目录

Vision-Language Models are Strong Noisy Label Detectors

摘要：关于微调视觉-语言模型的研究在各种下游任务中都表现出优越的性能，但是在真实世界中，要获取大量完全准确标注的数据几乎不可能，这也成为微调过程中的一个重大挑战。为了解决这一问题，本文提出了一种去噪微调框架，叫做 DEFT（Denoising Fine-Tuning）用于适配视觉-语言模型。DEFT 利用在数百万辅助图文对上预训练的文本和视觉特征的鲁棒对齐，筛选出带噪声的标签。该框架通过对每个类别学习正向和负向文本提示来建立一个噪声标签检测器。正向提示旨在揭示类别的显著特征，而负向提示则充当一个可学习的阈值，用于区分干净样本和噪声样本，采用参数高效的微调方法来调整预训练的视觉编码器，以促进其与所学习的文本提示的对齐。作为一个通用框架，DEFT 能够通过使用精心选择的干净样本，轻松地将许多预训练模型微调到下游任务。在七个合成和真实世界的带噪数据集上的实验结果验证了 DEFT 在噪声标签检测和图像分类任务中的有效性。

关键词：微调；视觉-语言模型；噪声标签检测器；图像分类

1 引言

大规模图文对上的视觉-语言模型（如 CLIP）的预训练在已广泛应用于少样本学习、多标签学习和长尾识别等多种机器学习任务。研究表明，CLIP 在许多下游任务中表现出色的泛化性能，无需适配即可实现高效表现。零样本 CLIP 通过对图像嵌入与文本类别提示之间的相似性进行比较，利用视觉与文本特征的强对齐能力对未训练过的图像进行分类。然而，当下游任务的数据分布显著偏离 CLIP 的训练来源时，微调变得必要。常见的微调范式包括全量微调（FFT），即修改所有模型参数，以及参数高效微调（PEFT），即固定预训练参数，仅添加少量可学习参数用于适配。尽管这些方法表现良好，但微调 CLIP 需要完美标注的数据集，而许多实际任务中难以获得此类高质量数据，从而限制了其广泛应用。

为了解决这一问题，本文研究了在带有噪声标签的数据集上微调 CLIP 的方法。直观来看，直接使用带噪声标签进行适配可能显著降低性能。为了减轻噪声标签的负面影响，研究者提出了多种带噪声标签学习方法，包括样本选择技术和抗噪声学习方法。然而，在 CLIP 适配背景下对这一问题的探索仍然有限。本文旨在填补这一研究空白，展示如何利用 CLIP 的强视觉与文本特征对齐能力，将其作为高效的噪声标签检测器。

2 相关工作

2.1 初步调查

2.1.1 零样本CLIP

零样本图像分类（Zero-shot image classification）指的是使用一个模型对图像进行分类，而这个模型并没有在包含那些特定类别的标记样本的数据上进行过显式训练。传统的图像分类方法需要在一组特定的带标签的图像上训练模型。这个模型通过学习，将图像的某些特征与标签相对应。当需要使用这种模型来处理引入了新标签集的分类任务时，通常需要进行模型的微调，以适应新的标签。与此相反，零样本图像分类模型通常是多模态模型，这些模型在包含大量图像及其相关描述的数据集上进行训练。这些模型学习了视觉和语言之间对齐的表示方法，可以应用于包括零样本图像分类在内的许多下游任务。这是一种更为灵活的图像分类方法，它允许模型在不需要额外训练数据的情况下，泛化到新的和未见过的类别。同时，它也使用户能够用自由形式的文本描述来查询他们目标对象的图像。

CLIP,即对比比语言-图像预训练，是一种自然语言监督学习的高效方法。CLIP是一个联合图像和文本的嵌入模型，通过4亿个图像和文本对以自监督的方式进行训练，将文本和图像映射到同一个嵌入空间中，图像与其文本描述有相近的向量。同时，CLIP可以用于各种未经过训练的任务，在多个基准测试上取得了显著的零样本性能，如 ImageNet，CLIP模型未明确训练于 ImageNet 数据集中的任何 128 万个训练样本。尽管如此，CLIP 的准确度与原始的基于该数据进行训练的 ResNet-50 相当。

