预训练中每个分支任务的作用和二阶微调对性能的影响。此外，本文对所提出方法的适用性和鲁棒性进行了分析。在 miniImageNet 和 CUB 上的消融实验分别使用 ResNet12 和 ResNet18 架构。

4.4.1　预训练不同分支任务对性能的影响

在预训练阶段，不同损失的作用如表 4 所示。仅使用对比式自监督任务即能学习到可迁移的图像表示，并在新类取得良好的测试结果。基于对比式自监督任务，依次增加生成式自监督任务和有监督分类任务，在 miniImageNet 上可以进一步带来 3.40%(1-shot)和 2.61%(5-shot)的提升；在 CUB 上可以进一步带来 12.83%(1-shot)和 10.34%(5-shot)的提升。实验结果显示，对比式自监督、生成式自监督和有监督学习三个任务分支呈现相互促进的效果。采用多任务协同预训练的结果显著超越有监督预训练的结果，在 miniImageNet 上带来 4.75%(1-shot)和 3.57%(5-shot)的提升；在 CUB 上可以进一步带来 5.28%(1-shot)和 2.13%(5-shot)的提升。这表明多任务协同训练能够有效地提升模型在未见类别上的测试结果和泛化性能。

表 4　预训练损失的作用评估　(单位：%)

对比损失	生成损失	分类损失	miniImageNet 1-shot	miniImageNet 5-shot	CUB 1-shot	CUB 5-shot
×	×	√	66.10	82.97	81.39	92.67
√	×	×	67.45	83.93	73.84	84.46
√	√	×	68.14	84.10	75.94	85.95
√	×	√	70.20	86.24	86.61	**94.82**
√	√	√	**70.85**	**86.54**	**86.67**	94.80

多任务协同训练在提高性能的同时也会带来额外的计算代价，如表 5 所示。在 miniImageNet 上进行多任务协同预训练的时间是有监督预训练方案[5-8,10]的 2.7 倍(2.5h vs. 6.7h)。自监督任务往往需要更多的迭代才能发挥预期的作用，例如 FewTURE 等[11,13-14]方法进行了 400～1600 轮次的预训练。相比之下，多任务协同的预训练优化 200 轮次即能实现领先的性能。骨干网络 ResNet12 的参数量为 12 M，在生成式任务分支中参数量仅增加 1.5 M，对比式任务分支的投影头沿用 DINO[18] 的主体设计，参数量增加 22 M。预训练结束后，所有任务分支的参数都会直接丢弃。在二阶微调阶段仅使用预训练的骨干网络，因此预训练新增的投影头参数不会影响后续阶段的训练或推理效率。

表 5　预训练方法对训练时间、参数量和元测试性能的影响

预训练范式	训练轮次	训练时间/h	参数量/M 骨干网络	参数量/M 监督分支	参数量/M 对比分支	参数量/M 生成分支	推理速度/(ms・task^{-1}) 1-shot	推理速度/(ms・task^{-1}) 5-shot	准确度/% 1-shot	准确度/% 5-shot
有监督训练	200	2.5	12	0.05	—	—	100	135	66.10	82.97
多任务协同训练	200	6.7	12	0.05	22	1.5	100	135	**70.85**	**86.54**

未来的工作将进一步优化对比式自监督分支和生成式自监督分支投影头的结构设计，并充分探讨其对预训练效率和小样本分类准确度的影响。

4.4.2　二阶微调的作用评估

在元学习阶段，Baseline[5] 方法使用 GAP 聚合图像的一阶表示，并以线性探测的范式直接进行元测试。基于本文的预训练模型，重新实现了 Baseline[5] 方法，并以此作为评估二阶微调效果的基线。相关结果使用 † 标记。在二阶微调阶段，随着映射维度 D 的增加，CBP 对二次多项式映射的近似误差降低，这有助于提高二阶表示的分辨力和模型的非线性建模能力。然而，在元测试阶段 D 越高，CBP 的计算成本和逻辑回归器的训练代价越大。特征维度 D 对准确度和推理速度的影响，如表 6 所示。在 $D\geq 1024$ 时，二阶微调显著领先于一阶的 Baseline[5] 方法，在 miniImageNet 上分别带来 3.37%～4.36%(1-shot)和 1.39%～2.46%(5-shot)的性能提升，在 CUB 上分别带来 18.85%～19.60%(1-shot)和 7.53%～7.84%(5-shot)的性能提升。随着 D 的增加，推理速度下降，且识别准确度趋于稳定。平衡速

