语义辅助少样本学习图像分类的复现与改进

摘要

少样本学习旨在利用少量标注数据实现高效的图像分类，其关键在于如何提取和融合视觉特征与语义特征以提高分类性能。本文在复现已有基于语义辅助的少样本学习方法的基础上，提出了进一步的尝试性改进方案。复现工作包括通过大语言模型扩展生成高质量语义源，然后结合语义对齐网络将文本特征与视觉特征进行融合，并在推理阶段通过特征对齐机制提升分类性能。本文尝试将主干视觉编码器替换为 ResNet50，充分利用其深层次特征提取能力，进一步提高了分类正确率，该结果在 CIFAR-FS 和 FC100 两个数据集上的 12 组实验得到了验证。

关键词：少样本学习；图像分类；语义融合

1 引言

深度学习模型在利用大规模标注数据时取得了显著进展，然而在许多现实应用中，标注数据的缺乏限制了传统深度学习方法的广泛应用。与此相对，人在学习新概念和进行图像识别时具有非凡的能力，能够从极少量的标注样本中学习和推理。这一现象启发了少样本学习（Few-Shot Learning, FSL）[4] 的提出，旨在模仿人类的学习能力，通过少量标注数据来进行有效的图像分类。

在 FSL 的典型设置中，支持集由 N 个类别组成，每个类别包含 K 个样本，模型通过在支持集上学习，要求能够将测试样本（查询集）正确分类为这 N 个类别中的一个。然而，传统的分类方法往往通过将支持集和查询集投影到预设的度量空间来进行分类，通过度量相似性来对查询样本进行标注。问题在于，少量的支持样本可能无法提供足够的判别特征，尤其当这些样本位于度量空间的外围时，距离评估可能不稳定，导致分类准确率降低。

少样本学习还面临着多方面的挑战：首先，数据稀缺导致模型无法学习到足够的区分性特征，使得模型在面对新的、未见过的样本时表现出泛化能力不足；其次，类别内部的样本可能表现出较大的变异性，进一步增加了构建鲁棒特征表示的难度；最后，不同类别之间的样本可能在视觉特征上存在高度相似性，这加剧了分类任务的复杂性。在这一背景下，设计能够有效应对这些挑战的模型和方法成为少样本学习研究的核心任务。

2 相关工作

少样本学习在许多现实应用中面临着诸多挑战，这使得传统的深度学习方法难以在极少量标注数据的情况下取得良好的表现。为了应对这一问题，研究者们提出了多种方法以提升

模型在少样本场景下的分类能力。总体而言，这些方法可以分为基于视觉特征的传统方法和利用多模态信息的语义辅助方法。下面将分别介绍这两类方法的核心思路及其研究进展。

2.1 基于视觉特征的传统方法

为了应对少样本学习中因数据稀缺和特征不足导致的分类挑战，研究者们首先提出了一系列基于视觉的方法。这些方法旨在从样本中提取与类别相关的特征，减少类内变化，从而构建鲁棒的特征表示。例如 Hao Cheng 等 [2] 通过引入频率引导的掩模机制和多层次度量学习策略，有效突出类别判别性强的频率信息，从而显著提升了少样本学习的分类性能和泛化能力；例如 Markus Hiller 等 [3] 提出了一种基于 vision transformer（ViTs）的少样本学习方法，通过将输入图像分割为小块并利用无监督的掩蔽图像建模来学习数据的更一般统计结构，从而克服了图像级标签的限制，并通过在线优化选择最具信息性的图像块，并提供视觉解释，展示哪些区域对分类任务最为重要；这类方法主要是通过数据增强、改进特征提取网络等手段提高少样本的效果，在许多研究中得到了广泛的探索并取得了显著的成功。然而，当样本极为稀少甚至每个类别仅有一个样本时（one-shot 任务），无论通过何种方法进行数据增强，或是改进特征提取网络，都难以学习到足够的图像特征，这种基于视觉的方法仍然难以满足需求。

