

突破简单管道的极限，实现快速学习：外部数据和微调大有裨益

Shell Xu Hu¹Da Li¹ *Jan Stühmer^{1*}Minyoung Kim^{1*}

Timothy M. Hospedales

^{1,21} 剑桥三星人工智能中心 ² 爱丁堡大学

{shell.hu, da.li1, jan.stuhmer, k.minyoung, t.hospedales}@samsung.com

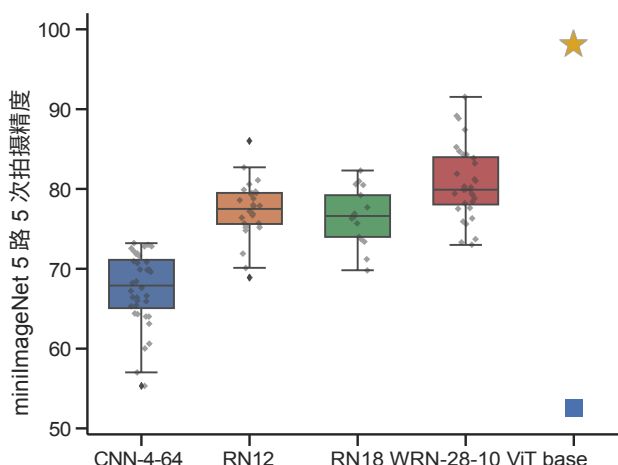
摘要

少镜头学习 (FSL) 是计算机视觉领域的一个重要而热门的问题，促使人们对从复杂的元学习方法到简单的迁移学习基线等众多方法进行了广泛的研究。我们力图突破简单但有效的管道极限，使其适用于更现实、更实用的少帧图像分类设置。为此，我们从神经网络架构的角度探索了少拍学习，以及在不同数据供应条件下网络更新的三阶段管道，其中无监督外部数据用于预训练，基础类别用于模拟少拍任务以进行元训练，而无标记任务的稀缺数据则用于微调。我们对以下问题进行了研究

这些问题包括^① 外部数据的预培训如何有利于

FSL？^② 如何利用最先进的变换器架构？最终，我们展示了基于转换器的简单流水线在 Mini-ImageNet、CIFAR-FS、CDFSL 和 Meta-Dataset 等标准基准上产生了令人惊讶的良好性能。我们的代码和演示可在

<https://hushell.github.io/pmf> 上获取。



" (FSL) 的大量且不断增长的研究，其目的是模仿人类从少量训练示例中学习新概念的能力。事实证明，FSL 挑战是开发和测试大量复杂研究想法的沃土，这些想法包括度量学习 [59, 61]、基于梯度的元学习 [29]、程序归纳 [41]、可微分优化层 [42]、hy-

*平等贡献。

1. 引言

在有大量注释数据集的应用中，主流的有监督深度学习取得了优异的成绩。然而，在许多应用中，数据（如稀有类别）或人工标注成本是令人望而却步的瓶颈，因此无法满足这一假设。这就推动了对 "少量学习

图 1. 预训练和架构如何影响少镜头学习？通过以下方法可以实现少镜头学习：a) 元学习（meta-learning）[66, 72]；b) 在大规模外部数据上预先训练的自监督基础模型的迁移学习[18, 53]。虽然大多数 FSL 社区都将重点放在前者上，但我们的研究表明，后者可能更有效，因为它可以使用更强大的架构，如视觉转换器（ViT）[25]--并且可以与 ProtoNet 等简单的元学习器相结合。图中显示了过去 5 年 FSL 研究中数十项研究的结果汇总，以及 ProtoNet + ViT 骨干的结果 + 对比语言图像预训练 (CLIP) [53]（黄星）。为了强调预训练的重要性，还比较了 ProtoNet + 随机初始化 ViT（蓝色方块）。

神经网络[9]、神经优化器[54]、传导标签传播[55]、神经损失学习[4]、贝叶斯神经先验[72]等[69]。但是，在所有这些技术进步的基础上，我们又取得了多少实际进展呢？

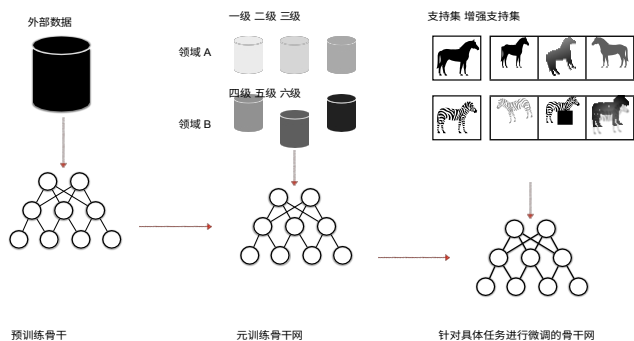
有一些研究[19, 20, 23, 51, 63, 68]探讨了更简单的基线是否能提供与最先进的几发学习器相媲美的性能。虽然目前还没有定论，但由于复杂学习器[72]和简单基线的不断发展，有一种趋势是，简单方法的性能往往令人吃惊

与复杂的同类方法相比，它们的效果更好。它们的简单性和有效性使得这些简单方法被广泛应用于从医疗数据分析[11]到电子工程[40]等许多实际应用中。

我们沿用了这一思路，但更进一步研究了以前未充分研究的、影响简单少量学习管道性能的因素。特别是，我们从 ProtoNet [59] 少量学习器入手，研究了三个实际重要的设计选择：预训练数据、神经网络架构和元测试时间微调。

源数据 虽然 FSL 解决的是小数据机制，但实际上 FSL 研究几乎总是关于从大规模源任务（又称元训练）向小规模目标任务（又称元测试）转移知识的算法。现有文献几乎总是控制源数据，以便仔细比较从超网络[9]到基于梯度的元学习器[29]等不同知识转移机制的影响。虽然这有助于推动对复杂算法的研究，但并不能回答源数据的选择如何影响性能？视觉和模式识别的其他领域已经对这个问题进行了研究[10, 31, 60]，但 FSL 却没有。这对于计算机视觉 FSL 研究的消费者来说毫无帮助，因为他们很想知道简单地改变源数据能在多大程度上改善他们的应用？特别是因为免费提供的大型数据集已经存在 [21, 62]，而且在实践中利用更多外部源数据比实施复杂的先进元学习器更容易。为此，我们研究了外部数据无监督预训练（最近称为利用基础模型的工作流程[10]）对 FSL 任务的影响。与 5 年的 FSL 研究相比，这一微小的变化产生了巨大的影响（图 1）。虽然这可能违反了严格规定源集的 FSL 问题定义，但该方法的功效可能会促使人们反思这是否是最好的问题定义。

神经架构 与源数据的情况类似，FSL 研究通常将神经架构控制在 CNN-4-64 和 ResNet-12 等少数几个小型网络中。这样做的部分原因是为了公平比较 FSL 算法，但这种特定的网络套件也是迷你图像网络等常用基准中用于训练的源数据集规模较小的结果。因此，就最先进的计算机视觉技术而言，FSL 通常研究的架构有些过时。因此，我们要问的是，视觉转换器 [25] 等最先进的架构能在多大程度上提高少拍性能，尤其是与



更大的预训练数据集结合使用时？

微调 FSL 文献中的许多研究在是否主张 [29,54,65] 在针对单个任务的模型部署（又称元测试）过程中进行某种微调，或者固定的特征表示是否就足够了方面存在一定分歧 [42, 59, 68]。我们还研究了

图 2：概览概览--我们所考虑的简单而有效的流程示意图：

预训练 → 元训练 → 微调 ($P > M > F$)。按照红色箭头所示，该流程将与类无关的特征骨干转换为通用特征骨干，并最终转换为与特定任务相关的特征骨干。

这个问题，并提出*微调对于将基础模型用于分布外任务是必要的*。我们还介绍了微调的算法改进，通过验证自动选择学习率，从而为跨域 FSL 提供性能更高的管道。

总之，我们通过研究一个简单管道的设计选择 [59] (图 2)，而不是开发新的算法，来推进少量学习。我们回答的问题包括*预训练对 FSL 有何影响？最近的变压器架构能否适应 FSL？如何最好地利用微调？*基于上述分析，我们展示了一种新的 FSL 基线，其性能超越了最先进的技术，同时简单易实现。

2. 相关工作

快速学习 快速学习目前是一个深入而广泛的研究领域，其规模之大无法在此详述，我们将参考相关调查报告以了解其概况[35, 69]。关键的一点是，尽管名称相同，但几乎所有的 FSL 方法都提供了将知识从大量源数据转移到一组稀疏注释的目标兴趣类别的算法。该领域的许多活动都属于元学习 (meta-learning) 的范畴[35]，其目的是通过模拟少量学习问题，从源数据集 (又称元训练数据集) 构建数据高效的学习器，然后在目标数据集 (又称元测试数据集) 上部署定制的学习器。由此产生的学习器可以是初始化[29]、学习度量[59]、贝叶斯先验[72]或优化器[54]等形式。

简单而有效的基线 与上述大量复杂的 "少量学习"[35, 69]相比，最近的一些研究提出了性能相当好而又比较简单的强大基线。这些学习器通常基于迁移学习[70]管道。它们在源数据上应用传统的深度学习器，然后通过训练一个简单的线性[19, 51, 63]或中心点[68]分类器来适应少量目标数据。

这些方法大多使用标准化的 FSL 源数据集（如 miniImageNet）和架构（如 ResNet-12 和 WRN-10-28）。这些方法大多使用标准化的 FSL 源数据集（如 miniImageNet）和架构（如 ResNet-12 和 WRN-10-28），以便将所提倡的简单基线与先进的学习器进行直接比较。相比之下，我们的具体目标是通过利用其他可用的预训练数据集和架构，探索 FSL 的实用性能可以提升到什么程度。

一些研究利用 ImageNet1K [20] 或 ImageNet21K [23] 等数据集对 FSL 进行了更大规模的评估。然而，由于同时改变了源数据集和目标数据集，这并不能清楚地说明源数据的选择/规模对特定目标问题的影响，而这正是我们要回答的问题。还有一些人探讨了元学习之前的传统预训练[20]或元学习过程中的正则化[30]的影响，但没有利用额外的数据。

更大的数据和架构 在视觉领域的 标准监督[60]和自我监督[10, 31]学习以及视觉之外的模式识别应用[3,10,13,22]中，源数据集的影响被广泛研究。然而，它并没有在 FSL 中得到广泛评估，这是一个令人惊讶的疏忽，因为正如我们将看到的那样，它很可能是提高实际 FSL 性能的最简单方法。同样，正在使用的 FSL 方法几乎都是基于一些不太常见的架构（如 Conv-4-64 和 ResNet-12），这可能是由于在 Omniglot [29, 66] 等小型数据集上的首次实验设置所致。变换器在 FSL 中的应用有限，主要用于度量学习 [24]，但不用于特征提取。我们将探讨如何训练最近的变换器特征提取器并将其应用于 FSL，尤其是与在较大源数据集上预先训练好的基础模型 [10] 相结合时。

