PET-TURTLE：面向不平衡数据簇的深度无监督支持向量机

Javier Salazar Cavazos, *IEEE* 研究生会员

arXiv:2601.03237v1 [cs.LG] 6 Jan 2026

摘要——基础视觉、音频和语言模型通过其潜在表示，在下游任务中实现零样本性能。近年来，利用深度学习方法无监督地学习数据分组结构日益受到关注。作为一种先进的深度聚类算法，TURTLE 通过交替进行数据标注和超平面更新来揭示数据标签，无需监督，并以类似支持向量机（SVMs）的方式最大化超平面间隔。然而，TURTLE 假设各簇均衡；当数据不平衡时，它会产生不理想的超平面，导致更高的聚类误差。我们提出 PET-TURTLE，通过幂律先验推广代价函数，以处理不平衡数据分布。此外，在标注过程中引入稀疏逻辑值后，PET-TURTLE 能够优化更简单的搜索空间，进而提高均衡数据集上的准确率。在合成数据和真实数据上的实验表明，PET-TURTLE 提高了不平衡数据源上的准确率，避免了对少数类簇的过度预测，并提升了整体聚类效果。

索引词—聚类、不平衡数据、基础模型、支持向量机（SVMs）、无监督学习。

I. 引言

迁移学习利用预训练神经网络，在数据通常有限的情况下提升模型在下游任务中的性能 [1]。近期研究表明，与使用冻结骨干网络搭配线性分类器相比，对整个模型进行微调带来的提升微乎其微[2] [3] [4]。这些自监督基础模型采用 CLIP 或对比学习范式进行训练 [5] [6]，，通过学习通用表示并在零样本场景中加以运用，在多种下游任务中展现出有竞争力的性能。然而，在某些场景中，数据标签未知，因此无法在这些模型之上训练线性分类器。因此，人们自然会想到在潜在表示上应用 K-Means 等聚类方法 [7] 或子空间变体 [8] 。已有研究表明，与弱监督方法相比，这种策略的性能较差 [9]。近期提出的 TURTLE 算法 [10] 旨在克服这些挑战：通过寻找能够最大化超平面间隔的标签划分，以无监督的方式发现数据组。与 DEC [11]，、DAC [12]，、DeepCluster [13]，和 SPICE [14]，等其他深度聚类方法相比，TURTLE 不采用特定任务表示学习，而这类学习对于现代基础模型而言通常代价高昂。与这些方法相比，TURTLE 在无监督聚类学习方面创下了新的最先进性能。