2.2 语义辅助方法

针对这一局限，一些研究者开始探索如何通过多模态的方式增加特征信息，也就是基于语义辅助的方法，利用语言和视觉模式之间的协同作用来增强模型对样本的理解能力。基于语义的方法通过引入不同类型的语义信息来提升分类效果。其中，一种常见方法是将类名作为语义信息的来源，Chen Xing 等 [5] 首先利用类名提供的先验知识帮助模型学习。然而，类名并非总是最佳的语义来源。例如，对于从未见过“斑马”这一类的人，通过其定义（如“有条纹的马”）来理解和识别斑马可能更加有效。此外，类名在某些场景中可能存在歧义，例如在 MiniImageNet 数据集中，“耳朵”这一类的实际图像是玉米，而非人体器官。语义源中包含的信息数量是非常重要的，如图1所示，左图中由于美洲鸵被枝叶遮挡，导致可学习的特征很少，干扰项过多，但提供了图片上方这段文本作为语义源后，可学习的特征非常丰富，效果可以等价于右边这种没有遮挡的图片。

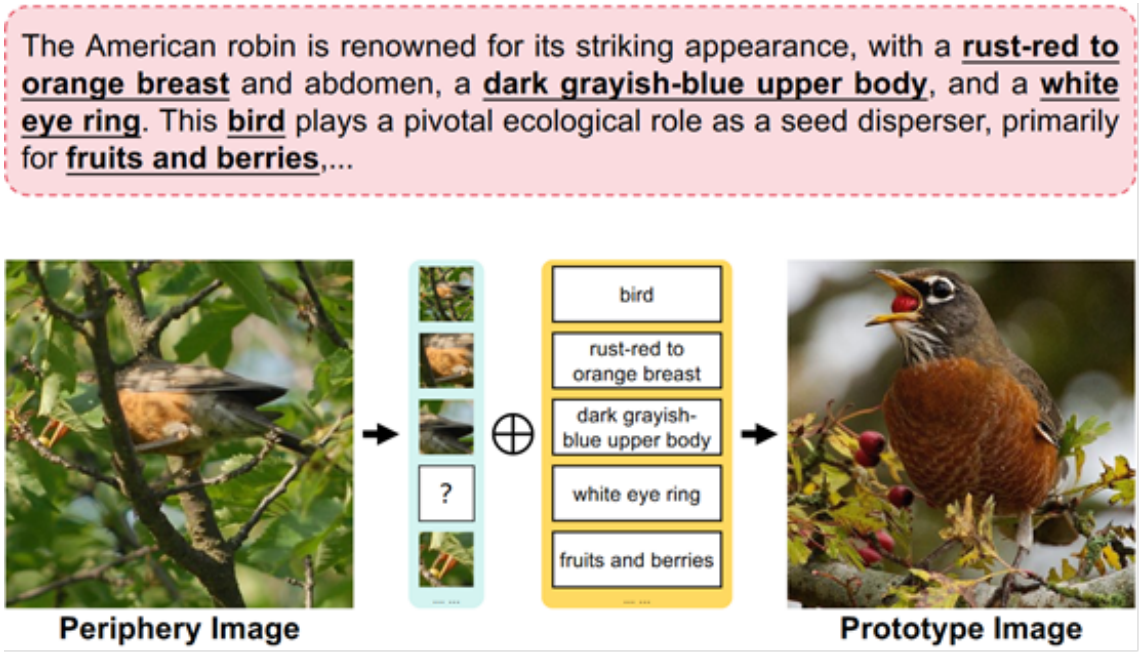

图 1. 语义源与图像结合案例

Wentao Chen 等 [1] 的工作就是采用了这一思路，他们通过在 WordNet 词典中对类名进行查询，将查询得到后的定义作为语义源，提取文本特征，然后将文本特征与图像特征进行融合，从而提高少样本学习图像分类的正确率。这一工作证明了高质量语义源对少样本学习的正面作用。本文复现并尝试改进的论文，也就是 Hai Zhang 等 [6] 的工作中，他们进一步使用大语言模型（LLM，large language model）生成内容更加丰富的高质量语义源辅助少样本图像分类，并得到了良好的实验结果。