自监督和少镜头 我们的管道扩展了自监督研究界典型的无监督预训练→监督微调工作流程[28, 39]，该流程最近在少镜头监督学习方面表现出色[15, 18, 27]。然而，由于典型的评估实践和基准不同，自我监督（SSL）和 FSL 社区方法在数据高效学习方面的直接比较还很有限。例如，许多 SSL 评估在 ImageNet 上执行无监督代表学习，然后再在 ImageNet 内执行少量监督学习

[15,18]，这违反了 FSL 社区通常要求的源数据和目标数据不相交的规定。本文的贡献之一是对 SSL 和 FSL 方法进行了一定程度的比较和组合。例如，我们的 MetaDataset、CDFSL 和预告图 1 结果都使用了不相连的源数据和目标数据，但都受益于外部自监督预训练。

跨域少拍 FSL 的一个特别实用的变体是跨域少拍[33]。

目标/元测试数据集。这比标准的域内设置更具挑战性，但更切合实际。这是因为在医学或地球观测成像 [33] 等许多需要使用 FSL 的场景中，FSL 的目标数据与可用的源数据（如（小型）ImageNet [21]）有很大不同。这类基准主要有 CDFSL [33] 和元数据集 [65]。

3. 用于 FSL 的简单管道

问题表述 少量学习（FSL）的目的是只用少量注释示例来学习模型。Vinyals 等人 [66] 从元学习的角度提出了一种被广泛采用的 FSL 问题表述，其假设是，我们应该根据以前所见的许多类似的少量任务的经验来学习解决新的少量任务。因此，FSL 问题通常分为两个阶段：在训练任务分布的基础上元训练少次元学习器，以及通过在新的少次元任务上评估学习器来元测试学习器的结果。在每个阶段中，数据以偶发方式到达，每个任务的“训练集”和“测试集”分别称为支持集和查询集，以避免术语混淆。就分类而言，每集的难度被描述为 K -way- N -shot，相当于在支持集中每个类别有 N 个示例的情况下，学习 K 个类别的分类器。为每个难度级别学习一个模型很常见，但更现实的设置 [65] 是为不同的 K 和 N 学习一个全局模型。这种方法有时被称为“多路-多射”，我们在这里讨论的就是这种更实用的方法。这也是我们更倾向于选择简单管道而非复杂元学习器的原因，因为复杂元学习器可能不容易扩展到各种路径-变量-射程设置。

小数据学习的另一种方法出现在迁移学习 [12, 70] 和自我监督 [10, 17] 文献中。在这种情况下，人们使用一些大型源数据对模型进行预训练，然后将其重新用于感兴趣的稀疏数据目标任务。预训练步骤旨在降低适应步骤中学习目标问题的样本复杂度。

虽然这两类方法通常是分开研究的，但它们都提供了从源数据到目标少数问题的知识转移机制。为了实现高性能少量数据学习的目标，我们将预训练（通常是在无标签的辅助数据上进行，这些数据可以免费获

得）和元学习（带有标签的偶发训练）结合在一起，使用一个单一的特征提取器--后骨，形成一个简单的顺序管道。我们的管道包括三个阶段：1) 使用自我监督损失在未标记的外部数据上对特征骨干进行**预训练**；2) 使用 ProtoNet [59] 损失在已标记的模拟少量任务上对特征骨干进行**元训练**；3) 在新颖的少量任务上部署特征骨干。

任务，并可对每个任务的增强支持集进行微调。我们的流程示意图如图 2 所示，我们称之为 P>M>F（即流程 预培训

→ 元训练 → 微调）。接下来，我们将概述如何在不同阶段更新特征骨架。

3.1. 骨干预培训

我们认为 ResNet [34] 或 ViT [25] 的特征骨干可为我们的管道提供基础模型。在预训练步骤中，有几种成熟的自监督学习算法：DINO 算法 [15] 使用 ImageNet1K，利用同一图像的大面积裁剪和多个局部裁剪之间的预测一致性，在 ImageNet 图像中，大面积裁剪极有可能与前景物体重叠；BEiT 算法 [6] 使用 ImageNet1K，利用同一图像的大面积裁剪和多个局部裁剪之间的预测一致性，在 ImageNet 图像中，大面积裁剪极有可能与前景物体重叠。

相当于解决了在

与针对文本数据的原始 BERT 预训练 [22] 保持一致；CLIP [53] 利用 YFCC100m 数据集中的图像标题来对齐图像和文本数据。

在共同的特征空间中捕捉表征。对于像 ViT [25] 这样更灵活的架构，在外部数据上进行预训练非常重要，因为它们很难在常见的小型 FSL 基准上进行训练（图 1 和表 1）。

3.2. 使用 ProtoNet 进行元培训

由于我们的目标是建立一个简单的管道，因此我们考虑使用原型网络（ProtoNet） [59]，它为每个情节动态构建类中心点，然后执行最近中心点分类。具体来说，ProtoNet 只需要一个特征骨干 f 来将数据点映射到 m 维特征空间： $f: X \rightarrow R^m$ ，查询图像 x 属于类别 k 的概率为

$$p(y = k|x) = \frac{\exp(-d(f(x), c_k))}{\sum_k \exp(-d(f(x), c_k))} \quad (1)$$

其中， d 在我们的工作中通过余弦距离来实现

\sum

\sum

算法 1 PyTorch 微调伪代码

```
# 输入: 包括 supp_x, supp_y 和 query_x 的任务 #
backbone_state: 经过元训练的主权重
# 优化器: Adam optimizer #
Outputs: logits

backbone = create_model_from_checkpoint(backbone_state)

def single_step(z):
    supp_f = backbone(supp_x)
    proto = compute_prototypes(supp_f, supp_y)
    f = backbone(z)

    logits = f.norm() @ proto.norm().T # cos 相似性损失 =
    cross_entropy_loss(logits, supp_y)
    返回对数, 损失

# 微调循环
for i in range(num_steps):
    aug_supp_x = rand_data_augment(supp_x)
    loss = single_step(aug_supp_x)
    loss.backward() # back-prop
    optimizer.step() # 梯度下降法

logits, _ = single_step(query_x) # 分类
```

我们的微调算法与文献 [33, 43] 类似，后者使用支持集微调模型权重，因为支持集是元测试时唯一可访问的标注数据。我们对支持集的利用方式略有不同：我们使用数据扩增来创建源自支持集的伪查询集；因此，我们不需要使用支持集计算原型，然后再使用公式 (1) 在同一支持集上应用原型。此外，我们只需更新整个骨干网，而无需探索部分模型适应性。

学习率选择 我们发现，微调性能对学习率的选择相对敏感（更多分析见补充材料）。然而

现有的几次学习问题表述无法为每个任务提供一个验证集来选择最佳学习率进行微调。之前的研究 [33, 43] 选择了一个学习率。

而不是通常选择的欧氏距离， c_k 是类别 k 的原型，定

义为 $c_k = \frac{1}{N_k} \sum_{i: y_i = k} f(x_i)$ 和 $N_k = \sum_{i: y_i = k} 1$

N_k i i

这就使得 ProtoNet 可以在各种不同的拍摄设置下进行训练和部署。

3.3. 元测试与微调

为了与元训练保持一致，默认情况下，我们会在所有新任务中直接使用元训练的 ProtoNet。但是，如果新任务来自未曾见过的领域，学习到的特征表征可能会因为数据分布的巨大变化而无法泛化。为此，我们建议在数据增强的帮助下，通过几个梯度步骤对特征骨干进行微调。具体细节以 PyTorch 伪代码的形式总结在算法 1 中。

并为每项任务固定它。这种策略需要对主干网架构有很好的了解，但一般情况下仍会导致性能达不到最优。

。给定任务

在只有极少数标注图像（即支持集）的情况下，它是

几乎不可能确定哪种学习率能对无标签图像（即查询集）产生良好的泛化效果。好消息是，我们根据经验发现，在同一领域的不同任务中，最佳学习率是相对稳定的。为此，我们建议从每个领域中抽取 $N = 5$ 个额外任务，并在合理范围内（如 $\{0.01, 0.001, 0.0001, 0\}$ ）

自动搜索领域学习率。然后将最佳学习率用于领域内的每个任务。这一额外步骤相当于为每个领域准备几张带标签的图像来创建验证集，这在实践中很有意义，因为我们可以很容易地按领域组织任务，并在搜索后为单个任务识别领域以查找相应的学习率。

4. 实验

miniImageNet [66] 包含 ImageNet-1k 中的 100 个类别，然后将其分为 64 个训练类、16 个验证类和 20 个测试类；每幅图像的采样率降至 84×84 。**CIFAR-FS** [8] 是通过将原始 CIFAR-100 分成 64 个训练类、16 个验证类和 20 个测试类而创建的。图像大小为 32×32 。**Meta-Dataset** [65] 包含 10 个不同领域的公共图像数据集：这些数据集包括：ImageNet-1k、Omniglot、FGVC- Aircraft、CUB-200-2011、Describable Textures、QuickDraw、FGVCx Fungi、VGG Flower、Traffic Signs 和 MSCOCO。每个数据集都分为训练/验证/测试三个部分。我们分别采用了 [65] 和 [24] 提出的两种训练协议。对于前者，前 8 个数据集（域内）的训练/验证分集用于元训练和验证，所有数据集的测试分集用于元测试。元测试只使用 ImageNet-1k 的训练分集进行元训练，其他设置保持不变。有关元数据集的更多详情，请参阅 [65] 的附录 3。

评估 为了评估少镜头分类的性能，我们模拟了每个相关数据集的 600 个测试分集/任务。评估指标是任务的平均分类准确率。对于 miniImageNet 和 CIFAR-FS，常规方法是评估 5 路-1-镜头 (5w1s) 和 5 路-5-镜头剧集，每个剧集的查询集大小固定为 15×5 。对于 Meta-Dataset，除 ImageNet-1k 和 Omniglot 外（它们根据类的层次结构有特定的抽样策略），其他方式、镜头和查询图像的数量都是根据数据集的规格随机抽取的。此外，我们还对 miniImageNet 的 (5w5s) 元训练模型进行了跨域评估 (CDFSL) [33]，其中考虑了 4 个域外数据集，并报告了 5 路-5/20/50 张设置下的重新结果。

训练细节 为了避免针对不同数据集和架构进行过度训练，我们采用了一种通用的训练策略，即从预先训练好的模型检查点 (ResNet 和 ViT) 对骨干进行元训练。这在某些情况下可能会导致次优结果，但却简化了比较。具体来说，我们对骨干网进行 100 个历元的训练，每个历元包含 2000 个集/任务。我们采用热