表 6　特征映射维度 D 对元测试准确度(*ACC*,%)和推理延迟(*Lat.*, ms)的影响

方法	D	mini. 1-shot *ACC*	mini. 1-shot *Lat.*	mini. 5-shot *ACC*	mini. 5-shot *Lat.*	CUB 1-shot *ACC*	CUB 1-shot *Lat.*	CUB 5-shot *ACC*	CUB 5-shot *Lat.*
Baseline[5]†	640	66.58	39	84.19	49	67.07	37	86.97	55
本文方法	1024	69.95	45	85.58	62	85.92	68	94.50	84
	2048	70.49	54	86.08	73	86.30	150	94.62	172
	4096	70.74	76	86.36	105	86.39	186	94.70	210
	8192	**70.85**	**100**	**86.54**	**135**	**86.67**	**257**	**94.80**	**318**
	16384	70.94	164	86.65	222	86.60	414	94.81	519

度和准确度，后续实验默认采用 $D=8192$ 的设置。Baseline[5] 方法代表了领先的速度基准。在 $D=1024$ 时，本文方法在CUB上的准确度比 Baseline[5] 高 18.85%(1-shot)，推理延迟增加 80%；在 miniImageNet 上的准确度比 Baseline[5] 高 3.37%(1-shot)，而推理延迟仅增加 15%，具有一定的工程应用前景。

本文利用 CBP 聚合图像的二阶表示进行微调，其对准确度的影响如表 7 所示。由于 CBP 算法没有可学习的参数，因此基于一阶表示训练的模型可以直接利用 CBP 提取图像的二阶表示进行元测试。相较于仅使用一阶表示的识别方案，本文采用一阶表示训练，二阶表示元测试的方案在 miniImageNet 上分别提升了 0.68%(1-shot)和 0.91%(5-shot)；在 CUB 上分别提升了 8.59%(1-shot)和 4.07%(5-shot)。利用二阶表示进行微调的结果更具竞争力，在 miniImageNet 上分别带来 4.27%(1-shot)和 2.35%(5-shot)的性能提升；在 CUB 上分别带来 13.50%(1-shot)和 5.69%(5-shot)的性能提升。

表 7 二阶微调对准确度的影响

图像表示阶数		miniImageNet		CUB	
微调	元测试	1-shot	5-shot	1-shot	5-shot
1st	1st	66.58	84.19	73.17	89.11
1st	2nd	67.26	85.10	81.76	93.18
2nd	2nd	**70.85**	**86.54**	**86.67**	**94.80**

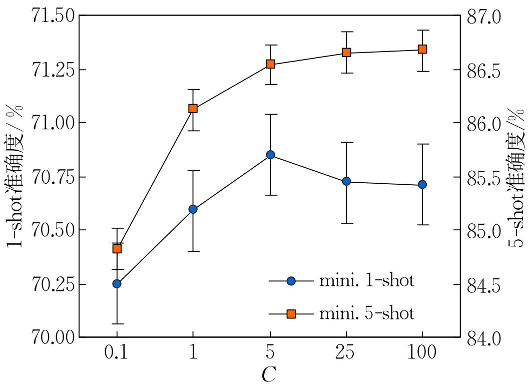
4.4.3 鲁棒性和通用性分析

与文献[10,33]相同，为了关注更多的图像细节，本文移除了网络的最后一个降采样以获得更多的卷积特征。本文在两种架构下评估了特征数量对模型性能的影响，如表 8 所示。在 ResNet18 中，当特征数量由 7×7 变为 14×14 时，1-shot 和 5-shot 的准确度分别增加 0.63%和 0.37%。增加特征数量有利于模型关注更多的空间细节特征，这对细粒度图像的识别更有帮助。即使不取消降采样，本文的结果依然具有竞争力。

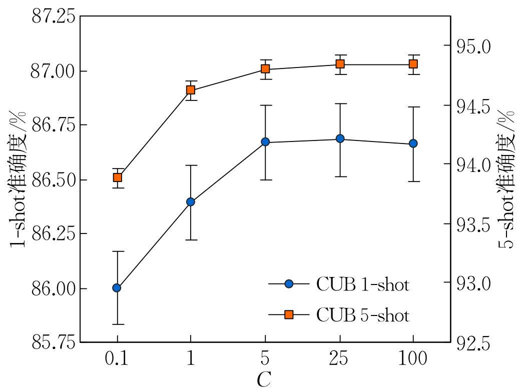
表 8 模型输出特征数量对准确度的影响

网络	数据集	图像分辨率	特征数量	*ACC*/%	
				1-shot	5-shot
ResNet12	mini.	84×84	5×5	70.52	**86.71**
			10×10	**70.85**	86.54
ResNet18	CUB	224×224	7×7	86.05	94.30
			14×14	**86.68**	**94.67**