总体而言，如何高效地收集和利用高质量的语义信息是当前少样本学习领域亟待解决的问题。未来的研究应致力于进一步提升语义信息的丰富性和表示能力，以更好地支持少样本学习任务。

3 本文方法

3.1 生成高质量语义源

在少样本学习任务中，语义信息的质量对模型性能具有重要影响。为了有效增强样本的语义表达能力，需要一种获取高质量语义源的方法，具体步骤如图2所示。首先从少样本数据集中提取样本的类名（标签）。这些类名提供了类别的基本信息，是生成语义的起点。然后将获取的类名输入至 WordNet 中，查询与该类名相关的定义。例如，对于类名 “Robin”，WordNet 返回其定义为 “Large American thrush having a rust-red breast and abdomen”，然后使用大型语言模型（LLM）对从 WordNet 获得的定义进行扩展，即通过图2中的提示词让大语言模型对定义进行扩写，最终得到了一个内容丰富的文本段落，可以用作高质量的语义源。这种方法有效地利用了储存在预训练的大语言模型中的大量知识。

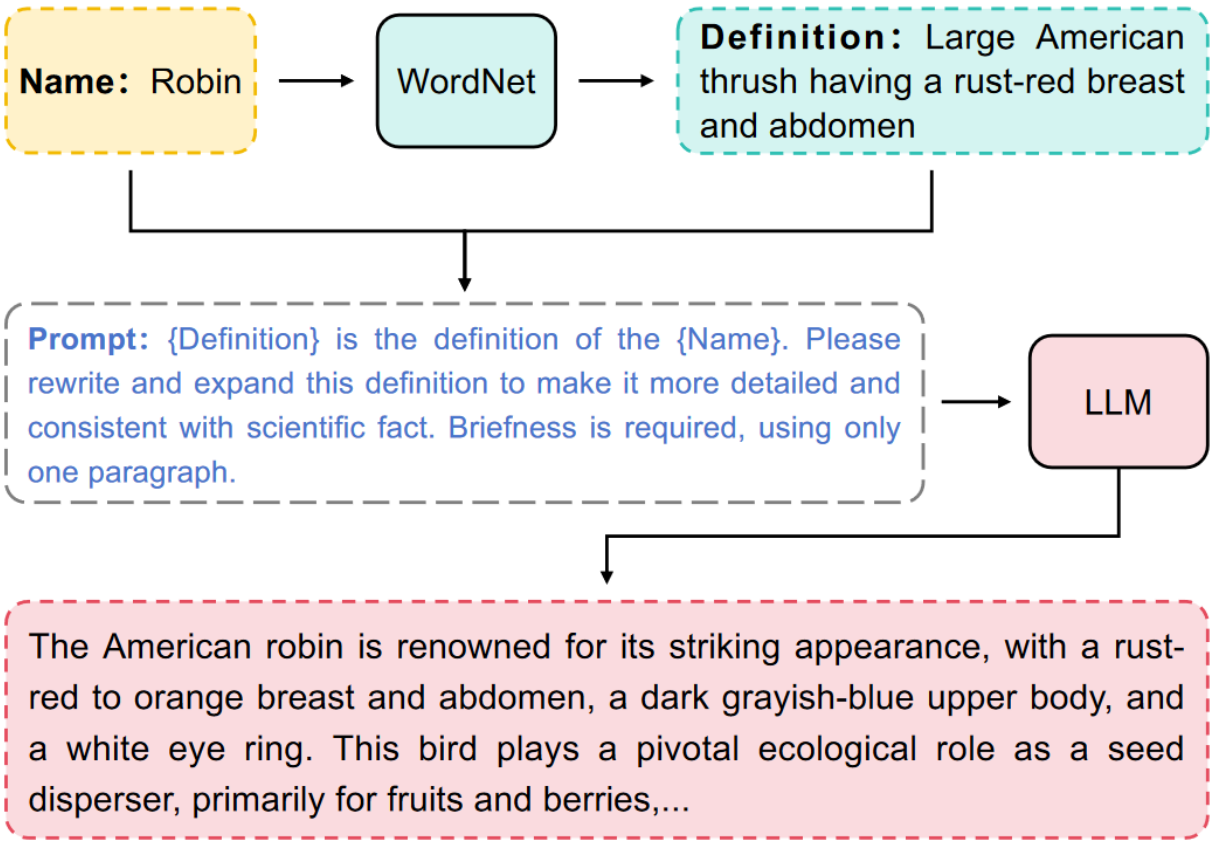

图 2. 使用大语言模型生成高质量语义的过程

3.2 提取文本特征与视觉特征