本研究获得美国国立卫生研究院（NIH）R21 AG082204 项目资助，以及密歇根大学拉克姆优秀奖学金的支持。

Javier Salazar Cavazos 任职于密歇根大学电气与计算机工程系（ECE），地址为美国密歇根州安娜堡，美国密歇根州48109。

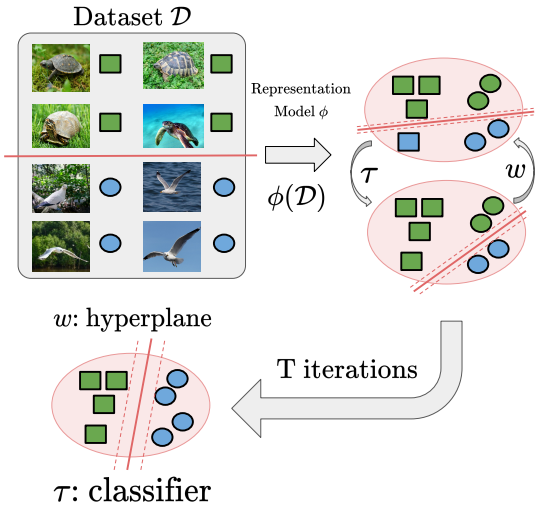

图 1. 无监督支持向量机核心思想的示意图：在标签更新与超平面估计之间交替进行。

进一步深入探讨问题陈述，令 ϕ 表示某个骨干模型，使得 $\phi(x) = z \in \mathbb{R}^d$ 表示属于 d 维空间且与数据集 $\mathcal{D}$ 中的原始数据样本 x 相关联的潜在特征。令 C 表示簇的数量。令 $\tau_\theta(z)$：$\{\mathcal{Z} = \phi(\mathcal{D})\} \to \{1, \ldots, C\}$表示具有连续参数 θ 的分类器，用于揭示 $z \in \mathcal{Z}$的底层标签。具体而言，令 $\tau_\theta(z) = A_\theta z + b_\theta$，其中 $A_\theta \in \mathbb{R}^{C\times d}$ 且 $b_\theta \in \mathbb{R}^C$。令 $w_\theta \in \mathbb{R}^{C\times d}$表示潜在空间中的超平面，其中还包含一个偏置项。于是，给定单个表征空间，TURTLE 的优化目标为

$$\mathcal{L}_{\text{TURTLE}}(\theta) = \sum_{z\in\phi(\mathcal{D})} \mathcal{L}_{\text{CE}}(w_\theta^M z; \sigma(\tau_\theta(z)))$$
$$\text{s.t. } w_\theta^M = \Xi^{(M)}(w_\theta^M, \phi(\mathcal{D})) \qquad (1)$$

其中 $\sigma(\cdot)$ 表示 softmax 运算，而 $\mathcal{L}_{\text{CE}}(\cdot)$ 是交叉熵损失 [15]。内层项 $\Xi^{(M)}(w_\theta^M, \phi(\mathcal{D}))$表示一个迭代优化算法 Ξ，例如梯度下降，运行 M 步，起始点为随机初始化的w_θ^0。在线性可分数据的二分类场景中，(1) 对应无正则化逻辑回归。在文献 [16],中，作者表明，将梯度下降应用于带有已知标签的 (1) 时，会使迭代过程偏向最大硬间隔超平面的方向。后续研究表明，在不可分场景下，软间隔超平面也存在类似结果 [17]。TURTLE 利用双层优化问题，通过交替寻找超平面并用其更新标签，这与 K-Means 类似

的方式，但应用于支持向量机的场景中。

然而，分类器需要进行正则化，τ 以防止出现全局且平凡的解，即编码器将所有样本都分类到同一个簇中，从而在技术上使（1）达到最小值。在此 $C=1$ 情形下，由于对应于较低代价函数值的单一簇数据距离超平面可能很远，因此很容易找到最大间隔，如参考文献 [18]所示。令

$$\bar{\tau}_\theta = \frac{1}{|\mathcal{D}|}\sum_{z\in\phi(\mathcal{D})}\tau_\theta(z) \tag{2}$$

为分类器对 $\mathcal{D}$ 预测得到的经验标签分布。于是，TURTLE公式的最终目标函数为

$$\min_\theta \mathcal{L}_{\text{TURTLE}}(\theta) - \gamma\mathbb{H}(\bar{\tau}_\theta) \tag{3}$$

其中，$\mathbb{H}(\cdot)$对应于离散分布的熵函数。令 $h_i \in \bar{\tau}_\theta$ 表示向量 $\bar{\tau}_\theta$的第 i个元素。于是，$\mathbb{H}(\cdot)$逐元素执行以下运算：

$$\mathbb{H}(h_i) = \begin{cases} -h_i \ln h_i & h_i > 0 \\ 0 & h_i = 0 \\ -\infty & h_i < 0 \end{cases} \tag{4}$$

这可以避免退化解 $C=1$。然而，这也会促使簇趋于均衡，因为所有类别受到的惩罚相同。因此，TURTLE公式中明确包含均衡簇假设，这会导致其在不平衡数据情形下的聚类质量较差。这是自监督学习中普遍观察到的现象[19]。例如，其他人解决这一问题的方法包括：通过集成方法融合不同专家的信息 [20]，或随机编码子任务，以在训练过程中找到自适应加权 [21]。

在本研究中，我们通过推广（3），来应对均衡簇假设，从而更好地处理不平衡数据集带来的分布，提升准确率并增强簇内聚性。我们提出的新方法 PET-TURTLE 会根据数据不平衡的严重程度，在不平衡情形下取得更高性能。此外，由于分类器采用稀疏logit建模，我们的方法甚至能在平衡数据场景下降低聚类误差 τ：它将超平面更新时需要考虑的潜在类别数量限制为概率较高的类别。我们推测，这使超平面搜索更加高效，不会受到低概率预测的拖累。这与子空间学习问题类似：在主成分分析（PCA）中，噪声极大的数据样本可能导致子空间基估计效果变差 [22]。在第II节中，我们将详细介绍所提出的PET-TURTLE算法，首先从