在元测试阶段，逻辑回归器的正则化参数 C 影响分类边界的拟合效果，其对分类结果的评估如图 3 所示。本文利用 scikit-learn 库中实现的逻辑回归器，C 越大代表正则化程度越小。在 5-way 的设置下，C 较小时($C=0.1$)会出现欠拟合，导致分类准确度下降约 1.0%。此外，基于二阶表示训练的逻辑回归器对 C 的设置具有鲁棒性。C 在一个较大的范围内($C\geq 5.0$)在 miniImageNet 和 CUB 上准确度波动较小，且趋于稳定。



(a)



(b)

图 3 逻辑回归正则项参数 C 对元测试结果的影响

多任务协同的预训练可以自然地与一些两阶段的小样本分类方法相结合。本文在 miniImageNet 上对 ProtoNet[26]、Baseline[5] 和 GoodEmbed[7] 三种方法进行了复现，如表 9 所示。具体而言，本文首先采用本文的预训练范式对模型进行预训练，而不是沿用原文[5,7,26]的有监督预训练范式。随后，本文按照上述三种方法的设计进行后续的训练或测试。在元学习阶段，ProtoNet[26] 通过计算特征间的欧氏距离进行相似性判别；Baseline[5] 和 GoodEmbed[7] 在支撑集上训练线性分类器用于预测查询集样本。此外，GoodEmbed[7] 在本文预训练的基础上使用 KL 散度[53]进行了一次额外的自蒸馏。实验发现，本文实现的结果均高于原文报告的结果。相较于原文的有监督预训练，本文的预训练范式带来 1.24%～4.56%(1-shot)和 1.49%～4.55%(5-shot)的性能提升，这充分展现了本文利用自监督学习和有监督

学习进行多任务预训练的优势和通用性。

表 9　预训练方法的通用性分析

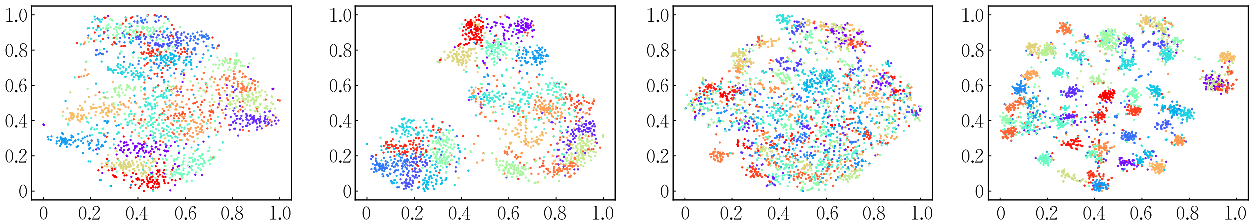
方法	miniImageNet	
	1-shot/%	5-shot/%
ProtoNet[26,10]	62.11±0.44	80.77±0.30
+本文预训练	**63.35±0.19**	**82.26±0.13**
Baseline[5,7]	62.02±0.63	79.64±0.44
+本文预训练	**66.58±0.19**	**84.19±0.13**
GoodEmbed[7]	64.82±0.44	82.14±0.30
+本文预训练	**68.01±0.19**	**84.91±0.13**

4.5　二阶表示在小样本分类中的作用和优势

（1）二阶表示的作用分析

利用 t-SNE 将 mini-ImageNet 和 CUB 测试集中所有未见类别的图像表示向量投影到二维平面，对一阶和二阶表示的聚类效果进行可视化，如图 4 所示。在一阶表示的空间中，相似类别样本的特征分布存在交集的可能性更大，这将增加分类边界的非线性程度。相反，二阶表示的类内样本分布更集中，类间特征的距离增大，类间边界更加线性可分。因此，二阶表示能够有效地增强分类器的判别能力。