在生成高质量语义文本后，下一步是提取文本特征和视觉特征，并通过特征融合机制实现少样本图像分类。该过程如图3所示，具体而言，首先使用 Vision Encoder 提取样本图像的视觉特征向量。Vision Encoder 可以采用 ResNet 或 Swin-Transformer 等模型，通过捕捉图像的全局语义和细粒度信息生成深层次的图像特征。这些视觉特征为分类任务提供了丰富的视觉表达能力，是模型决策的重要依据。同时，通过 Text Encoder 提取高质量语义文本的特征向量。Text Encoder 使用 ViT-B/16 CLIP 或 BERT-Base 模型对扩展生成的语义段落进行编码，从而生成具有语义丰富性和判别能力的文本特征。

得到视觉特征和文本特征后，可以使用图3中的 SemAlign 网络进行特征融合，该网络的具体实现在 3.3 节中进行展示。在推理阶段，首先需要通过支持集生成类别的原型向量。具体而言，模型分别提取支持集中每个样本的视觉特征和文本特征，并通过 SemAlign 模块将两类特征融合，然后计算每个类别的特征平均值以生成类别原型。这些类原型代表了每个类别的全局特征，是模型在查询样本分类中参考的标准。对于查询样本，模型提取其视觉特征后，通过 SemAlign 模块与支持集的文本特征进行对齐。在此基础上，查询样本的特征与支持集类别原型进行最近邻匹配。匹配过程中，使用欧几里得距离来衡量查询样本与各类别原型的相似性，最终将查询样本分类到距离最近的类别。这种特征对齐与匹配的过程，确保了模型能够在少量样本的情况下实现高精度分类。