CLIP 利用视觉和文本特征的余弦相似度进行零样本分类，其性能可以通过在标注数据集上微调进一步提高，但由于现实场景中的标注数据集通常包含噪声标签，引起的特征扭曲，如何在微调过程中获得鲁棒且可区分的特征表示是核心挑战。

2.1.2 下游数据集上的CLIP微调

为了确定最有效的 CLIP 微调方法，在多个数据集上进行了实证研究，使用三种微调方法对预训练的 CLIP 模型进行适配，并比较它们在带噪声和干净数据集上的性能。其中，FFT是全参数微调，即更新模型的全部参数；VPT是视觉提示调优，即固定预训练模型参数，在视觉编码器前添加一小部分可学习参数进行微调。VLPT是视觉-语言提示调优，即结合视觉和文本的可学习提示，并在固定预训练模型的基础上进行微调。对于 FFT 和 VPT，额外训练了一个线性分类器，而 VLPT 则直接使用学习到的文本提示进行分类。

通过对三种微调方式方法的结果进行分析得到，VPT 在大量噪声标签条件下有利于表示学习，通过 FFT 适配 CLIP 能够利用其强大的模型容量学习任务特定表示，从而提升性能。然而，当数据集中包含噪声标签时，FFT 的效果显著下降。这是由于随着噪声比例增加，特征表示出现严重扭曲。相比之下，在噪声数据上适配 CLIP 时，VPT 有助于表示学习，尤其是在高噪声条件下。由于引入的参数量很小，VPT 能够高效保留图文预训练的泛化能力，同时提升下游任务的分类性能，成为处理带噪声下游数据集的一种稳健且有效的微调方法。在冻结视觉编码器的情况下，文本分类器在小样本学习场景中表现出色。从带有文本分类器的VLPT在分类任务中始终优于带传统线性分类器的 VPT，尤其是在噪声严重的情况下。这种在不同噪声比例下的性能提升进一步验证了可学习文本提示在缓解标签噪声影响方面的鲁棒性。

尽管FFT 比 VPT 和 VLPT 更易受到噪声标签的影响，但是当训练数据干净时，FFT 的判别能力显著提升，这种优势在细粒度数据集上尤为明显。

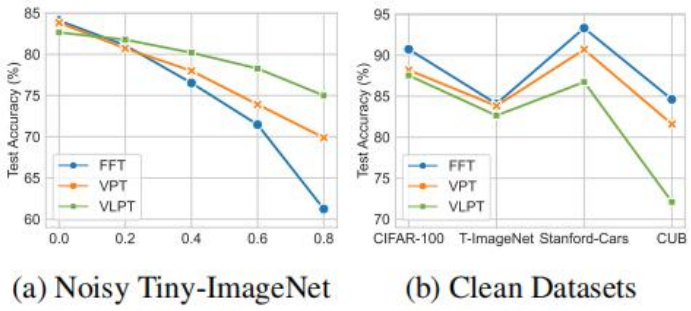

图1 在噪声数据集和干净数据上对三种微调方式的对比

3 本文方法

3.1 本文方法概述

去噪微调框架DEFT的核心设计分为两个部分：第一阶段学习双文本提示（dual textual prompts），用于区分干净样本和噪声样本，使用 PEFT （参数高效微调）方法来适配视觉编码器。第二阶段利用第一阶段筛选出的清洁样本，通过 FFT（全参数微调）重新适配预训练模型，进一步提升视觉识别性能。