每一个小样本任务的类别和样本都随机采样于所有测试图像。在二阶表示的特征空间中，任务图像的类间可分性增强，因而模型整体的识别性能更高。如表 7 所示，在 miniImageNet 的 1-shot/5-shot 设置下，二阶表示带来 4.27% 和 2.35% 的性能提升；在 CUB 的 1-shot/5-shot 设置下，二阶表示带来 13.50% 和 5.6% 的性能提升。如表 10 所示，在 3 个跨域细粒度任务的 1-shot/5-shot 设置下，二阶表示带来 2.28%～11.15% 和 5.02%～11.92% 的性能提升。



(a) 1st mini.　(b) 2nd mini.　(c) 1st CUB　(d) 2nd CUB

图 4　一阶表示和二阶表示的可视化（图(a)和(b)：在 miniImageNet 测试集 20 类上的可视化；图(c)和(d)：在 CUB 测试集 50 类上的可视化。每类随机选择 100 张图像，并分别提取一阶和二阶表示进行可视化）

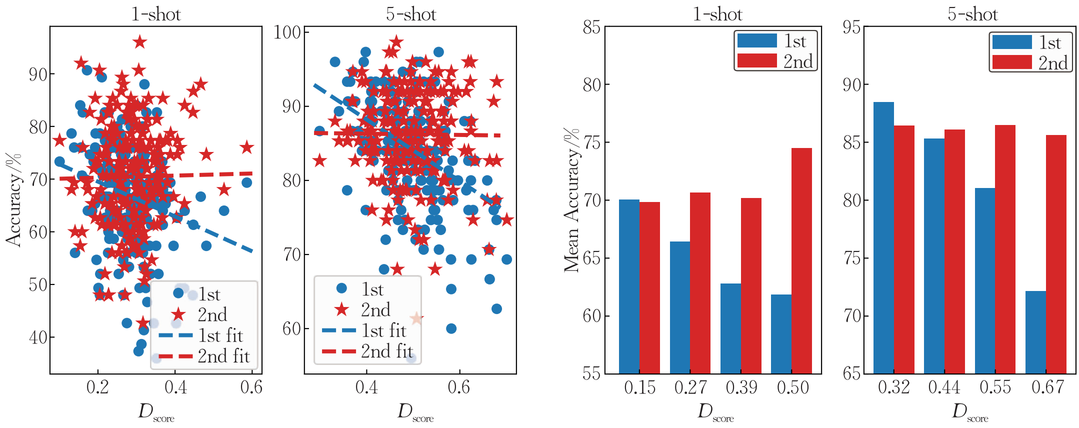
表 10　一阶和二阶表示在跨域任务上的结果对比

方法	模型	表示阶次	miniImageNet→CUB		miniImageNet→Aircraft		miniImageNet→Cars	
			1-shot/%	5-shot/%	1-shot/%	5-shot/%	1-shot/%	5-shot/%
ProtoNet[10,26]	ResNet12	一阶	—	67.19	—	55.96	—	46.30
Baseline[5]	ResNet12	一阶	—	—	—	59.04	—	50.29
GoodEmbed[7,10]	ResNet12	一阶	—	67.43	—	58.95	—	50.18
ADM[10,27]	ResNet12	一阶+二阶	—	70.55	—	65.40	—	53.94
CovNet[6,10]	ResNet12	二阶	—	76.77	—	63.56	—	52.90
Meta DeepBDC[10]	ResNet12	二阶	—	77.87	—	68.67	—	54.61
Baseline[5]†	ResNet12	一阶	49.30	70.64	41.91	63.23	36.08	53.93
本文方法	ResNet12	二阶	**60.45**	**82.60**	**47.14**	**70.63**	**38.36**	**58.95**

（2）二阶表示的比较优势

一个小样本任务的类间样本越相似，分类越细粒度，类间边界拟合越困难，任务难度指标越高。通过任务难度指标 D_{score} 量化每个任务的分类粒度，以此定量地探究了一阶和二阶表示在不同粒度的小样本分类任务上的表现。具体而言，利用预训练模型提取一阶表示；利用二阶微调的模型提取二阶表示。从 miniImageNet 的测试集中随机选取 1000 个任务，利用一阶表示计算每个任务的难度分数 D_{score}。对每个小样本分类任务分别利用一阶和二阶表示进行元测试，并统计每个难度范围内的平均分类准确度，结果展示如图 5 所示。为了避免样本点的重合，在散点图中仅展示了 200 个随机采样任务的结果。