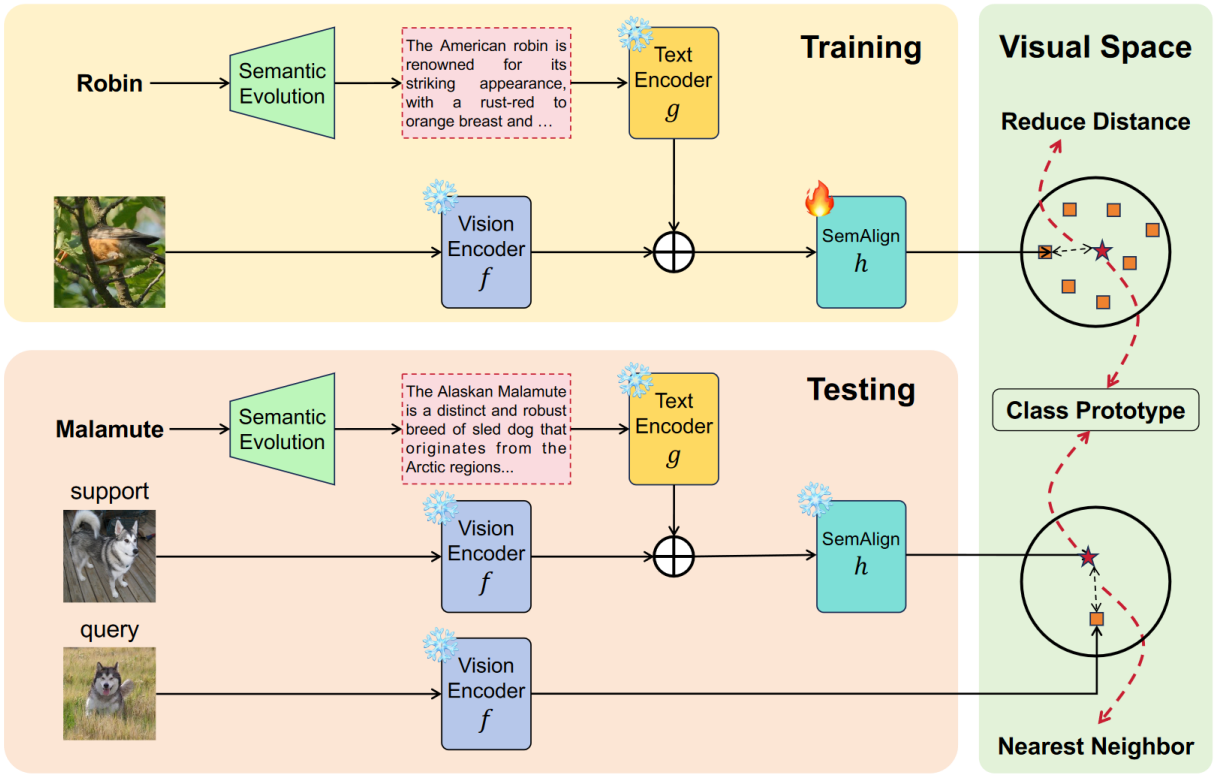

图 3. 提取文本特征与视觉特征用于少样本图像分类

3.3 特征融合网络

SemAlign 模块的核心流程包括以下几个步骤：首先，拼接层将文本语义特征与视觉上下文特征进行拼接，以形成融合输入；接着，线性层将拼接后的输入映射到隐层空间，为后续特征融合奠定基础。随后，经过激活函数的非线性变换与 Dropout 层，模型能够有效防止过拟合并提升特征的泛化能力；最后，通过全连接层将融合后的特征映射回输出的目标维度，从而获得最终的融合表示。该网络的训练损失函数为：

$$\min_{W_1,W_2} \mathbb{E}\left[\|h(f(x_i), g(s_i)) - c_y\|_1\right]$$

其中 f 为提取图片的视觉特征向量，g 为提取语义源的文本特征向量，h 即为 SemAlign 特征对齐网络，c 为样本所属类型的原型（也就是所有视觉特征的平均值）。

4 复现细节

4.1 与已有开源代码对比

原论文开源仓库地址为 https://github.com/zhangdoudou123/SemFew。在该仓库中，已包含文中特征融合网络 SemAlign 的模型代码、从其他开源仓库中复制的 ResNet12 和 Swin-Transformer 模型代码，以及 SemAlign 的训练和测试脚本。本复现工作参考了该仓库中提供的各种模型代码，实现了 ResNet50 特征提取网络，并修改了训练脚本使 SemAlign 特征融合网络的输入与 ResNet50 的输出以及语义特征进行对齐，最终完成了将文中的主干网络替换为 ResNet50，从而优化了特征融合的效果。

4.2 创新点

在少样本学习任务中，视觉编码器（Vision Encoder）的选择对模型性能具有至关重要的影响。为进一步优化模型，我将上述流程中的视觉编码器替换为 ResNet50，以提升视觉特征的表达能力并促进语义与视觉特征的深度融合。

ResNet50 是一种广泛应用于图像分类任务的深度卷积神经网络，具有更深的网络层次和强大的特征提取能力。相比原有的视觉编码器（如 ResNet12 或 Swin-Transformer），ResNet50 的复杂性使其能够捕捉更丰富、更细粒度的图像特征。这种特征的多样性不仅增强了类别特征的区分性，还为文本特征的嵌入提供了更多的语义上下文空间。具体来说，视觉特征越丰富，文本特征与视觉特征的匹配点越多，从而使文本信息能够更精准地作用于图像特征，最终提升特征融合后的表达能力。