身加余弦退火的学习率计划：学习率从 10^{-6} 开始，在 5 个 epoch 中增加到 5×10^{-5} ，然后通过余弦退火逐渐降低到 10^{-6} 。我们使用验证集来决定何时提前停止，并关闭强正则化和数据增强技术以简化操作。

4.1. 分析

现在，我们将利用第 3 节中概述的管道来回答一系列有关少量学习者管道设计的问题。

身份 证	培训配置			基准结果		
	拱门	火车前	MetaTr	MD	迷你IN	CIFAR
0	ViT-small	DINO (IN1K)	-	67.4	97.0	79.8
1	ViT-small	DeiT (IN1K)	-	67.5	98.8	84.6
2	ResNet50	DINO (IN1K)	-	63.8	91.5	76.1
3	ResNet50	Sup. (IN1K)	-	62.4	96.4	82.3
4	ViT-small	DINO (IN1K)	PN	78.4	98.0	92.5
5	ViT-small	DeiT (IN1K)	PN	79.3	99.4	93.6
6	ViT-small	-	PN	52.8	49.1	59.8
7	ResNet50	DINO (IN1K)	PN	72.4	92.0	84.0
8	ResNet50	Sup. (IN1K)	PN	70.2	97.4	87.6
9	ResNet50	-	PN	62.9	72.2	68.4
10	ResNet18	-	PN	63.3	73.7	70.2
11	ViT 基础	DINO (IN1K)	PN	79.2	98.4	92.2
12	ViT 基础	CLIP (YFCC)	PN	80.0	98.1	93.2
13	ViT 基础	超级 (IN21K)	PN	81.4	99.2	96.7
14	ViT 基础	BEiT (IN21K)	PN	82.8	99.0	97.5
15	ResNet50	CLIP (YFCC)	PN	75.0	92.2	82.6

表 1.架构和预训练算法（数据集）对 Meta 数据集（MD）、miniImageNet（miniIN）和 CIFAR-FS 的下游 few-shot 学习性能的影响。Meta 数据集的结果是所有目标数据集的平均值，而 minIN 和 CIFAR 的结果则是 5 路 5 次学习的结果。在元测试期间，ProtoNet (PN) 最近中心分类器始终用于支持集上的少量学习。MetaTr 表示在相应基准上用于偶发学习的算法。

值得注意的是 *预培训制度如何影响 FSL？当代架构（如 ViT）能否适应 FSL？*

4.1.1 预培训和架构

我们首先评估了预训练机制（包括算法和数据集）以及神经架构对 FSL 基准 Meta-Dataset [65]（在 8 个数据集上训练）、miniImageNet [66] 和 CIFAR-FS [8] 的影响。为了清楚地了解每个实验的配置，表 1 中的结果按照架构、预训练算法（和数据集）以及元训练算法进行了分类。我们假定 ProtoNet（最近中心点）分类器是元测试的标准方法，并比较了预训练和元测试之间的元学习步骤（MetaTr 列），要么是经过偶发训练的 ProtoNet，要么是什么都不训练。

预训练机制对 FSL 有何影响？ 根据表 1 中的结果，我们可以得出以下结论：(i) 与之前未使用预训练的工作所使用的传统管道相比，ImageNet1K 上的预训练一般都有显著改善（将模型 M9 与 M7 和 M8 等进行

比较）。(ii) 我们主要关注的是无监督预训练，而将有监督预训练作为一个不公平的上限。然而，使用 DINO 进行无监督预训练的最新技术表现接近于有监督预训练（比较 M3 与 M2 等）。这一点值得注意，因为虽然某些源语言之间存在一些语义重叠，但这些语义重叠并不意味着没有监督预训练。

①
②
③

①

(ii) 在本文考虑的目标数据集（ImageNet1K）和 Meta-

Dataset、miniImageNet、CI-FAR）中，不使用源标签也能实现良好的性能，在这种情况下，不存在训练-测试标签泄漏¹。(iii) 在 DINO 等强大的预训练机制下，基于预训练特征的简单最近中心点分类性能良好（顶部区块包括 M2 等）。特别是，与 ProtoNet-ResNet18 的传统特定数据集训练（M2 对 M10）相比，没有特定数据集元学习的 foundation 模型中的现成特征表现更佳，而 ProtoNet-ResNet18 可以说是最接近 FSL 行业标准的。(iv) 尽管如此，特定数据集元学习确实能进一步提高性能（M7 对 M2 等）。冻结基础模型的简单线性读出[18, 27]不具竞争力。

② ViT 等最先进的架构能否适用于 FSL？ 利用表 1 中的结果，我们也可以回答这个问题。特别是，与较小的架构相比，ViT 在较小的元训练基准（miniImageNet、CIFAR）上的训练效果不佳（见 M6 vs M9、M10），但在受益于大型预训练数据时（M6 vs M4），ViT 通常表现优异。总体而言，当受益于预训练时，ViT 的表现全面优于行业标准 ResNet18 和我们的 ResNet50 基线。我们注意到，我们的 ResNet50 基线在没有预训练的情况下表现也相对较差，尤其是在较小的 miniImageNet 和 CIFAR 上，这表明它也过于庞大，无法单独在目标数据集上进行良好的训练。

其他基础模型 总体而言，我们可以看到，较大的预训练数据源和最新的体系结构对下游 FSL 在标准基准上的性能影响巨大。我们还比较了 M11-15 中的一些其他基础模型[10]。我们可以看到：(i) 所有基础模型在标准数据集内训练（M10, M9）上都有大幅提高；(ii) 使用 ViT-base 和 ImageNet21K 或 YFCC 数据源等的最大基础模型在各方面都有最强的表现，但并没有明显优于更经济的基于 DINO+ImageNet1K 的 ViT-small（M4）。为了提高预训练和部署的效率，我们在下文中将其作为默认模型。

①+② **预训练和架构对其他 Few-Shot Learners 有何影响**？ 我们的主要实验基于广泛使用的行业标准 ProtoNet

。我们接下来

¹ 在 miniImageNet 和 Meta-Dataset 中，元训练和元测试拆分中都使用了 ImageNet1K 的部分内容。例如：由于 Meta-Dataset 的 ImageNet 使用 712/288 源类/目标类分割，这意味着对于 Meta-Dataset 的 10 个领域之一，某些基础模型的预训练和元测试之间存在一些数据（但不是标签）重叠。正如第 2 章所讨论的，这种重叠在典型的自我监督评估管道中是普遍存在的 [15, 17]。在 FSL 评估管道中，这种情况并不常见，但它相当于在数据访问方面做出了半监督或转导假设，如文献 [38, 45, 49, 55] 所述。尽管如此，我们并不认为这是导致结果优异的重要因素，因为 CLIP 的 YFCC 没有这种重叠，其表现与基于 ImageNet1K 的模型类似。

列车配置		基准		
身份证		ArchPre 火车	MetaTr	miniIN
CIFAR				
	0ViT-small DINO (IN1K) -88		.8 97.0 59.1 79.8	
1	ViT-small DINO (IN1K) ProtoNet93		.1 98.0 81.1 92.5	
2	ResNet18	-MetaQDA	65.1 81.0 -	-
	3ViT-small DINO (IN1K) MetaQDA	92.0 97.0 77.2 90.1		
	4ResNet12	-MetaOptNet	64.1 80.0 72.8 85.0	
	5ViT-small DINO (IN1K) MetaOptNet	92.2 97.8 70.2 84.1		

表 2.架构和预训练对最先进的少量学习器的影响：MetaQDA [72], MetaOptNet [42].

我们将探索我们的管道如何影响两个更能代表最新技术水平的少量学习器，即 MetaOptNet [42] 和 MetaQDA [72]。从表 2 中的结果可以看出 (i) MetaQDA 和 MetaOptNet 在直接特征转移（M5 和 M3 对 M0）以及在最初评估时使用的较简单 ResNet 特征（M5 对 M4，M3 对 M2）方面确实有所改进。但是 (ii) 在使用更强的特征时，它们的表现要优于更简单的 ProtoNet 学习器（M3 和 M5 对 M1）。这表明，在使用更强特征的新系统中，可能需要重新评估以前关于元学习器性能比较的结论。

少量学习与自我监督学习 现有文献通常无法直接比较少量学习领域的算法（如 ProtoNet [59]、MAML [29]、MetaOptNet [42] 等）与自我监督领域的算法（如 DINO [15]、Sim-CLR [17, 18] 等）。部分原因是流行的评估协议不同：例如，FSL 社区流行 5 路-1-shot 机制，而 SSL 社区流行 1%标签（ImageNet 的情况下≈ 1000 路-10-shot）；网络架构不同（≤ResNet18 vs ≥ResNet50 再光谱）；图像分辨率不同（84× vs full）。我们的结果为这种直接比较提供了一个尝试。总的来说，这些结果表明，与标准少量学习器（仅使用元训练数据）相比，冷冻自监督基础模型（使用额外的预训练数据）具有开箱即用的竞争力。不过，更有趣的是，我们将这两种范式结合起来，很容易就能在典型的 FSL 指标上取得一流的性能。

预训练和元测试之间的类别重叠 虽然无监督预训练不使用标签，但预训练中使用的某些类别也很可能在元

测试中使用。这种类别重叠是否违背了"少量学习"的定义？从元学习的角度来看，答案是肯定的。但我们认为，除非对数据进行仔细的分割模拟，否则类重叠几乎是不可避免的。例如，就元数据集而言，CUB 数据集 [67]、飞机数据集 [50] 和 COCO 数据集 [47] 与 ImageNet [24, 32] 存在类重叠，但它们仍被用于元测试。当我们考虑更实际的大规模实验时，类重叠的问题就会出现。

M Arch	PreTr	元技术	MetaTe	平均值	输出-D
1 ViT-	小型 DINO	PN (IN)	PN	68.38	67.68
2 ViT-	小型 DINO	PN (IN)	PN+FT (lr=0.01)	76.05	76.54
3 ViT-	小型 DINO	PN (IN)	PN+FT (lr=0.001)	74.47	74.51
4 ViT-	小型 DINO	PN (IN)	PN+FT (已调试)	77.33	77.85
5	ViT-small DINO	PN (MD)	PN78	.43	55.71
6	ViT-small DINO	PN (MD)	PN+FT(lr=0.01)	76.09	73.26
7	ViT-small DINO	PN (MD)	PN+FT(lr=0.001)	74.64	69.97
8	ViT-small DINO	PN (MD)	PN+FT (调谐)	83.13	75.72

表 3.元数据集元测试期间的微调 (FT)。元训练 (MetaTr) 设置表示源数据集仅为 ImageNet (IN) 或全部元数据集 (MD)。结果是元数据集中所有域 (Avg) 的平均值, 以及仅分布外子集 (Out-D) 的平均值。