由图 5(a)可知，一阶和二阶表示对每个小样本任务的分类准确度有所差异，但整体上二阶表示的效果更优。由图 5(b)可知，随着任务难度的增加，一阶表示的识别准确度出现显著下降，而二阶表示的准确度保持稳定。在高难度的任务中，二阶表示展现出稳健的线性判别能力。任务难度越高、类间样本越相似，二阶表示的提升越大。

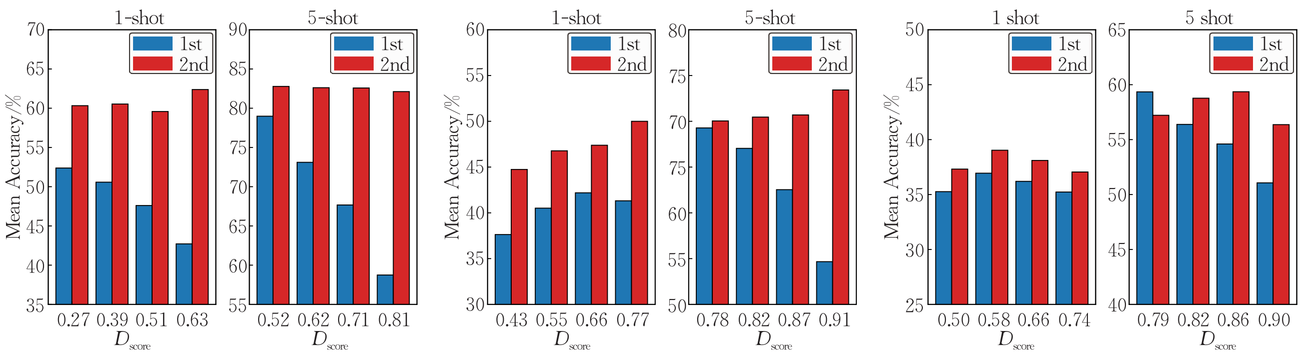


(a) 任务难度-准确度的散点图 (b) 任务难度-平均准确度的直方图

图 5 在通用识别数据集 miniImageNet 上，一阶和二阶表示对不同粒度任务的识别准确度（一阶和二阶表示的元测试结果分别使用 1st 和 2nd 标记）

(3) 应用效果与启示

跨域场景的训练和测试数据存在分布偏移，更加考验图像表示在未见类别上的泛化效果和分辨力。如图 6 所示，在跨域细粒度任务 miniImageNet→CUB、miniImageNet→Aircraft 和 miniImageNet→Cars 上，二阶表示的识别效果在各个任务难度区间整体都领先于一阶表示，且任务越难领先优势越大。而在局部上，一阶和二阶表示在不同数据集上的表现存在差异。例如，一阶表示在 miniImageNet→Aircrafts 和 miniImageNet→Cars 的低难度 5-shot 任务上的准确度与二阶表示持平或略优。因此，在实际应用中，通过分析任务难度和识别准确度的关系可以确定方法的优势场景。综合考虑识别准确度和推理速度等方面，能够帮助从业者选择出更符合实际的小样本识别方案。



(a) miniImageNet→CUB (b) miniImageNet→Aircraft (c) miniImageNet→Cars

图 6 在跨域任务上，一阶和二阶表示对不同粒度任务的识别准确度

5 结 论

本文基于对比学习框架提出了一种多任务协同优化的预训练方法，实现了对比式自监督任务、生成式自监督任务和监督分类任务的联合训练，用于学习可迁移到未见类别的图像表示。随后，本文利用 CBP 对预训练模型进行微调，用于提取图像更具分辨力的二阶表示。为了进一步探究二阶表示在小样本分类任务中发挥的作用，基于小样本任务中图像的类间相似度关系，本文设计了一种衡量小样本任务分类难度的指标。基于线性探测的元测试方法，本文对一阶和二阶表示在小样本分类中的表现进行了对比分析。实验发现，在预训练中不同的任务呈现相互促进的效果，多任务协同的预训练能够有效提高模型的泛化性能。二阶表示在高难度的小样本任务上具有更强的线性可分优势。本文方法在六个主流的基准任务上取得了有竞争力的结果。未来的工作将专注于探索本文方法在 Transformer 架构上的适应性。

致 谢 感谢各位审稿专家和编辑老师在百忙之中审阅本文！

参　考　文　献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012, 25: 1-9

[2] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[3] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-22