簇不平衡问题。

II. 所提出的方法

A. 不平衡数据的先验约束项

为解决簇不平衡带来的局限，我们考虑以下一类优化问题

$$\min_\theta \mathcal{L}_{\text{TURTLE}}(\theta) + \gamma D_{\text{KL}}[\bar{\tau}_\theta \parallel \Pi] \tag{5}$$

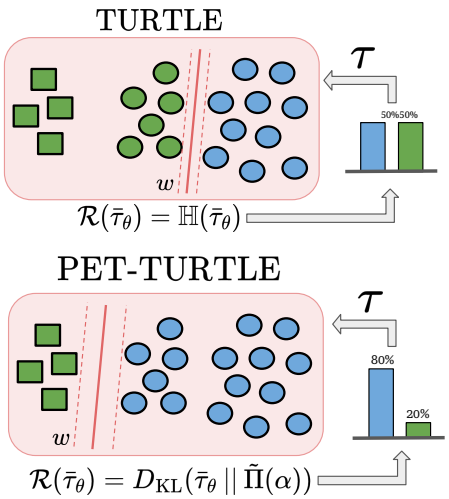

图2. TURTLE和PET-TURTLE目标函数中正则化项效果的直观比较。

其中，$D_{\text{KL}}(\cdot)$对应于两个分布 $\bar{\tau}_\theta$ 与先验数据分布 Π之间的KL散度 [23]。在平衡聚类情形下，Π 可以假设为均匀分布。在不平衡情形下，对于某些应用，Π 可能是已知的，例如场景图生成问题，其中在分类场景中，交叉熵损失中加入了KL先验 [24]。在许多情况下，分布更有可能是未知的，但可以假设其是不平衡的。

我们提出使用幂律分布 [25] 作为代理，在未知情况下对不平衡数据进行建模。由于其在物理学、生物学、医学、化学、天文学和经济学等领域有广泛的应用，因此幂律分布是描述不平衡分布的自然选择 [25]。给定衰减因子 $\alpha \in \mathbb{R}_{\geq 0}$，幂律概率质量函数定义如下

$$p_{\text{powerlaw}}(c;\alpha) = \frac{c^{-\alpha}}{\sum_{x=1}^{C+1} c^{-\alpha}} \tag{6}$$

其中 $c \in \{1,\ldots,C\}$。令 $\tilde{\Pi}(\alpha) \in \mathbb{R}_{\geq 0}^{C}$ 表示幂律分布的向量化概率质量函数，使得 $\forall c,\ p_{\text{powerlaw}}(c;\alpha) \in \Pi(\alpha)$。图3 展示了 $C=10$ 在不同 α 取值下的幂律分布。

由于该分布包含一个描述数据不平衡程度的统计参数 α，因此可以利用领域知识或已知标签来选取合理的取值。对于不平衡程度未知的应用场景，可以估计 α。通过使用估计标签$\{y_1,\ldots,y_N\}$，可以找到最优的（γ，α）取值 $y_i = \arg\max \tau_\theta(\phi(x_i))\ \ \forall i \in \mathcal{D}$，并在交叉验证设置下测量基于这些估计标签训练的线性分类器的泛化误差；这一做法与 γ 在参考文献 [10]中的方法类似。该过程无需真实标签，而是根据算法推导出的标签来测量超平面间隔。因此，在给定的问题设置中，只需怀疑数据存在不平衡，无需其他先验知识。

B. 用于超平面估计的稀疏逻辑值

由（1）可知，$\mathcal{L}_{\text{CE}}(w_\theta^M z;\sigma(\tau_\theta(z)))$ 使用软标签来寻找最优超平面。这是因为离散的