此外，ResNet50 的深层结构允许模型在更高维度上学习到多尺度和多层次的图像语义信息。这对于少样本学习中的难分类别（例如跨类别间相似性较高的类别）尤为重要。通过深度特征表征，ResNet50 提供了更具判别力的特征空间，有助于减少类间混淆并提高分类准确率。

5 实验结果分析

为了验证本文提出方法的有效性，我们在 CIFAR-FS 和 FC100 两个数据集上进行了系统的实验。实验使用了 ResNet12、Swin-Transformer 和 ResNet50 三种视觉编码器，并分别在 1-shot-5-way 和 5-shot-5-way 的任务设置下进行了测试，共计 12 组实验。其中使用 ResNet12 和 Swin-Transformer 的八组实验如图4所示，每一组图中，左侧折线图为训练过程中，特征融合后图像分类正确率（黄色线）与特征融合前图像分类正确率（蓝色线），右侧折线图为特征融合网络的损失值。

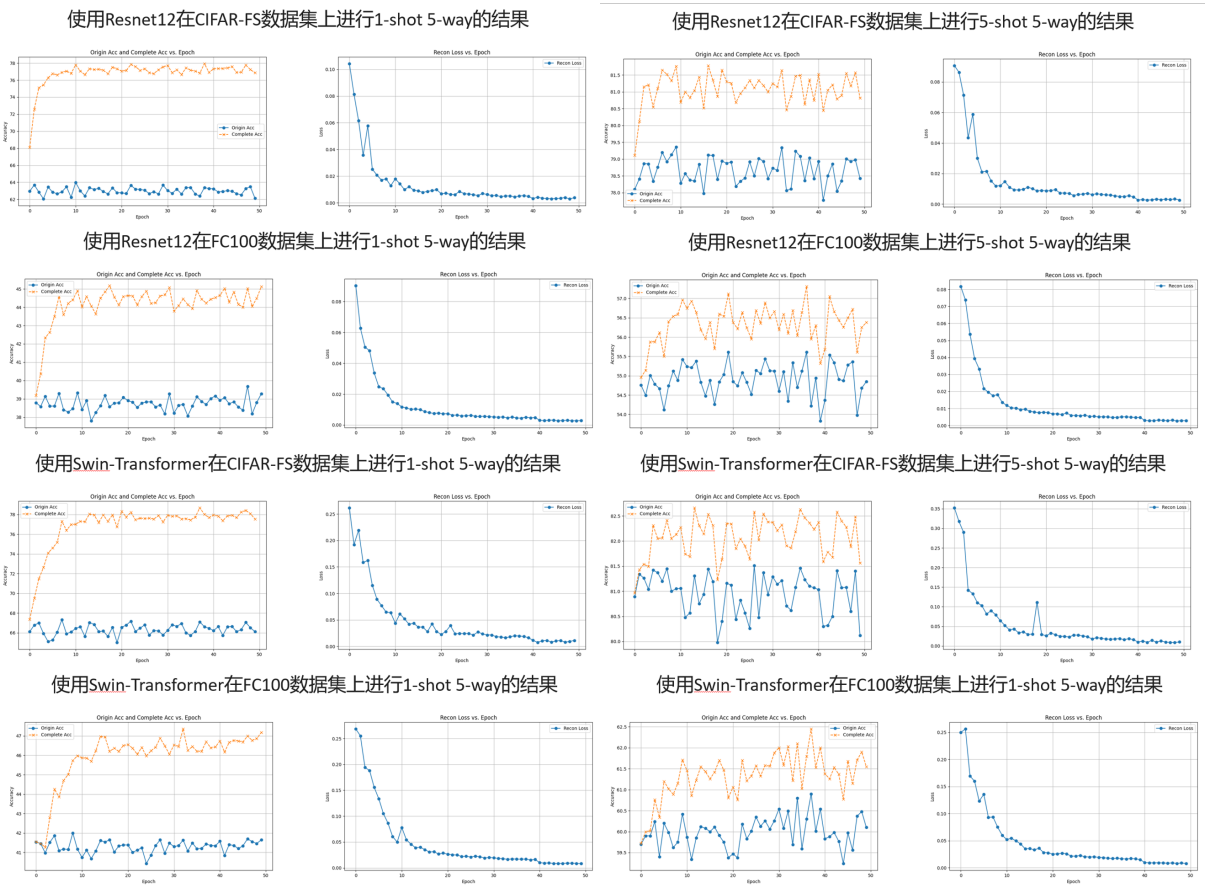

图 4. 使用原论文模型的实验结果

可以发现，在两种主干网络下，特征融合后少样本图像分类的正确率都有稳定的提升，这充分证明了复现结果的正确性。将主干网络替换为 ResNet50 后的实验结果如图5所示，Gap 这一列表示的是该特征融合方法对少样本图像分类正确率的提升。观察 ResNet12 和 ResNet50 在 CIFAR-FS 数据集上实验得到的 Gap，可以发现将主干网络替换为 ResNet50 后，少样本图像分类正确率的提升幅度确实变大了，这一改进的结果符合预期。

Backbone	Dataset	1-shot-5-way	Gap
ResNet12	CIFAR-FS	Origin Acc: 62.11%±1.01%, Complete Acc: 76.86%±0.72%	14.75
ResNet12	FC100	Origin Acc: 38.83%±0.74%, Complete Acc: 44.83%±0.73%	6.00
Swin	CIFAR-FS	Origin Acc: 66.33%±0.98%, Complete Acc: 78.24%±0.74%	11.91
Swin	FC100	Origin Acc: 41.65%±0.71%, Complete Acc: 47.18%±0.72%	5.53
ResNet50	CIFAR-FS	Origin Acc: 43.26%±0.79%, Complete Acc: 61.50%±0.74%	18.24
ResNet50	FC100	Origin Acc: 33.52%±0.61%, Complete Acc: 38.11%±0.67%	4.60
Backbone	Dataset	5-shot-5-way	Gap
ResNet12	CIFAR-FS	Origin Acc: 78.11%±0.70%, Complete Acc: 80.87%±0.64%	2.76
ResNet12	FC100	Origin Acc: 53.98%±0.73%, Complete Acc: 55.61%±0.72%	1.63
Swin	CIFAR-FS	Origin Acc: 80.12%±0.75%, Complete Acc: 81.56%±0.70%	1.44