这一点无处不在。如果我们要对元学习算法进行基准测试, 我们就应该担心这个问题, 但就少量学习的本质而言, 对从极少数标签中快速构建分类器的能力进行基准测试并不会受到类重叠的阻碍。这也是自监督学习界完全不担心这个问题的原因。值得一提的是, 文献[46, 71]也提出了类似的 "少数几个标签学习" (few-shot few-shot learning) 方法, 即通过从不同领域仔细收集预训练数据或从互联网上抓取基础类别的预训练数据来避免重叠。或者, 也可以通过使用不同的模式来避免重叠。我们提倡元学习研究人员将这种受控环境视为测试平台, 以纳入强大的预训练特征骨干。

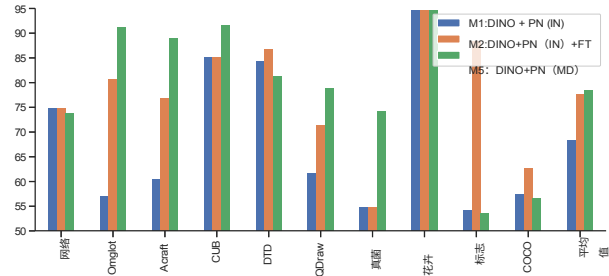


图 3.元测试期间微调对元数据集的影响。微调对 Signs 和 COCO 等数据集以及 omniglot 和 QuickDraw 等与 ImageNet 截然不同的数据集都有好处。

4.1.2 微调

之前的实验使用固定的特征提取器和 ProtoNet 进行元

方法 (骨干)	扩展 dat.	扩展 实验室	CIFAR-FS		迷你图像网	
			5w1s	5w5s	5w1s	5w5s
感应式						
ProtoNet (CNN-4-64) [59]			49.4	68.2	55.5	72.0
基线++ (CNN-4-64) [19]					48.2	66.4
MetaOpt-SVM (ResNet12) [42]			72.0	84.3	61.4	77.9
元基线 (ResNet12) [20]					68.6	83.7
RS-FSL (ResNet12) [2]					65.3	
传导式						
微调 (WRN-28-10) [23]			76.6	85.8	65.7	78.4
国际基础结构、水利和环境工程学院 (WRN-28-10) [36]			80.0	85.3	70.0	79.2
PT-MAP (WRN-28-10) [37]			87.7	90.7	82.9	88.8
CNAPS + FETI (ResNet18) [7]	✓	✓			79.9	91.5
自我监督						
ProtoNet (WRN-28-10) [30]			73.6	86.1	62.9	79.9
ProtoNet (AMDIM ResNet) [16]	✓				76.8	91.0
EPNet + SSL (WRN-28-10) [57]	✓				79.2	88.1
半监督						
LST (ResNet12) [45]	✓				70.1	78.7
PLCM (ResNet12) [38]	✓		77.6	86.1	70.1	83.7
$P>M>F$ (IN1K, RN50)	✓		73.7	84.0	79.2	92.0
$P>M>F$ (IN1K, ViT-Small)	✓		81.1	92.5	93.1	98.0
$P>M>F$ (IN1K, ViT-碱基)	✓		84.3	92.2	95.3	98.4

测试。接下来, 我们将研究如何在元测试过程中使用微调来进一步提高性能。基于 DINO 预训练 ViT 模型在第 4.1.1 节中的出色表现, 我们将重点放在这些模型上。

如何更好地利用微调进行元测试?

miniImageNet 和 CIFAR - 与代表性 SOTA FSL 算法的比较
。使用外部数据和/或标签的方法已标出。

为了回答这个问题，我们比较了之前探讨过的虚构特征转移与 ProtoNet，以及第 3.3 节中概述的对支持集（ProtoNet+FT）进行随集微调的 ProtoNet。我们使用 Meta-Dataset，包括将 ImageNet 单独作为源和对所有 Meta-Dataset 进行联合元训练这两种情况。根据图 3 和表 3 中的结果，我们可以得出以下结论：(i) 在全部元数据集上进行元训练比单独在 ImageNet 上进行元训练效果更好（M5 对 M1）。

(ii) 在元测试期间进行的微调大大改善了分布外数据集，尤其是在 ImageNet 上进行元训练，然后跨域部署到所有其他元数据集任务的情况：参见表 3 中的 Out-D 栏和 M2 与 M1 的对比；图 3 中 OmniGlott、QuickDraw、交通标志等的蓝色与橙色条。然而，在使用更多元数据集领域进行训练和测试的情况下，微调对各领域的影响并不一致：虽然微调对其余的 OOD 数据集有帮助，但总体上却没有帮助（M5 与 M6 对 Avg 和 Out-D）。通过微调更新的总体特征骨干对元训练期间未见过的领域更有帮助，这与文献[43, 65]一致。在分析微调的不一致影响时，我们发现这是由于难以选择合适的学习率造成的。像我们上面所做的（ $lr=0.01$ ）那样自始至终使用任何一个学习率，对于某些数据集来说都是不合适的。因此，我们还探索了第 3.3 节中提出的学习率选择启发式，结果发现它能带来最好的性能（M4 vs M2）。

4.2. 标准基准的结果

我们将我们的管道称为 $P>M>F$ ，它可以与任何预训练算法和骨干架构实例化、

8 个域内数据集	域内								域外		平均值	
	网络	Omglot	Acrafft	CUB	DTD	QDraw	真菌	花卉	标志	COCO		
ProtoNet [65] (RN18)	67.01	44.5	79.56	71.14	67.01	65.18	64.88	40.26	86.85	46.48	63.29	
CNAPs [56] (RN18+Adapter)	50.8	91.7	83.7	73.6	59.5	74.7	50.2	88.9	56.5	39.4	66.90	
SUR [26] (RN18+适配器)	57.2	93.2	90.1	82.3	73.5	81.9	67.9	88.4	67.4	51.3	75.32	
T-SCNAPs [7] (RN18+适配器)	58.8	93.9	84.1	76.8	69.0	78.6	48.8	91.6	76.1	48.7	72.64	
URT [48] (RN18+适配器)	55.7	94.4	85.8	76.3	71.8	82.5	63.5	88.2	69.4	52.2	73.98	
长笛 [64] (RN18)	51.8	93.2	87.2	79.2	68.8	79.5	58.1	91.6	58.4	50.0	71.78	
URL [44] (RN18+适配器)	57.51	94.51	88.59	80.54	76.17	81.94	68.75	92.11	63.34	54.03	75.75	
国际热带木材组织 [43] (RN18+适配器)	57.35	94.96	89.33	81.42	76.74	82.01	67.4	92.18	83.55	55.75	78.07	
P>M>F (DINO/IN1K, RN50)	67.51	85.91	80.3	81.67	87.08	72.84	60.03	94.69	87.17	58.92	77.61	
P>M>F (DINO/IN1K, ViT-small)	74.59	91.79	88.33	91.02	86.61	79.23	74.2	94.12	88.85	62.59	83.13	
P>M>F (DINO/IN1K, ViT-碱基)	77.02	91.76	89.73	92.94	86.94	80.2	78.28	95.79	89.86	64.97	84.75	
域内 = ImageNet	域内	域外										
	INet	Omglot	Acrafft	CUB	DTD	QDraw	真菌	花卉	标志	COCO	平均值	
ProtoNet [65] (RN18)	50.5	59.98	53.1	68.79	66.56	48.96	39.71	85.27	47.12	41	56.10	
alfa+fp-maml [5] (rn12)	52.8	61.87	63.43	69.75	70.78	59.17	41.49	85.96	60.78	48.11	61.41	
博博[58] (RN18)	51.92	67.57	54.12	70.69	68.34	50.33	41.38	87.34	51.8	48.03	59.15	
CTX [24] (RN34)	62.76	82.21	79.49	80.63	75.57	72.68	51.58	95.34	82.65	59.9	74.28	
P>M>F (DINO/IN1K, RN50)	67.08	75.33	75.39	72.08	86.42	66.79	50.53	94.14	86.54	58.2	73.25	
P>M>F (DINO/IN1K, ViT-small)	74.69	80.68	76.78	85.04	86.63	71.25	54.78	94.57	88.33	62.57	77.53	
P>M>F (DINO/IN1K, ViT-碱基)	76.69	80.68	76.78	84.38	86.63	71.25	55.93	95.14	89.68	65.01	79.09	
表 5. 元数据集 80 与 SOTA FSL 算法的比较。												
	胸部 X			ISIC			欧洲卫星			作物病害		
	5w5s	5w20s	5w50s	5w5s	5w20s	5w50s	5w5s	5w20s	5w50s	5w5s	5w20s	5w50s
ProtoNet [59] (RN10)	24.05	28.21	29.32	39.57	49.50	51.99	73.29	82.27	80.48	79.72	88.15	90.81
关系网[61] (RN10)	22.96	26.63	28.45	39.41	41.77	49.32	61.31	74.43	74.91	68.99	80.45	85.08
MetaOptNet [42] (RN10)	22.53	25.53	29.35	36.28	49.42	54.80	64.44	79.19	83.62	68.41	82.89	91.76
微调[33] (RN10)	25.97	31.32	35.49	48.11	59.31	66.48	79.08	87.64	90.89	89.25	95.51	97.68
厨师 [1] (rn10)	24.72	29.71	31.25	41.26	54.30	60.86	74.15	83.31	86.55	86.87	94.78	96.77
启动 [52] (rn10)	26.94	33.19	36.91	47.22	58.63	64.16	82.29	89.26	91.99	93.02	97.51	98.45
DeepCluster2 [14, 27] (IN1K, RN50)	26.51	31.51	34.17	40.73	49.91	53.65	88.39	92.02	93.07	93.63	96.63	97.04
P>M>F (DINO/IN1K, ResNet50)	27.13	31.57	34.17	43.78	54.06	57.86	89.18	93.08	96.06	95.06	97.25	97.77
P>M>F (DINO/IN1K, ViT-small)	27.27	35.33	41.39	50.12	65.78	73.50	85.98	91.32	95.40	92.96	98.12	99.24

表 6.跨域少量学习的更广泛研究--与 SOTA FSL 算法的比较。

例如，DINO > ProtoNet (PN) > Fine-tuning (FT)。接下来，我们将我们的管道与现有技术进行比较。**我们强调，在架构和外部数据的使用方面，我们的结果与之前的 SOTA 并无直接可比性。**我们进行这种比较是为了了解简单的变化（例如将特征骨干网升级为现代网络架构，以及利用公开数据进行大规模预训练）与 5 年来对 FSL 算法的深入研究相比有何不同。表 4 总结了单域案例（即 mini-ImageNet 和 CIFAR-FS）的结果，表 5 和表 6 分别显示了 跨域数据集（即 Meta-Dataset 和 Broader Study CDFSL）的结果。从结果中我们可以看出，我们的框架在域内和跨域条件下的表现都远远优于目前的技术水平，尽管它比一些复杂的竞争者要简单得多。我们注意到，对于表 4 中的单一来源基准，