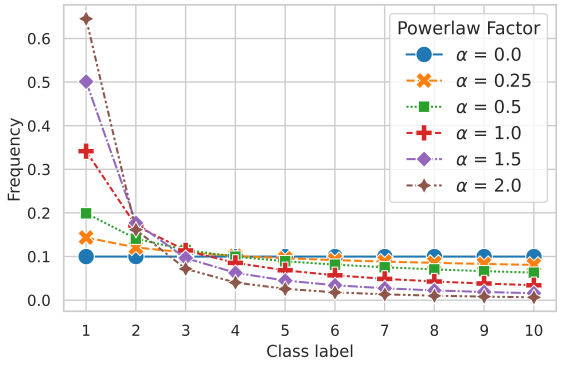

图3. 幂律分布在不同衰减率下的概率质量函数 α 当 $C = 10$ 如表I所示。

搜索空间包含 $\mathcal{O}(C^N)$ 种可能的标注，这是一个 NP难问题 [26]。通过在 θ 中的连续参数 $\tau_\theta(x) = A_\theta x + b_\theta$ 限制搜索空间，以便进行高效的梯度优化。然而，由于 softmax 运算$\sigma(\tau_\theta(z))$，每个 logit 都会参与超平面的更新；即使是数值较低、对于某个固定数据点而言不太可能是正确标签的 logit 也是如此。这是由于 softmax 函数具有全支撑。根据表 II 的结果，我们通过实验推测，这可能导致超平面的估计不够理想。

相反，我们提出应用 sparsemax 函数 [27]来过滤较小的逻辑值，该函数定义为求解

$$\text{sparsemax}(z) = \underset{p \in \Delta^{C-1}}{\arg\min} \|p - z\|_2^2 \tag{7}$$

其中，Δ^{C-1} 对应于概率单纯形。此操作返回 z 在单纯形上的欧几里得投影。单纯形投影问题有高效的求解方法，其中包括使用自动选取的阈值对较小的值进行软阈值处理 [28]。我们在损失函数中采用这一思路，并提出一种采用稀疏单纯形投影的改进聚类损失：

$$\begin{aligned} \mathcal{L}_{\text{SSP}}(\theta) = \sum_{z \in \phi(\mathcal{D})} \mathcal{L}_{\text{CE}}(w_\theta^M z; \text{sparsemax}(\tau_\theta(z))) \\ \text{s.t. } w_\theta^M = \Xi^{(M)}(w_\theta^M, \phi(\mathcal{D})) \end{aligned} \tag{8}$$

包含先验项的整体目标函数为

$$\min_\theta \mathcal{L}_{\text{SSP}}(\theta) + \gamma D_{\text{KL}}[\bar{\tau}_\theta \parallel \tilde{\Pi}(\alpha)]. \tag{9}$$

该变体称为**PET-TURTLE**(**先**验**强**化**项**TURTLE**TURTLE**）。为完整起见，PET-TURTLE 的算法伪代码见算法1。此外，该方法的 Pytorch 代码可在此处获取：github.com/javiersc1/pet-turtle。

III. 实验与结果

A. 实验设置

在我们的所有实验中，均使用 CLIP-RN50x64 [5]进行特征提取，这意味着仅使用单一表征空间。我们将 PET-TURTLE 与 TURTLE[10]进行比较，并纳入 K-Means++ 算法 [7] 以建立基线。此外，还加入线性探测 [29], 监督式线性

算法1PET-TURTLE

输入：数据集 $\mathcal{D}$、表示模型 ϕ、类别 C，以及正则化参数（γ，α）
参数：迭代次数 $T = 6000$、学习率 $\eta = 10^{-3}$，以及内部步骤 $M = 10$
提取潜变量 $\mathcal{Z} = \phi(\mathcal{D})$
对于 $t = 1$ 到 T **执行**
  采样小批量表示 $z \sim \mathcal{Z}$
  // 在逻辑值固定时更新超平面 sparsemax($\tau(z)$)
  $w_\theta^M \leftarrow w_\theta^M - \eta \frac{\partial}{\partial w_\theta}$ [代价函数（9）] 针对 1，.. .，M
  // 在超平面固定时更新分类器 ω^M
  $\tau_\theta \leftarrow \tau_\theta - \eta \frac{\partial}{\partial \tau_\theta}$ [代价函数（9）]
  如果热启动**则**
    // 使用最新的超平面初始化下一次迭代 w_θ^0
    更新起始点 $w_\theta^0 \leftarrow w_\theta^M$
结束循环
输出：簇标签 $\arg\max \tau_\theta(\mathcal{Z})$