Swin	FC100	Origin Acc: 59.97%±0.69%, Complete Acc: 61.68%±0.69%	1.7
ResNet50	CIFAR-FS	Origin Acc: 61.18%±0.77%, Complete Acc: 66.82%±0.70%	5.64
ResNet50	FC100	Origin Acc: 48.22%±0.71%, Complete Acc: 49.74%±0.69%	1.52

图 5. 使用 ResNet50 的实验结果

6 总结与展望

本文通过 12 组实验验证了复现结果的正确性，并在已有方法的基础上进行了优化与扩展。通过大语言模型生成高质量语义源，模型的确能够获取更丰富的语义信息，从而更好地结合视觉特征。将主干视觉编码器替换为 ResNet50，使得模型能够捕捉更细粒度的视觉特征，为语义与视觉特征的融合提供更大的对齐空间。实验结果表明，这种改进方案提升了分类正确率的提升幅度，尤其是在 CIFAR-FS 数据集上的表现尤为突出。

综合来看，语义辅助的方法有效地解决了少样本学习中数据稀缺和特征不足的问题，验证了语义与视觉特征融合的重要性和复杂编码器的潜力。未来的研究可以进一步探索如何结合多模态信息，设计更加高效的特征对齐机制，以及在其他复杂场景下的扩展应用。

参考文献

[1] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic prompt for few-shot image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23581–23591, 2023.

[2] Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, and Bihan Wen. Frequency guidance matters in few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11814–11824, 2023.

[3] Markus Hiller, Rongkai Ma, Mehrtash Harandi, and Tom Drummond. Rethinking generalization in few-shot classification. *Advances in Neural Information Processing Systems*, 35:3582–3595, 2022.

[4] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[5] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in neural information processing systems*, 32, 2019.

[6] Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He. Simple semantic-aided few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28588–28597, 2024.