一些竞争对手也使用了外部数据或 ImageNet 预训练。而我们的混合

表 6 显示，在 CDFSL 方面，流水线的性能优于 SOTA 纯外部自监督 [14, 27]。我们的代码见 https://github.com/hushell/pmf_cvpr22。

4.3. 讨论

总之，这些结果表明，我们利用现有预训练数据和现代架构的简单方法，往往能在少数几次学习中超越先进技术。在元测试阶段，利用我们提出的自适应微调机制，这一优势还将进一步扩大。基于这些观察结果，我们向从业人员和少次元学习研究人员提出了建议。

从业人员：增加预训练数据量或简单地使用基础模型 [10, 15] 并升级到现代架构可能比跟上和实施最先进的少量学习算法更有成效（也更容易实施）。如果感兴趣的几项目标任务与预训练和元训练数据的相似度较低，那么微调可能会很重要。

FSL 研究人员：我们的研究表明，使用外部数据和现代架构是实现强大的 FSL 性能的一种简单而有效的方法，而且一些 SOTA 元学习器在这种情况下也无法提供预期的改进。虽然外部数据违反了坚持使用特定有限元训练集的 FSL 问题定义，但面对不断进步的自我监督 [15,28,39,53]，我们应该认真对待这种设置，以保持实用性。特别是，我们建议对所有标准 FSL 基准进行新的评估设置，即自由选择预训练数据和架构，并明确报告。然后，针对给定的外部数据集和架构，对少量元学习方法进行评估，看其是否能改进线性读出、微调或我们的 PMF 基线。

5. 结论

我们从数据集、架构和微调策略的角度出发，推动了简单预训练 + ProtoNet 管道的极限。我们发现，源数据集和神经架构是影响 FSL 性能的主导因素。当训练和测试之间存在领域转换时，我们发现通过数据增强对特征骨干进行微调也很重要。我们验证了我们的简单管道在四个 FSL 基准中取得了极具竞争力的性能。

局限性和未来工作 我们的实证研究有几个局限性。我们只是初步了解了外部数据和相应的大型架构对 FSL 的影响。我们对外部数据的重新关注强调了将 FSL 社区的算法 [29, 42, 59] 与自我监督社区的算法 [10, 17] 进行直接比较的必要性，或者像我们在这里尝试的那样，将两者进行协同组合的必要性。我们提出的混合管道显然仅限于已经存在大型外部数据集的模式，如果没有预先训练的基础模型，则需要在计算和能源成本方面进行大量前期投资。基础模型内部可能存在的偏差也是一个潜在风险[10]。最后，我们的自适应微调策略虽然有效，但在元测试时的计算成本相当高，在没有反向传播的嵌入式平台上可能无法支持。前馈表示自适应方法 [56] 可能对未来的工作很重要。

我们感谢 CVPR2022 的匿名审稿人和元审稿人对我们稿件的仔细阅读和深入讨论。我们还要感谢上汽集团剑桥分公司的同事，特别是 Gabor Gyorkei、Taekwon Jang 和 Brais Martinez 的帮助和支持。

鸣谢

参考资料

- [1] Thomas Adler、Johannes Brandstetter、Michael Widrich、Andreas Mayr、David Kreil、Michael Kopp、Günter Klambauer 和 Sepp Hochreiter。通过表征融合进行跨域少量学习。载于 *arXiv*, 2021 年。8
- [2] Mohamed Afham、Salman Khan、Muhammad Haris Khan、Muzammal Naseer 和 Fahad Shahbaz Khan。丰富的语义改善了少量学习。*BMVC*, 2021.7
- [3] Wav2vec 2.0: 语音表征的自监督学习框架。*NeurIPS*, 2020.3
- [4] Sungyong Baik、Janghoon Choi、Heewon Kim、Dohee Cho、Jaesik Min 和 Kyoung Mu Lee。利用任务自适应损失函数的元学习 (Meta-learning with task-adaptive loss function for few-shot learning)。*ICCV*, 2021.1
- [5] Sungyong Baik、Myungsub Choi、Janghoon Choi、Heewon Kim 和 Kyoung Mu Lee。具有自适应超参数的元学习。*NeurIPS*, 2020.8
- [6] Hangbo Bao, Li Dong, and Furu Wei. Beit: 图像变换器的伯特预训练。2022 年, *ICLR*。4
- [7] Peyman Bateni、Jarred Barber、Jan-Willem van de Meent 和 Frank Wood。利用无标签示例增强少镜头图像分类。*WACV*, 2022.7, 8
- [8] Luca Bertinetto, João F. Henriques, Philip H.S. Torr, and Andrea Vedaldi. 可微分闭式求解器的元学习。在 *ICLR*, 2019 年。5
- [9] Luca Bertinetto、Joao F. Henriques、Jack Valmadre、Philip H.S. Torr 和 Andrea Vedaldi. 学习前馈单次学习器。In *NIPS*, 2016.1, 2
- [10] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *ArXiv preprint arXiv:2108.07258*, 2021.2, 3, 6, 8, 9
- [11] Myriam Bontou, Nicolas Farrugia 和 Vincent Gripon。解码大脑信号的少量学习。*CoRR*, abs/2010.12500, 2020。2
- [12] Stevo Bozinski. 关于神经网络迁移学习的第一篇论文的提醒, 1976. *Informatica*, 44 (3), 2020. 3
- [13] Tom B Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prfulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少数学习者。*NeurIPS*, 2020.3
- [14] 玛蒂尔德-卡隆、伊山-米斯拉、朱利安-梅拉尔、普里亚-戈亚尔、皮奥特-博扬诺夫斯基和阿曼德-朱林。通过对比聚类分配实现视觉特征的无监督学习。*NeurIPS*, 2020.8
- [15] 玛蒂尔德-卡隆 (Mathilde Caron)、雨果-图夫隆 (Hugo Touvron)、伊山-米斯拉 (Ishan Misra)、埃尔韦-热古 (Hervé Jégou)、朱利安-梅拉尔 (Julien Mairal)、皮奥特-博扬诺夫斯基 (Piotr Bojanowski) 和阿曼德-朱林 (Armand Joulin)。自监督视觉转换器的新兴特性。*ICCV*, 2021.3, 4, 6, 8, 9
- [16] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. 用于少镜头图像分类的自监督学习。*ICASSP*, 2021.7

- [17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 视觉表征对比学习的简单框架。In *ICML*, 2020.3, 6, 9
- [18] Ting Chen、Simon Kornblith、Kevin Swersky、Mohammad Norouzi 和 Geoffrey Hinton。Big self-supervised models are strong semi-supervised learners.*NeurIPS*, 2020.1, 3, 6
- [19] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 几发分类的近距离观察。*ICLR*, 2019.1, 2, 7
- [20] 陈银波、刘壮、徐慧娟、特雷弗·达雷尔、王喜龙。元基线：探索用于少量学习的简单元学习。*ICCV*, 2021.1, 3, 7
- [21] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet：大规模分层图像数据库。2009 年，*CVPR*。2, 3
- [22] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向转换器预训练。In *ACL*, 2019.3, 4
- [23] Guneet Singh Dhillon、Pratik Chaudhari、Avinash Ravichandran 和 Stefano Soatto。少镜头图像分类基线。在 *ICLR*, 2020。1, 3, 7
- [24] 卡尔·多尔施、安库什·古普塔和安德鲁·齐瑟曼。Crosstransformers: spatially-aware few-shot transfer.*NeurIPS*, 2021.3, 5, 6, 8
- [25] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一幅图像胜过 16x16 个单词：大规模图像识别变换器。In *ICLR*, 2021.1, 2, 4
- [26] Nikita Dvornik、Cordelia Schmid 和 Julien Mairal。从多域表示中选择相关特征，用于少镜头分类。在 *ECCV*, 2020。8
- [27] Linus Ericsson、Henry Gouk 和 Timothy M Hospedales。自我监督模型的转移效果如何？In *CVPR*, 2021.3, 6, 8
- [28] Linus Ericsson、Henry Gouk、Chen Change Loy 和 Timothy M Hospedales。自我监督表征学习：简介、进展与挑战》。*IEEE Signal Processing Magazine*, 2022.3, 9
- [29] Chelsea Finn、Pieter Abbeel 和 Sergey Levine。用于深度网络快速适应的模型无关元学习。*ICML*, 2017.1, 2, 3, 6, 9
- [30] Spyros Gidaris、Andrei Bursuc、Nikos Komodakis、Patrick Pérez 和 Matthieu Cord。利用自我监督提升少镜头视觉学习。*ICCV*, 2019.3, 7
- [31] Priya Goyal、Dhruv Mahajan、Abhinav Gupta 和 Ishan Misra。自监督视觉再现学习的规模化和基准化。In *ICCV*, 2019.2, 3
- [32] Pei Guo. imagenet 和 cub 之间的重叠。6
- [33] 郭云辉 Noel C Codella Leonid Karlinsky James V Codella John R Smith Kate Saenko Tajana Rosing and Rogerio Feris。跨域少量学习的广泛研究。*ECCV*, 2020.3, 4, 5, 8
- [34] 何开明、张翔宇、任少清和孙健。图像识别的深度残差学习。In *CVPR*, 2016.4