		K-means++ [7]	TURTLE [10]	PET-TURTLE	Linear Probe [29]
CIFAR10-PL	幂律($\alpha = 0.25$)	65.5 ± 4.0	72.8 ± 0.3	**78.7*** ± 2.6	93.8
	幂律($\alpha = 0.50$)	60.3 ± 4.3	68.2 ± 1.3	**74.0*** ± 2.0	93.8
	幂律($\alpha = 1.0$)	50.9 ± 2.8	54.9 ± 3.0	**71.5*** ± 3.5	94.6
	幂律($\alpha = 1.50$)	42.5 ± 3.0	48.3 ± 4.8	**67.1*** ± 4.1	95.0
	幂律($\alpha = 2.0$)	32.8 ± 2.6	42.8 ± 5.0	**60.6*** ± 3.9	96.2

表 I 不同幂律衰减率下聚类方法在 CIFAR10-PL 数据集上的准确率结果（%）。

分类器方法，以确定在已知相关潜在特征标签的情况下所能达到的最高准确率。值得注意的是，一般而言，聚类比分类更难，因此结果可能会因问题难度而有很大差异。对于所有结果，我们报告了 10 次试验的平均值和标准差；在适用的情况下，各次试验使用不同的模型随机种子初始化，以确保透明度。此外，我们对 TURTLE 和 PET-TURTLE 的试验进行了配对 t 检验，p 值小于 0.01 的结果用“*”符号标示。

B. 合成结果

我们基于 CIFAR 生成了一个幂律不平衡数据集10 [30]，其衰减率各不相同$\alpha \in \{0.25, 0.50, 1.0, 1.50, 2.0\}$，并将这一变体记为“CIFAR10-PL”。由于 TURTLE 和 PET-TURTLE 都有一个正则化参数 γ，因此使用交叉验证从集合 $\gamma \in \{1, 5, 10, 25, 50, 100, 250, 500\}$中找出能使泛化误差最低的参数。由于在这项合成实验中衰减率 α 是已知的，因此将其固定用于 PET-TURTLE。表I 比较了 PET-TURTLE 与其他方法在不同 α 衰减率下的表现。从结果可以明显看出，PET-TURTLE 的聚类准确率更高，尤其是在数据不平衡程度较高时。

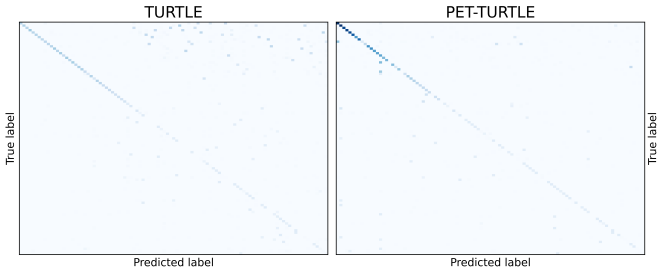

图 4. 在衰减率固定为 $\alpha = 1.0$ 时，TURTLE 和 PET-TURTLE 方法在 Food101-PL 数据集 ($C = 101$) 上的混淆矩阵。

同样，我们基于 Food101 生成了一个合成幂律数据集[31]，固定 $\alpha = 1.0$，其中包含 $C = 101$个类别，并将其记为“Food101-PL”。在这项实验中，仅比较 TURTLE 和 PET-TURTLE，以研究过度预测的影响。图4展示了真实标签与预测簇标签之间的混淆矩阵。我们使用“匈牙利算法” [32] 解决簇标签置换问题，以确保真实标签与簇标签之间的对应关系最优。图4中，TURTLE 的混淆矩阵在上三角区域包含许多元素，表明许多属于多数类簇的样本被划分到了少数类簇中，因而导致聚类效果较差。相比之下，PET-TURTLE 不仅非对角元素更少，而且对于各簇所含样本数，其颜色也更符合预期地从深蓝色过渡到浅蓝色，这正是不平衡数据问题中应有的表现。