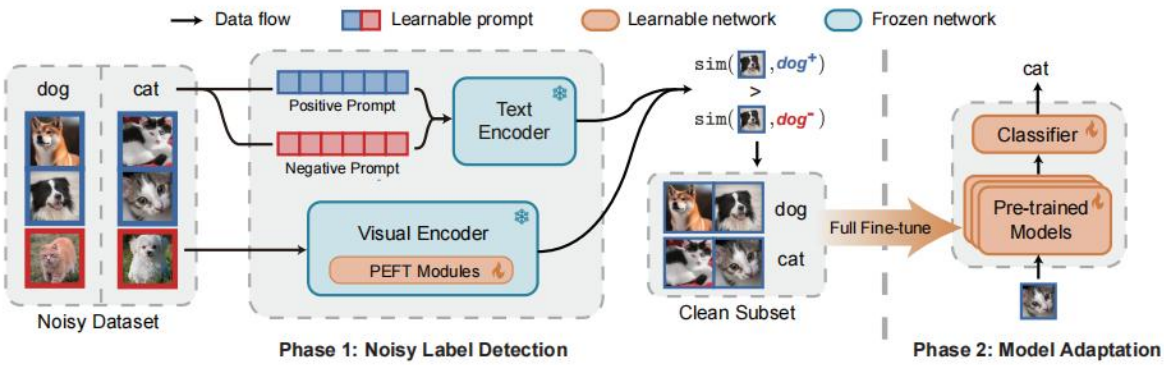

图2　DEFT架构设计

3.2 噪声识别

3.2.1 双提示符识别噪声标签

在带噪标签数据集的下游任务中，标签噪声是模型微调的一大挑战。为了应对这一问题，本文提出了一种基于双文本提示的噪声标签检测方法，通过视觉和文本模态的特性，筛选出干净样本，为模型的进一步训练提供可靠的数据支持。现有噪声标签检测方法通常依赖两种策略。第一种是基于损失的样本选择，通过计算样本的损失值，选择损失较小的样本作为干净样本。这种方法对噪声比例的先验知识敏感，并且容易忽略困难但重要的样本。第二种是固定阈值筛选，通过设定固定的相似性阈值筛选干净样本，但阈值的选取通常需要人工调整，实用性有限。为了克服上述局限性，本文创新性地提出利用正向提示和负向提示的双文本提示机制，以更为自适应和鲁棒的方式识别噪声标签。

双文本提示机制通过结合正向提示和负向提示，充分利用图像和文本模态的特性。正向提示的目标是最大化图像特征与文本特征之间的相似性，用于揭示类别的关键特征，从而增强模型对正确样本的识别能力。负向提示则设定样本相似性的可学习阈值，通过区分清洁样本和噪声样本，为模型提供判定标准。每个样本的阈值由图像特征与负向提示之间的余弦相似度定义清洁样本筛选规则为当正向提示的相似性超过阈值且标签匹配时，样本被认为是清洁样本。为每个类别设计对应的正向提示和负向提示，使用 CLIP 提供的预训练文本编码器初始化提示，随后通过下游任务数据进行优化，计算样本特征与提示相似度，利用 CLIP 的视觉编码器提取图像特征，根据正向相似性与负向相似性的比较，筛选符合条件的清洁样本。使用筛选出的清洁样本，通过全参数微调（FFT）进一步提升模型在下游任务中的性能。

3.2.2 优化噪声标签检测器

端到端优化阈值是一个复杂任务，因为其间接参与了前向传播。但可以引出一个清洁概率p_{ik}^{clean}来表示第i个图像的清洁概率，将原始基于阈值的选择方法可以被转化为K 个二分类任务，用于判断样本是否为清洁样本。p_{ik}^{clean}>0.5时，将样本视为干净样本，反之则视为噪声样本。

$$p_{\text{clean}}^{ik} = \frac{\exp(\text{sim}(I_i, T_k^+)/\tau)}{\exp(\text{sim}(I_i, T_k^+)/\tau) + \exp(\text{sim}(I_i, T_k^-)/\tau)}$$