- [35] Timothy Hospedales、Antreas Antoniou、Paul Micaelli 和 Amos Storkey. 神经网络中的元学习：概览。《电气和电子工程师学会《模式分析与机器学习》期刊》，2021 年。[2](#)
- [36] Shell Xu Hu, Pablo Moreno, Yang Xiao, Xi Shen, Guillaume Obozinski, Neil Lawrence, and Andreas Damianou. 具有合成梯度的经验贝叶斯传导元学习。《国际法律文献中心》，2020 年。[7](#)
- [37] 胡宇清、Vincent Gripon 和 Stéphane Pateux. 在基于转移的少量学习中利用特征分布。《ICANN》，2021。[7](#)
- [38] Kai Huang, Jie Geng, Wen Jiang, Xinyang Deng, and Zhe Xu. 半监督少点学习的伪损失置信度。In *ICCV*, 2021。[6](#), [7](#)
- [39] L.Jing 和 Y. Tian. 深度神经网络的自监督视觉特征学习：A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.[3](#), [9](#)
- [40] Arman Kazemi、Shubham Sahay、Ayush Saxena、Moham- mad Mehdi Sharifi、Michael Niemier 和 X. Sharon Hu. 基于闪存的多比特内容可寻址存储器的电子平方距离。《IEEE/ACM 低功耗电子与设计国际研讨会》，2021 年。[2](#)
- [41] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 通过概率性程序归纳的人类级概念学习。《科学》，2015 年。[1](#)
- [42] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. 可微凸优化元学习。In *CVPR*, 2019.[1](#), [2](#), [6](#), [7](#), [8](#), [9](#)
- [43] Wei-Hong Li, Xialei Liu, and Hakan Bilen. 提高跨域少量学习的任务适应性》，*arXiv preprint arXiv:2107.00358*, 2021.[4](#), [7](#), [8](#)
- [44] Wei-Hong Li, Xialei Liu, and Hakan Bilen. 从多个领域学习通用表述，实现少量分类。《ICCV》，2021。[8](#)
- [45] 李新哲、孙倩茹、刘瑶瑶、周琴、郑世宝、蔡达生和 Bernt Schiele. 学习自我训练，实现半监督式少拍分类。《NeurIPS》，2019。[6](#), [7](#)
- [46] Yann Lifchitz, Yannis Avrithis, and Sylvaine Picard. 少数几个镜头的学习和空间注意力的作用。《2020 年第25届国际模式识别大会 (ICPR)》，第2693-2700页。IEEE, 2021。[7](#)
- [47] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 微软 coco：上下文中的常见对象。《欧洲计算机视觉会议》，第 740-755 页。Springer, 2014。[6](#)
- [48] Lu Liu、William Hamilton、Guodong Long、Jing Jiang 和 Hugo Larochelle. 用于少量图像分类的通用表示变换层2021 年，《ICLR》。[8](#)
- [49] Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 学习传播：用于少量学习的传导传播网络。载于 *ICLR*，2019 年。[6](#)
- [50] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 细粒度飞机视觉分类。《ArXiv 预印本 arXiv:1306.5151》，2013。[6](#)

- [51] Puneet Mangla、Nupur Kumari、Abhishek Sinha、Mayank Singh、Balaji Krishnamurthy 和 Vineeth N Balasubramanian。绘制正确的歧管图：少量学习的流形混合。*WACV*, 2020.1, 2
- [52] Cheng Perng Phoo 和 Bharath Hariharan。跨越极端任务差异的少发转移自我训练。2021 年, *ICLR*。8
- [53] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Aspell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可转移的视觉模型。In *ICML*, 2021.1, 4, 9
- [54] Sachin Ravi 和 Hugo Larochelle。将优化作为少数几次学习的模型。*ICLR*, 2017.1, 2
- [55] Mengye Ren、Eleni Triantafillou、Jake Snell Sachin Ravi、Kevin Swersky、Joshua B. Tenenbaum、Hugo Larochelle 和 Richard S. Zemel。用于半监督少数镜头分类的元学习。In *ICLR*, 2018.1, 6
- [56] James Requeima、Jonathan Gordon、John Bronskill、Sebastian Nowozin 和 Richard E. Turner。使用条件神经自适应过程的快速灵活多任务分类。*NeurIPS*, 2020.8, 9
- [57] 保罗-罗德里格斯、伊萨姆-拉拉吉、亚历山大-德鲁安和亚历山大-德雷-拉科斯特嵌入传播：用于少镜头分类的更平滑流形。*ECCV*, 2020.7
- [58] Tonmoy Saikia、Thomas Brox 和 Cordelia Schmid。用于跨领域少量分类的开放式通用特征学习。载于 *arXiv*, 2020 年。8
- [59] Jake Snell、Kevin Swersky 和 Richard Zemel。用于少量学习的原型网络。In *NIPS*, 2017.1, 2, 3, 4, 6, 7, 8, 9
- [60] Chen Sun、Abhinav Shrivastava、Saurabh Singh 和 Abhinav Gupta。重新审视深度学习时代数据的不合理有效性。In *ICCV*, 2017.2, 3
- [61] 宋洪涛、杨永新、张莉、向涛、菲利普-托尔和蒂莫西-M-霍斯佩德莱斯。学会比较：用于少量学习的关系网络In *CVPR*, 2018.1, 8
- [62] 巴特-托米 (Bart Thomee)、大卫-A-沙马 (David A. Shamma)、杰拉尔德-弗里德兰 (Gerald Friedland)、本杰明-埃利萨尔德 (Benjamin Elizalde)、卡尔-倪 (Karl Ni)、道格拉斯-波兰 (Douglas Poland)、达米安-博思 (Damian Borth) 和李力嘉。Yfcc100m：多媒体研究的新数据。*Commun.ACM*, 59 (2) : 64-73, 2016 年 1 月。2
- [63] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola。反思少量图像分类：一个好的嵌入就够了吗？*ECCV*, 2020.1, 2
- [64] Eleni Triantafillou、Hugo Larochelle、Richard Zemel 和 Vincent Dumoulin。为少量数据集泛化学习通用模板。*ICML*, 2021.8
- [65] Eleni Triantafillou、Tyler Zhu、Vincent Dumoulin、Pascal Lamblin、Utku Evci、Kelvin Xu、Ross Goroshin、Carles Gelada、Kevin Swersky、Pierre-Antoine Manzagol 和 Hugo Larochelle。元数据集：从少量示例中学习的数据集。In *ICLR*, 2020.2, 3, 5, 7, 8
- [66] Oriol Vinyals、Charles Blundell、Timothy Lillicrap、Koray Kavukcuoglu 和 Daan Wierstra。单次学习的匹配网络。*NeurIPS*, 2016.1, 3, 5

- [67] Catherine Wah、Steve Branson、Peter Welinder、Pietro Perona 和 Serge Belongie。Caltech-ucsd birds-200-2011 数据集。
技术。技术报告, 2011 年。[6](#)
- [68] Yan Wang、Wei-Lun Chao、Kilian Q. Weinberger 和 Laurens van der Maaten。Simpleshot: 重新审视近邻分类的少量学习, 2019 年。[1](#), [2](#)
- [69] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni.从少量实例中归纳: 关于少数几个学习的调查。*ACM Computing Surveys (CSUR)*, 53(3):1-34, 2020.[1](#), [2](#)
- [70] Jason Yosinski、Jeff Clune、Yoshua Bengio 和 Hod Lipson。深度神经网络中的特征有多大转移性? In *NIPS*, 2014.[2](#), [3](#)
- [71] Jianhong Zhang, Manli Zhang, Zhiwu Lu, and Tao Xiang.Adargcn: 用于少量学习的自适应聚合GCN。《*计算机视觉应用冬季会议论文集*》, 第 3482-3491 页, 2021 年。[7](#)
- [72] X.Zhang, D. Meng, H. Gouk, and T. Hospedales.浅贝叶斯元学习用于真实世界的少镜头识别。In *ICCV*, 2021.[1](#), [2](#), [6](#)

突破简单管道的极限，实现快速学习：补充材料

在这份补充材料中，我们介绍了

- 在[第 1 节](#)中，我们为正文中的表 1 补充了一些结果。
- 在[第 2 节](#)中，我们将对本文表 1 和表 4 的结果进行补充。
- 在[第 3 节](#)中，我们将研究超参数对微调阶段的影响。
- 在[第 4 节](#)中，我们展示了 ProtoNet 元训练前后的 T-SNE 图。

1. 元数据集的其他结果

在本节中，我们将全面展示正文表 1 中的结果，包括不同预训练方法的结果（见表 [1](#)）、在 ImageNet 领域进行元训练的结果（见表 [2](#)）以及在八个预先指定的领域进行元训练的结果（见表 [3](#)）。

如本文正文所示，我们的管道以 "P > M > F (骨干)" 的形式命名，其中 "P"、"M" 和 "F" 分别取自预训练、元训练和微调的首字母。在本节中，我们只考察预训练和骨干架构部分，元训练固定为 ProtoNet。例如，在表 [2](#) 中，我们使用 "DINO > PN (ViT-small)" 来表示使用 DINO 预训练、ProtoNet 元训练和主干架构为 ViT-small 的管道。

为了澄清表 [1](#)、表 [2](#) 和表 [3](#) 中的简称，我们在此列出一个清单：

- DINO：由 [\[2\]](#) 在 ImageNet-1k 数据集上进行预训练。
- BEiT：由 [\[1\]](#) 在 ImageNet-21k 数据集上进行 BERT 预训练。
- CLIP：由 [\[3\]](#) 在 YFCC100M 数据集上进行语言-图像对比预训练。
- Sup21k：在 ImageNet-21k 数据集上进行有监督的预训练。
- Sup1k：在 ImageNet-1k 数据集上进行有监督的预训练。
- BEiT + Sup21k：BERT 先在 ImageNet-21k 数据集上进行无监督预训练，然后使用 ImageNet-21k 的标签对模型进行微调。

2. miniImageNet 和 CIFAR-FS 的其他结果

我们还在 miniImageNet 和 CIFAR-FS 上评估了不同的预训练方法和骨干，如表 [4](#) 所示。由于在 ImageNet 上进行有监督的预训练仅对检查性能上限有用，因此我们没有将部分结果写入正文。

3. 关于微调超参数的消融研究

微调阶段有三个超参数：学习率、梯度下降步数和切换支持集数据增强的概率。图 1 显示，学习率是最主要的超参数。从结果中我们还可以看出，切换到数据增强的概率越高越好，而 50 个梯度步数在学习率合适的情况下性能相对较好。因此，在微调阶段，我们将概率固定为 0.9，步数为 50 步。

	INet	Omglot	Acrafft	CUB	DTD	QDraw	真菌	花卉	标志	COCO	平均值
迪诺 (ViT-小号)	73.48	54.33	62.17	85.37	83.67	60.59	56.26	94.45	53.7	54.58	67.86
迪诺 (ViT-base)	74.85	59.44	55.36	80.08	84	59.61	56.65	94.84	51.81	57.1	67.374
BEiT (ViT-base)	17.12	23.96	17.21	18.59	39.79	23.89	13.69	45.81	16.16	16.36	23.258
CLIP (ViT-base)	60.66	62.12	54.08	80.26	76.51	62.90	30.76	68.43	47.33	41.95	58.5
DINO (ResNet50)	64.13	52.51	57.02	62.63	84.5	60.78	50.41	92.18	58.27	55.43	63.786
CLIP (ResNet50)	51.67	44.16	44.18	70.2	70.64	47.88	34.13	87.97	39.59	41.63	53.205
Sup21k (ViT 基础)	67.00	37.02	47.72	82.9	79.77	52.25	41.98	95.7	46.22	53.46	60.402
BEiT + Sup21k (ViT 基础)	33.85	23.95	33.92	52.07	63.79	32.60	28.19	67.3	27.18	29.65	39.25
Sup1k (ViT 基础)	89.1	60.71	55.36	79.8	79.75	61.28	47.45	88.44	56.3	57.20	67.539
Sup1k (ResNet50)	76.22	47.31	55.75	76.40	80.40	51.26	43.42	85.48	50.46	57.10	62.38