C. 真实数据结果

最后，表II对平衡和不平衡数据集进行了真实数据对比。对于平衡数据集，使用了：加州理工学院101 [33], CIFAR 10 [30], DTD [34],EuroSAT [35], 和 Food101 [31]。对于不平衡数据集，使用了：血细胞显微图像（血液） [36],、皮肤镜（皮肤） [37], 、iNaturalist 2017 [38], 、视网膜光学相干断层扫描（光学相干断层扫描） [39], 、腹部轴位计算机断层扫描（器官A） [40], 和肾皮质显微图像（组织） [41]。使用 MedMNIST [42]软件包加载医学影像数据集。所有数据集的类别数 C 均为已知。交叉验证方式与之前类似。然而，在此情况下，α 未知，必须进行估计。采用网格搜索确定 $\alpha \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1.0, 1.25, 1.50, 1.75, 2.0\}$，并选择使验证误差最低的（$\gamma$，$\alpha$）组合。在平衡数据场景下，PET-TURTLE 平均相比 TURTLE 实现了 ~3%的准确率提升，这得益于用于超平面估计的稀疏逻辑值思想。在不平衡场景下，PET-TURTLE 平均相比 TURTLE 实现了 ~15%的准确率提升，这得益于先验约束项。需要注意的是，这些不平衡数据集并不严格遵循幂律分布，因此聚类质量可能仍有进一步提升的空间。从某种意义上说，这也检验了我们的方法应对分布失配的稳健性。

	数据集	C (num. of classes)	K-means++ [7]	TURTLE [10]	PET-TURTLE	Linear Probe [29]
Balanced	加州理工学院 [33]	101	76.6 ± 4.5	85.3 ± 0.9	**88.2*** ± 0.5	96.9
	CIFAR [30]	10	61.7 ± 4.2	76.3 ± 0.1	**80.9*** ± 1.6	94.1
	DTD [34]	47	48.2 ± 1.8	57.6 ± 0.9	**59.4** ± 1.5	82.5
	EuroSAT [35]	10	60.0 ± 6.4	77.6 ± 0.2	**80.4*** ± 1.6	95.2
	食物 [31]	101	63.1 ± 1.6	85.4 ± 1.1	**88.1*** ± 1.0	94.4
Imbalanced	血液 [36]	8	44.9 ± 0.7	48.1 ± 1.8	**56.1*** ± 2.7	96.1
	皮肤 [37]	7	25.4 ± 0.7	34.1 ± 0.3	**67.1*** ± 0.3	81.8
	iNaturalist [38]	13	43.6 ± 1.7	55.9 ± 3.7	**77.5*** ± 1.0	98.1
	光学相干断层扫描 [39]	4	46.2 ± 0.9	54.1 ± 0.7	**61.4*** ± 0.7	93.4
	器官A [40]	11	43.5 ± 2.8	45.0 ± 1.1	**51.9*** ± 2.4	89.9
	组织 [41]	8	26.3 ± 0.2	30.6 ± 1.8	**38.7*** ± 1.2	60.5

表II 聚类方法在真实平衡和不平衡图像数据集上的准确率结果（%）。

IV. 结论

总之，本文提出了 TURTLE 聚类算法的一种有原则的扩展，通过引入先验分布项，解决了其在不平衡数据场景下的局限性。该方法在合成数据和真实世界场景中均展现出稳健的聚类准确率提升，尤其是在类别不平衡显著的场景下。这些实证收益凸显了 PET-TURTLE 对于基础模型之上无监督聚类应用的实际价值，为在存在不平衡数据分布的情况下实现更可靠的聚类方法铺平了道路。

然而，该方法依赖于从采用某种自监督范式训练的基础模型中提取的特征。因此，如果直接将 TURTLE 或 PET-TURTLE 应用于呈现非线性结构的环境空间数据（$x \in \mathcal{D}$），那么由于这两种方法期望数据位于更加线性可分的空间中（例如由基础模型诱导的空间），它们的表现都不会很好。类似地，这也是 核支持向量机在非线性环境下比传统支持向量机更强大的原因。探索 PET-TURTLE 在这一非线性区域中的扩展将会很有意义。鉴于潜在空间（$z \in \phi(\mathcal{D})$）中可能存在一定的非线性，这取决于问题的难度或训练分布，因此在基础模型背景下，这一扩展可能会很有用，因为这些非线性会使寻找最优超平面更具挑战性。进一步探究这一主题将是未来工作中一个有趣的方向。