图 3　对噪声分类器的优化

3.3 基于干净数据的模型适配

由复现初期的调查，确定了FFT在干净数据集上表现最好，因此选择FFT微调方式在筛选后的数据集上进行训练。

4　复现细节

4.1 本论文创新点

我们提出了一种简单而有效的带噪标签学习框架DEFT。与现有方法相比，DEFT具有多个显著优势：基于实例（无需整个训练数据的信息）、对多种类型的噪声标签具有鲁棒性，并且能够广泛适用于多种预训练模型。通过大量实验，证明了DEFT在噪声标签检测和图像分类任务中，在多种合成和真实数据集上均表现出色。提供了深入的实证分析，为理解DEFT的有效性提供了重要见解，并期望本研究能为基于多模态特征的噪声标签检测领域的未来研究奠定基础。

本论文提出了一种全新的带噪标签学习范式，从传统的单模态方法转向多模态框架，通过学习双文本提示，构建了一种创新的噪声标签检测机制。该方法具有多项显著优势，包括对各种类型标签噪声的鲁棒性、对多种预训练模型的广泛适用性，以及无需依赖训练样本动态。通过在合成数据集和真实数据集上的广泛实验，我们验证了所提框架DEFT在噪声标签检测和图像分类任务中的卓越性能。此外，我们进一步强调了参数高效微调（PEFT）相比全参数微调（FFT）在噪声标签场景下的优越性。我们希望本研究能够激发未来在多模态噪声标签检测领域的进一步探索与研究。

4.2 数据集准备

4.2.1 合成数据集

合成数据集的构建严格遵循论文方法，以模拟现实世界中不同类型和比例的噪声标签。对于广泛使用的CIFAR-100和Tiny-ImageNet数据集，首先获取其原始图像数据及对应的真实标签。在引入噪声时，依据噪声转移矩阵来精准控制标签噪声的生成方式。

针对对称噪声，假设噪声概率仅依赖于真实标签，按照论文设定的噪声比例为0.2，0.4或0.6，通过随机替换给定比例的训练样本的真实标签为其他随机标签，实现对称噪声的添加。例如，对于CIFAR-100数据集中原本标注为“cat”的部分样本，在噪声比例为0.4时，有40%的概率将其标签随机更改为其他99个类别中的任意一个，如“dog”“bird”等，从而模拟现实中因标注人员疏忽或错误导致的随机标签错误情况。

对于实例依赖噪声，考虑到样本自身特征对标签噪声的影响，结合样本的视觉特征和已有标签，以一定概率对标签进行调整。如在处理Tiny-ImageNet中一些外观相似但类别不同的花卉图像时，若模型在初步学习过程中对某类花卉的特征把握不够准确，根据当前模型对图像特征的理解以及预设的噪声概率，对部分样本标签进行有针对性的修改，使其更符合噪声分布假设，以模拟实际标注过程中因样本特征模糊引发的标注错误，确保合成数据集能够真实反映复杂的噪声场景。

同时，对于斯坦福汽车（Stanford-Cars）和CUB-200-2011等细粒度数据集，同样依据上述噪声生成策略进行处理。在斯坦福汽车数据集中，由于车辆品牌、型号繁多，不同车型之间的外观差异细微，标注错误时有发生。通过引入不同比例的对称噪声和实例依赖噪声，如在对称噪声比例为0.3时，对部分车辆图像的品牌或型号标签进行随机篡改，使模型在训练过程中面临更具挑战性的噪声干扰，检验其在细粒度分类任务下应对噪声标签的能力。在CUB-200-2011鸟类数据集方面，考虑到鸟类的种类丰富、形态相似，针对不同鸟类的特征差异点，如羽毛颜色、喙的形状等，结合噪声模型，精心调整标签，使合成数据集充分涵盖各种可能的噪声情况，为后续实验提供全面且真实的测试场景。