表 1.元数据集的预训练结果--不同预训练方法和主干架构的比较。

	域内	域外									平均值
	网络	Omglot	Acrafft	CUB	DTD	QDraw	真菌	花卉	标志	COCO	
DINO > PN (ViT-小型)	74.69	56.91	60.5	85.04	84.21	61.54	54.78	94.57	54.21	57.35	68.38
DINO > PN (ViT 基底)	76.69	62.2	54.76	81.58	84.48	60.64	55.93	95.14	56.81	60.27	68.85
CLIP > PN (ViT-base)	76.03	59	65.75	90.2	83.08	65.45	53.2	96.35	58.65	61.2	70.891
DINO > PN (ResNet50)	67.08	49.21	58.46	72.08	85.01	59.2	50.53	89.91	55.44	53.94	64.086
CLIP > PN (ResNet50)	69.41	60.72	57.53	83.66	80.03	55.58	50.07	93.39	48.56	50.14	64.909
Sup21k > PN (ViT 基础)	85.88	39.72	52.03	94.54	83.42	54.58	57.06	99.01	47.74	69.02	68.3
BEiT+Sup21k > PN (ViT 基础)	84.39	60.54	74.04	95.66	86.14	65.24	64.25	99.19	63.02	69.91	76.238
Sup1k > PN (ViT 基础)	90.48	62.96	54.89	78.88	80.02	61.81	45.52	88.56	55.61	59.12	67.785

表 2.元数据集（仅 ImageNet）上的元训练结果--不同预训练方法和骨干架构的比较。

	域内								域外		平均值
	网络	Omglot	Acrafft	CUB	DTD	QDraw	真菌	花卉	标志	COCO	
DINO > PN (ViT-小型)	73.54	91.79	88.33	91.02	81.64	79.23	74.2	94.12	54.37	57.04	78.528
DINO > PN (ViT 基底)	73.55	91.54	89.73	92.94	81.52	80.2	78.28	94.53	53.65	59.13	79.507
CLIP > PN (ViT-base)	74.76	92.26	91.42	93.55	80.97	80.8	79.13	95.64	54.52	56.8	79.985
DINO > PN (ResNet50)	63.7	85.91	80.3	81.67	82.69	72.84	60.03	91.75	54.26	50.67	72.382
CLIP > PN (ResNet50)	64.86	92.09	89.19	89.17	71.67	78.71	76.15	91.25	51.1	45.88	75.007
Sup21k > PN (ViT 基础)	84.86	85.71	83.77	95.89	85.1	78.47	74	99.17	59.86	67.57	81.44
BEiT+Sup21k > PN (ViT 基础)	81.96	94.19	91.62	93.76	81.3	83.48	81.76	98.84	58.83	61.81	82.755
Sup1k > PN (ViT-小型)	83.87	91.22	87.9	89.2	78.11	78.7	70.33	94	56.24	57.16	78.673
Sup1k > PN (ViT 基础)	89.75	93.48	91.15	92.48	78.52	80.65	75.97	95.78	53.47	55.89	80.714
Sup1k > PN (ResNet50)	68.04	86.17	80.72	80.48	71.65	70.78	59.58	84.33	50.06	50.29	70.21
无 > PN (ViT-小型)	37.25	74.14	45.25	49.66	61.49	70.24	43.23	72.03	39.33	35.43	52.805
无 > PN (ResNet50)	40.74	90.67	80.67	68.88	62.4	75.96	55.72	75.37	43.11	35.49	62.901

表 3.元数据集上的元训练结果--不同预训练方法和主干架构的比较。

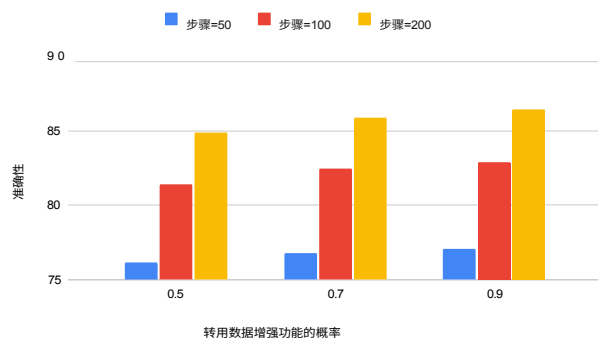
4. T-SNE 图：元训练前后

通过使用 T-SNE 可视化技术，我们发现 DINO 预训练的特征表示在多个领域都已达到很高的质量。图 2、图 3 和图 4 显示了三个例子。一般来说，尽管这些聚类所在的领域并不一定与 ImageNet 相似，但已经出现了许多语义聚类。这为 ProtoNet 提供了一个非常好的初始化，使其能够完善聚类，使其更加紧密。如果我们从头开始训练 ProtoNet，情况就会完全不同，表 3 中的无预训练结果证实了这一点。这可以从 K-means 聚类的角度来解释，K-means 聚类总是需要一个良好的初始化。

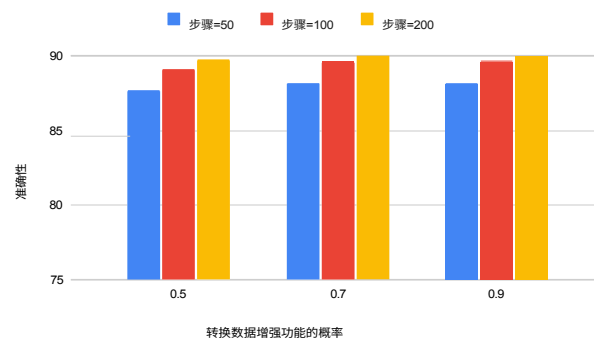
	迷你图像网		CIFAR-FS	
	5w1s	5w5s	5w1s	5w5s
DINO > PN (ViT-小型)	93.1	98.0	81.1	92.5
DINO > PN (ViT 基底)	95.3	98.4	84.3	92.2
CLIP > PN (ViT-base)	93.1	98.1	85.3	93.2
DINO > PN (ResNet50)	79.2	92.0	73.7	84.0
CLIP > PN (ResNet50)	78.9	92.2	71.4	82.6
Sup21k > PN (ViT 基础)	97.2	99.2	92.3	96.7
BEiT+Sup21k > PN (ViT 基础)	96.6	99	93.8	97.5
Sup1k > PN (ViT-小型)	97.7	99.4	86.2	93.6
Sup1k > PN (ViT 基础)	99.2	99.8	88.2	94.3
Sup1k > PN (ResNet50)	91.7	97.4	77	87.6
无 > PN (ViT-小型)	36.5	49.1	45.9	59.8
无 > PN (ResNet50)	46.1	60.3	54.1	68.4