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[5] Chen W Y, Liu Y C, Kira Z, et al. A closer look at few-shot classification//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-17

[6] Wertheimer D, Hariharan B. Few-shot learning with localization in realistic settings//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6558-6567

[7] Tian Y, Wang Y, Krishnan D, et al. Rethinking few-shot image classification: A good embedding is all you need?//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 1-13

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

[9] Wertheimer D, Tang L, Hariharan B. Few-shot classification with feature map reconstruction networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 8012-8021

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

[11] He Y, Liang W, Zhao D, et al. Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 9119-9129

[12] Yang Z, Wang J, Zhu Y. Few-shot classification with contrastive learning//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 293-309

[13] Hiller M, Ma R, Harandi M, et al. Rethinking generalization in few-shot classification//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 3582-3595

[14] Hao F, He F, Liu L, et al. Class-aware patch embedding adaptation for few-shot image classification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 18905-18915

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

[16] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 9729-9738

[17] Grill J B, Strub F, Altché F, et al. Bootstrap your own latent a new approach to self-supervised learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020, 33: 21271-21284

[18] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 9650-9660

[19] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16000-16009

[20] Zhou J, Wei C, Wang H, et al. iBOT: Image BERT pretraining with online tokenizer//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-29

[21] Gao P, Ma T, Li H, et al. MCMAE: Masked convolution meets masked autoencoders//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 35632-35644

[22] Woo S, Debnath S, Hu R, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 16133-16142

[23] Chen Z, Ge J, Zhan H, et al. Pareto self-supervised training for few-shot learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 13663-13672

[24] Su J C, Maji S, Hariharan B. When does self-supervision improve few-shot learning?//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 645-666

[25] Gidaris S, Bursuc A, Komodakis N, et al. Boosting few-shot visual learning with self-supervision//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 8059-8068

[26] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017, 30: 1-13

[27] Li W, Wang L, Huo J, et al. Asymmetric distribution measure for few-shot learning. arXiv preprint arXiv: 2002.00153, 2020

[28] Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 317-326

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

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

[31] Wang Duo-Rui, Du Yang, Dong Lan-Fang, et al. Feature transformation and metric networks for few-shot learning. Acta Automatica Sinica, 2024, 50(7): 1305-1314(in Chinese)
(王多瑞,杜杨,董兰芳等. 基于特征变换和度量网络的小样本学习算法. 自动化学报, 2024, 50(7): 1305-1314)

[32] Li Xiao-Xu, Liu Zhong-Yuan, Wu Ji-Jie, et al. Total relation network with attention for few-shot image classification. Chinese Journal of Computers, 2023, 46(2): 371-384(in Chinese)
(李晓旭,刘忠源,武继杰等. 小样本图像分类的注意力全关系网络. 计算机学报, 2023, 46(2): 371-384)

[33] Doersch C, Gupta A, Zisserman A. CrossTransformers: Spatially-aware few-shot transfer//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020, 33: 21981-21993

[34] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1449-1457

[35] Wang Q, Zhang Z, Gao M, et al. Towards a deeper understanding of global covariance pooling in deep learning: An optimization perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15802-15819

[36] Wang Q, Xie J, Zuo W, et al. Deep CNNs meet global covariance pooling: Better representation and generalization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2582-2597

[37] Pham N, Pagh R. Fast and scalable polynomial kernels via explicit feature maps//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 239-247

[38] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016, 29: 1-9

[39] Ren M, Triantafillou E, Ravi S, et al. Meta-learning for semi-supervised few-shot classification//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-15

[40] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD birds-200-2011 dataset. Computation & Neural Systems Technical Report: CNS-TR-2011-001, 2011

[41] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft. arXiv preprint arXiv: 1306.5151, 2013

[42] Krause J, Stark M, Deng J, et al. 3D object representations for fine-grained categorization//Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia, 2013: 554-561

[43] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019, 32: 1-12

[44] Oreshkin B, Rodríguez López P, Lacoste A. TADAM: Task dependent adaptive metric for improved few-shot learning//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2018, 31: 1-13

[45] Ye H J, Hu H, Zhan D C, et al. Few-shot learning via embedding adaptation with set-to-set functions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 8808-8817

[46] Liu Y, Zhang W, Xiang C, et al. Learning to affiliate: Mutual centralized learning for few-shot classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 14411-14420

[47] Wu J, Zhang T, Zhang Y, et al. Task-aware part mining network for few-shot learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 8433-8442