4.2.2 真实世界数据集

真实世界数据集的选用直接关乎实验结果的实用性与可靠性。CIFAR-100N作为CIFAR-100的变体，其标签来源于亚马逊的Mechanical Turk平台上的人类标注，由于标注人员的多样性和标注标准的不一致性，导致数据集中存在大量噪声标签，噪声比例约为0.4。在使用该数据集时，充分利用其丰富的图像类别和贴近现实的噪声特性，检验DEFT框架在处理实际人类标注错误时的有效性。

Clothing1M数据集聚焦于服装图像分类，包含100万张来自在线购物网站的14个类别的服装图像，同样面临着严重的噪声标签问题，约40%的标签不准确。这主要是由于服装款式繁多、时尚潮流变化快，以及不同商家或标注人员对服装类别的理解差异所致。在实验中，借助该数据集，深入探究DEFT框架在大规模、多类别且噪声复杂的真实场景下的性能表现，特别是在处理服装风格、颜色、款式相近的图像时，观察模型能否准确识别噪声标签并进行有效分类。

WebVision数据集则包含从Flickr和Google上抓取的240万张图像，涵盖了ImageNet ILSVRC12中的1000个概念，选取其中前50类进行实验，其噪声比例约为0.2。该数据集的优势在于数据来源广泛，反映了互联网上图像的多样性和标签的不确定性，通过在该数据集上的测试，能够全面评估DEFT框架在处理大规模网络图像数据时应对噪声的能力，验证其在不同领域、不同噪声水平下的泛化性能，确保复现实验结果能够真实反映模型在实际应用中的表现。

4.3 模型选择与初始化

在基于CLIP的DEFT框架实现过程中，首先精准加载预训练的CLIP模型，包括其基于Transformer的文本编码器和ViT - B/16架构图像编码器。对于文本编码器，依据论文设计，

构建正、负文本提示。以具体分类任务为例，针对“动物”类别下的“狗”类，正提示模板可设计为“a furry animal with four legs and a wagging tail, it is a [CLS]”，其中[CLS]为类别占位符，后续替换为“dog”，通过这种富含特征描述的文本提示，引导模型聚焦于狗类的典型特征；负提示模板则可设计为“a smooth-skinned animal with fins and gills, it is a [CLS]”，与狗类特征形成鲜明对比，以此作为区分干净样本与噪声样本的关键依据。

在训练过程中，分为两个关键阶段。噪声标签检测阶段，利用参数高效微调（PEFT）方法中的视觉提示调整（VPT）对视觉编码器进行适配。具体而言，在模型前向传播过程中，将可学习的视觉提示添加到图像编码器的输入层，通过反向传播优化这些提示参数，使得图像特征与正文本提示特征的对齐程度不断提升。同时，依据构建的正、负训练样本，结合损失函数进行迭代训练。在每一轮迭代中，仔细计算样本的干净概率，根据正、负提示与图像特征的相似度，精准筛选出潜在的干净样本，构建干净子集。

模型适配阶段，移除PEFT模块，采用全量微调（FFT）方法，利用筛选出的干净子集对预训练模型进行全面优化。在此阶段，学习一个线性分类器，通过最小化经典交叉熵损失，促使线性分类器与干净样本的特征分布高度适配。在代码实现中，精准设置优化器SGD的参数，包括学习率、动量和权重衰减，对模型参数进行精细调整，确保模型在下游任务中的性能得到显著提升。

4.4 实验设置细节

4.4.1 超参数调整策略

在复现实验中，除了论文中提及的基本超参数，如学习率、权重衰减、批次大小等，对其他超参数对模型性能的影响进行了测试。

对于文本提示长度，在一定范围内进行了细致的对比实验。当文本提示较短时，模型可能无法充分捕捉到类别特征的丰富语义信息，导致分类准确性受限。例如，将文本提示长度从4逐步增加到32，发现在CIFAR-100数据集上，当文本提示长度为16时，模型在噪声标签检测和图像分类任务中的综合性能达到最优。这是因为长度为16的文本提示既能涵盖足够的关键特征描述词汇，又避免了因过长导致的信息冗余和计算复杂度增加。