表 4. miniImageNet 和 CIFAR-FS - 不同预训练方法和骨干架构的比较。

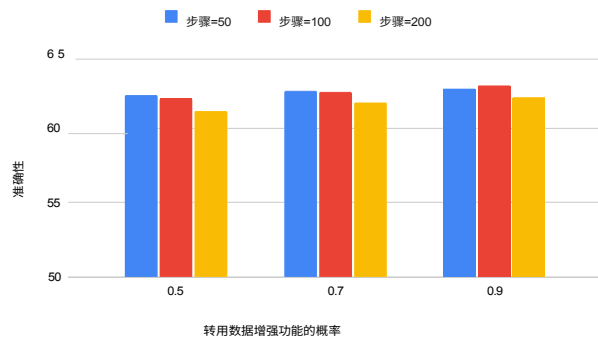
交通标志, lr = 0.001



交通标志, lr = 0.01



MSCOCO, lr = 0.001



MSCOCO, lr = 0.01

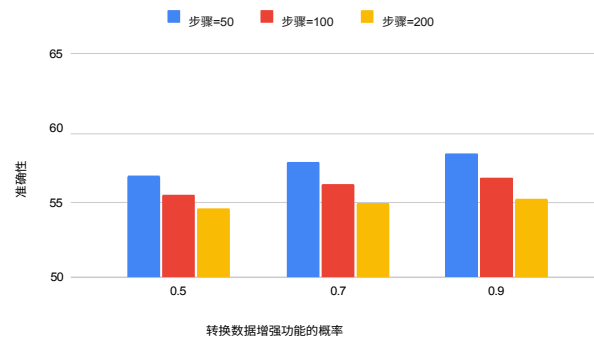


图 1 微调超参数的消融研究 - 实验是在交通标志域和 MSCOCO 域的验证集上进行的, 学习率固定为 0.001 或 0.01。

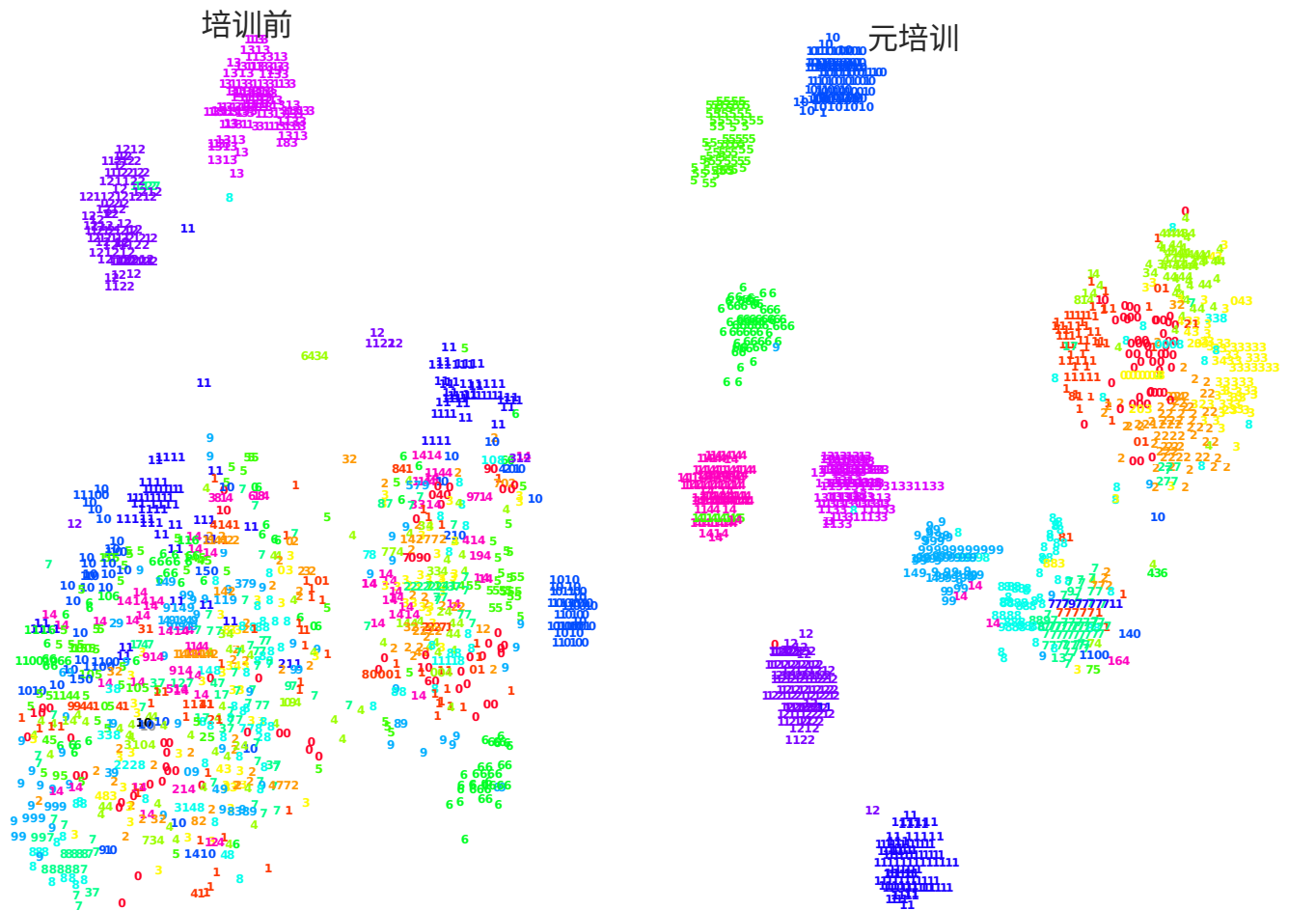
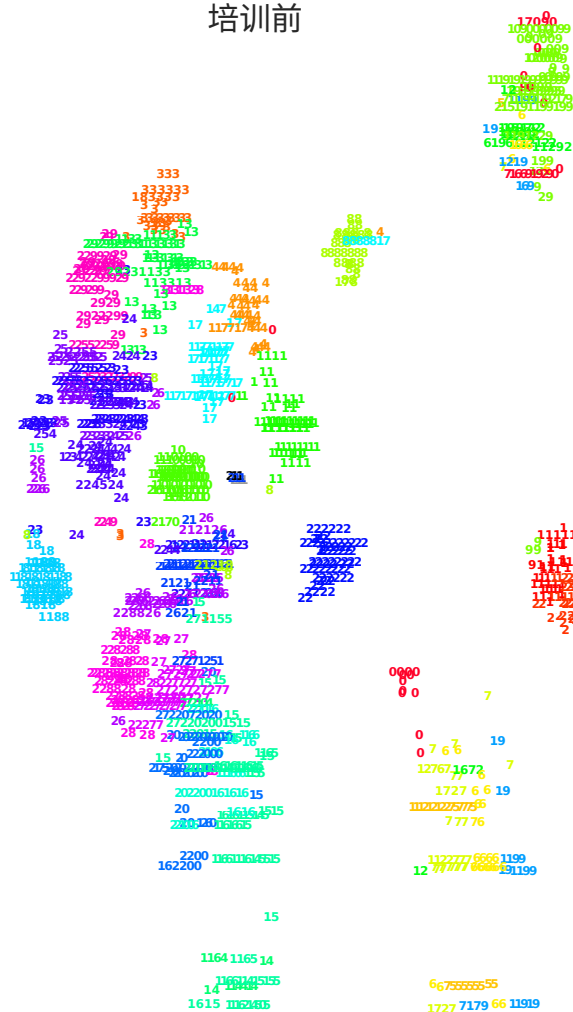
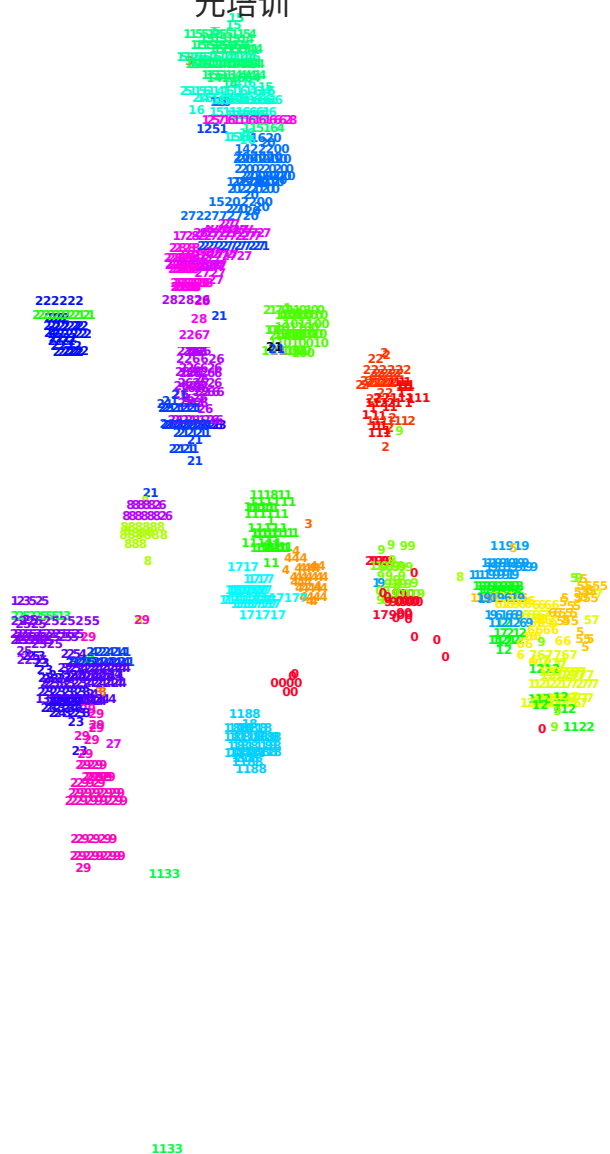


图 2.飞机领域

培训前



元培训



1010226 6 19

231311313318

6

15 15115515

12 7 75 5 66 19

10001000

1151611561514

1112221725

1411541151416
161415 14

11122 5555 61919
7 6 555 5 6

131110198133

13 13

16 116040116
16 1514151416

12 5555 59 6
6 9 5555

1131333
23113313
33313
33

16 1114144
16 22222
17 14

333 3
5 6 55
6 9 6
6 8 9
9

3 333
3333333
3

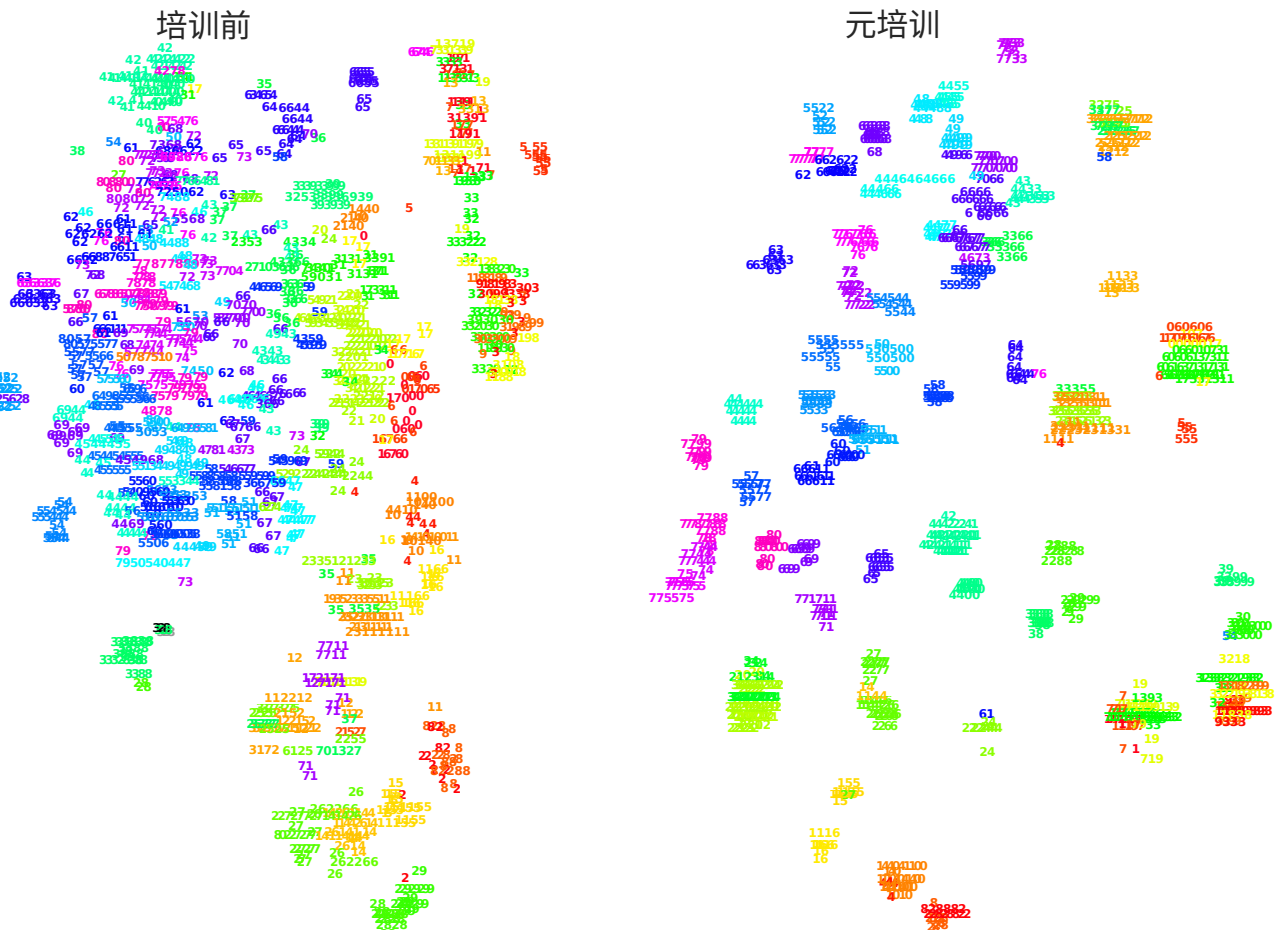


图 4.全图域

参考资料

- [1] Hangbo Bao、Li Dong 和 Furu Wei。Beit：图像变换器的伯特预训练。2022 年，*ICLR*。[1](#)
- [2] 玛蒂尔德-卡隆、雨果-图夫隆、伊善-米斯拉、埃尔韦-热古、朱利安-梅拉尔、皮奥特-博亚诺夫斯基和阿曼德-朱林。自监督视觉转换器的新特性。*ICCV*, 2021。[1](#)
- [3] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可转移的视觉模型。*ICML*, 2021。[1](#)