[48] Cheng H, Yang S, Zhou J T, et al. Frequency guidance matters in few-shot learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 11814-11824

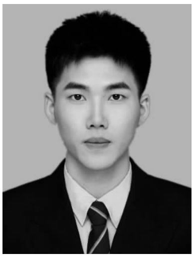
[49] He J, Kortylewski A, Yuille A. CORL: Compositional representation learning for few-shot classification//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Hawaii, USA, 2023: 3890-3899

[50] Ziko I, Dolz J, Granger E, et al. Laplacian regularized few-shot learning//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 11660-11670

[51] Wang H, Yue T, Ye X, et al. Revisit finetuning strategy for few-shot learning to transfer the emdeddings//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-11

[52] Zhou Z, Qiu X, Xie J, et al. Binocular mutual learning for improving few-shot classification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 8402-8411

[53] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2014: 1-9



LI Zhao-Liang, M. S. His research interests include few-shot image classification and self-supervised learning.

JIA Ling-Yao, Ph. D. candidate. His research interests include image classification and deep learning.

ZHANG Bing-Bing, Ph. D., lecturer. Her research interests include video action recognition, image classification and deep learning.

LI Pei-Hua, Ph. D., professor, Ph. D. supervisor. His research interests include deep learning, image/video recognition, object detection and semantic segmentation.

Background

Few-shot image classification has received widespread attention due to its potential applications in agricultural pest and disease identification, industrial defect classification, and other fields. Existing methods can be organized into the following two main-stream families, optimization-based methods and metric-based methods. Metric-based methods show great prospects and have achieved leading results on many benchmarks.

Currently, there are two key research routes for main-stream metric-based methods, namely pre-training and meta-learning. Pre-training is used to improve the model's representation power, thereby enhancing the performance of distance function in the meta-learning stage. Initially, most such studies used classification tasks for pre-training on the base set. Some works have explored different self-supervised tasks for pre-training, effectively improving their generalization performance. However, how to utilize self-supervised learning more effectively to enhance the model's expressiveness, is still a question that deserves deeper exploration. Especially for convolutional network-based methods, this exploration is clearly insufficient. Normally, in the meta-learning stage, they exploit the image's first-order representation, and focus on the design of distance functions. Instead, to leverage the rich statistical information in the features, some methods extract the images' second-order representations for similarity measurement, achieving leading performance. However, in different few-shot classification tasks, the inter-class similarity varies, resulting in different level of difficulty in classification. The concrete advantages of second-order representation in different few-shot classification tasks with different granularities are not yet clear.

This paper conducts research along the two key clues of pre-training and meta-learning. The training on the base set is divided into two stages. First, we have designed a multi-task cooptimization pre-training method to learn transferable representations. It is built upon a contrastive learning framework that realizes the joint training of contrastive learning tasks, generative self-supervised tasks, and supervised classification tasks. Compared to previous works, the proposed method can naturally integrate contrastive learning tasks and generative self-supervised tasks into the training of convolutional networks, demonstrating good complementarity. Subsequently, we finetune the pretrained model with parameter-free CBP to aggregate the second-order representations of the images and further enhance its detail perception. In the meta learning stage, we extract the second-order representations of images and evaluate their performance on few-shot classification using linear probing. To quantitatively analyze the role of second-order representation, we define a task difficulty score to characterize the level of nonlinearity in the inter-class boundaries for each few-shot classification task. Experimental results show that the second-order representations exhibit stronger linear separability compared to the first-order ones, making them more suitable for difficult few-shot classification tasks. We achieved leading results on four benchmarks, verifying the effectiveness of the proposed method.

Few-shot learning (FSL) and higher-order representation learning (HoRL) are the major research interests in our laboratory. DeepBDC is a representative FSL work and was accepted by oral presentation at CVPR2022. Our laboratory has been continuoussly exploring in the field of HoRL and related findings have published in IEEE transactions on Pattern Analysis and Machine Intelligence, IEEE transactions on Image Processing, CVPR, NeurIPS, ICCV etc. Representative works include but not limited to matrix normalization methods and theoretical study of HoRL (TPAMI2021, 43(8): 2582-2597) along with its analysis from optimization viewpoint (TPAMI2023, 45(12): 15802-15819). Inspired by our previous works, we explore FSL from the perspective of self-supervised learning and second-order representation. This work was supported by the National Natural Science Foundation of China under Grant No. 62471083.