视觉提示长度同样对模型性能有着不可忽视的影响。以VPT为例，在默认设置视觉提示长度为20的基础上，分别尝试了10、15、20、30等不同长度值。实验结果表明，在Tiny-ImageNet数据集上，当视觉提示长度为20时，模型在处理高噪声比例数据时表现更为稳健。这是由于合适的视觉提示长度有助于更好地引导视觉编码器学习与文本提示相匹配的特征，增强模型对噪声的鲁棒性。

温度参数在计算预测概率时起着调节作用。较小的值会使模型在相似度计算后得到的概率分布更加“尖锐”，即对高相似度样本赋予更高的概率，对低相似度样本赋予极低的概率；而较大的值则会使概率分布相对“平滑”。在不同数据集上通过调整发现，在斯坦福汽车数据集上，当取值为0.1时，模型在细粒度分类任务中对相似车型的区分能力更强，能够更精准地识别噪声标签，因为此时模型对图像与文本提示之间的相似度差异更为敏感，有助于筛选出更可靠的干净样本。

综合考虑这些超参数的相互作用，在复现过程中依据论文建议及实际实验结果，最终确定了一套最优的超参数组合，以实现模型复杂度与性能表现的精妙平衡，确保模型在各类数据集上均能发挥出最佳性能。

4.4.2 训练技巧应用

在训练过程中，运用多种训练技巧能够显著提升模型性能与训练效率。

在训练初期，采用相对较大的学习率，如论文中在噪声标签检测阶段初始学习率设置为0.03，这有助于模型快速探索参数空间，捕捉数据的大致特征。随着训练的推进，逐渐降低学习率，例如在训练后期将学习率衰减为原来的十分之一，即0.003。这是因为在训练后期，模型已经接近收敛，过大的学习率容易导致模型在最优解附近来回震荡，无法稳定收敛，而过小的学习率又会使训练过程过于缓慢。通过学习率衰减，模型能够在训练后期更加精细地调整参数，避免过拟合，进一步优化模型性能。

梯度裁剪也是重要的优化手段。由于在模型训练过程中，尤其是在面对复杂数据集和深度神经网络架构时，可能会出现梯度爆炸问题，导致模型参数更新幅度过大，无法收敛。采用梯度裁剪技术，如将梯度的范数限制在一定范围内，可有效避免这一问题。在复现实验中，根据不同模型和数据集的特点，将梯度范数限制在1.0或0.5左右，确保模型在训练过程中参数更新稳定、有序，避免因梯度异常导致的训练失败。

早停法的应用进一步提高了训练效率。在模型训练过程中，持续监测模型在验证集上的性能，如每经过一个训练周期，便评估模型在验证集上的准确率或损失值。一旦发现模型性能在连续多个周期（如5个周期）内不再提升，甚至出现下降趋势，便提前终止训练。这不仅节省了大量不必要的计算资源，还能有效防止模型因过度训练而出现过拟合现象，确保模型在最佳状态下完成训练，提高模型的泛化能力，使其能够更好地适应未知的测试数据。

5 实验结果分析

5.1 噪声标签检测性能评估

在噪声标签检测任务中，DEFT框架展现出卓越性能，与论文中的基线方法形成鲜明对比。表1详细列出了在CIFAR-100、Tiny-ImageNet、Stanford-Cars、CUB-200-2011等多个数据集上，DEFT与Label-match策略、Small-loss策略在精度和召回率方面的对比数据。

在CIFAR-100数据集上，当噪声类型为对称噪声且比例为0.2时，DEFT的精度达到99.51%，召回率为97.77%，而Small-loss策略的精度为97.24%，召回率为96.79%。此时，DEFT相较于Small-loss策略，精度提升了2.27个百分点，召回率提升了0.98个百分点。随着噪声比例增加到0.6，DEFT的优势愈发显著，精度为97.04%，召回率为94.08%，Small-loss策略的精度仅为92.93%，召回率为90.33%，DEFT的精度提升了4.11个百分点，召回率提升了3.75个百分点。

类似地，在Tiny-ImageNet数据集上，60%对称噪声时，DEFT的精度达到97.21%，召回率为95.44%，Small-loss策略的精度为92.63%，召回率为90.89%，DEFT在精度上提升了4.58个百分点，召回率提升了4.55个百分点。

为了更直观地呈现对比效果，绘制柱状图。从图中可以清晰看出，在不同数据集和噪声比例下，DEFT的精度和召回率大多高于基线方法。在Stanford-Cars数据集上，面对实例依赖噪声，DEFT始终保持较高的精度和召回率，相较于其他方法优势明显，这充分表明DEFT在复杂噪声场景下能够更精准地识别噪声标签，为后续模型训练筛选出更可靠的干净样本。

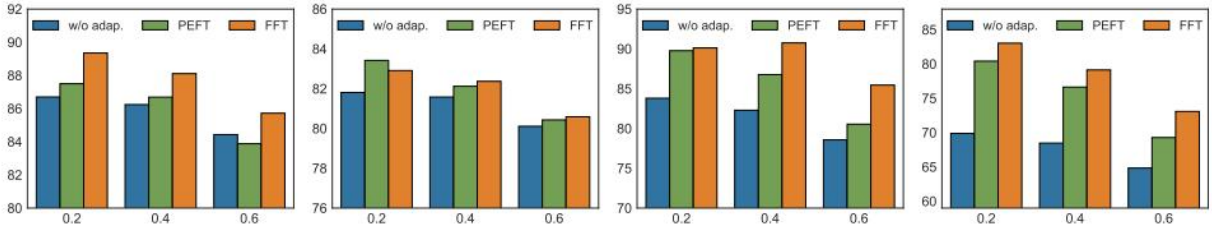

图4 实验部分结果可视化

5.2 对不同噪声场景的适应性分析

DEFT框架对不同噪声场景展现出出色的适应性，尤其是在高噪声比例、细粒度分类任务等复杂情况下。表1数据显示，在CUB-200-2011数据集上，面对实例依赖噪声且比例为0.4时，DEFT的精度为96.43%，召回率为97.11%，F1分数达到较高水平。相比之下，Label-match策略和Small-loss策略在该复杂噪声场景下的表现欠佳。

进一步分析发现，DEFT能够有效识别具有挑战性的硬噪声，得益于其利用预训练视觉语言模型中的多模态信息。在处理细粒度分类任务时，如鸟类图像分类，图像中的鸟类可能因姿态、羽毛颜色等细微差异导致标注困难，产生噪声标签。DEFT通过正、负文本提示，结合图像特征，能够精准捕捉到不同鸟类类别之间的细微特征差异。正提示聚焦于鸟类的独特特征，如某种鸟类特有的喙形、羽毛图案等，负提示则作为区分阈值，帮助筛选出与正提示特征匹配度高的干净样本，从而有效应对复杂噪声模式，确保模型在实际应用中的可靠性。

6 总结与展望

尽管 DEFT 相较于现有方法表现优异，但其仍存在一些局限性，并有多种未探索的研究方向。例如，当前算法仅处理单标签多分类任务中的噪声图像-标签对。未来研究的一个有前景方向是将 DEFT 泛化到处理噪声图像-文本对或多标签分类任务的场景。

参考文献

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[2] Natarajan, Nagarajan, et al. "Learning with noisy labels." Advances in neural information processing systems 26 (2013).

[3] Zhang, Jingyi, et al. "Vision-language models for vision tasks: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[4] Zhang, Pengchuan, et al. "Vinvl: Revisiting visual representations in vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In European Conference on Computer Vision, pages 709–727, 2022.

[6] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9676–9686, 2022.

[7] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19113–19122, 2023.