V. 影响声明

本工作的主要目标是推动聚类领域的发展，而所提出的方法依赖于基础模型的表示空间。众所周知，这些模型会继承训练数据中蕴含的偏差 [43]。因此，在医学影像等敏感领域使用 PET-TURTLE 时，建议谨慎行事。

参考文献

[1] S. J. Pan 和 Q. Yang，“迁移学习综述”，IEEE 知识与数据工程汇刊，第 22 卷，2009年，第 1345–1359 页。[2] T. Darcet、M. Oquab、J. Mairal 和 P. Bojanowski，“视觉变换器需要寄存器”，载于《第十二届国际学习表征会议》。国际学习表征会议，2024年。[3] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等，“Dinov2：在无监督条件下学习鲁棒视觉特征”，《机器学习研究汇刊》，第 1–31 页，2024年。[4] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick，“掩码自编码器是可扩展的视觉学习器”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》。国际学习表征会议，2022年，第 16 000–16 009 页。[5] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，“通过自然语言监督学习可迁移的视觉模型”，载于《国际机器学习会议》。PMLR，2021年，第 8748–8763 页。[6] H. Hu、X. Wang、Y. Zhang、Q. Chen 和 Q. Guan，“对比学习综合综述”，《神经计算》，第 128645 页，2024年。[7] D. Arthur 和 S. Vassilvitskii，“k-means++ 谨慎初始化的优势”，载于《第十八届ACM-SIAM离散算法研讨会论文集》，2007年，第 1027–1035 页。[8] J. Salazar Cavazos、J. A. Fessler 和 L. Balzano，“ALPCAHUS：异方差数据的子空间聚类”，arXiv 预印本 arXiv:2505 18918，2025年。[9] J. Zhou、C. Wei、H. Wang、W. Shen、C. Xie、A. Yuille 和 T. Kong，“iBOT：采用在线分词器的图像BERT预训练”，载于《国际学习表征会议》。国际学习表征会议，2022年。[10] A. Gadetsky、Y. Jiang 和 M. Brbic，“摆脱标签，进行无监督迁移”，载于《国际机器学习会议》。PMLR，2024年，第 14 382–14 407 页。[11] J. Xie、R. Girshick 和 A. Farhadi，“用于聚类分析的无监督深度嵌入”，载于《国际机器学习会议》。国际机器学习会议，2016年，第 478–487 页。[12] J. Chang、L. Wang、G. Meng、S. Xiang 和 C. Pan，“深度自适应图像聚类”，载于《IEEE国际计算机视觉会议论文集》，2017年，第 5879–5887 页。[13] M. Caron、P. Bojanowski、A. Joulin 和 M. Douze，“用于无监督视觉特征学习的深度聚类”，载于《欧洲计算机视觉会议论文集（欧洲计算机视觉会议）》，2018年，第 132–149 页。[14] C. Niu、H. Shan 和 G. Wang，“SPICE：用于图像聚类的语义伪标签”，《IEEE图像处理汇刊》，第 31 卷，第 7264–7278 页，2022年。[15] A. Mao、M. Mohri 和 Y. Zhong，“交叉熵损失函数：理论分析与应用”，载于《国际机器学习会议》。PMLR，2023年，第 23 803–23 828 页。[16] D. Soudry、E. Hoffer、M. S. Nacson、S. Gunasekar 和 N. Srebro，“梯度下降在可分数据上的隐式偏差”，《机器学习研究杂志》，第 19 卷，第 70 期，第 1–57 页，2018年。[17] Z. Ji 和 M. Telgarsky，“梯度下降在不可分数据上的隐式偏差”，载于《学习理论会议》。PMLR，2019年，第 1772–1798 页。[18] A. Gadetsky 和 M. Brbic，“追寻人工标注：无监督学习的新视角”，《神经信息处理系统进展》，第 36 卷，2024年。[19] M. Assran、R. Balestriero、Q. Duval、F. Bordes、I. Misra、P. Bojanowski、P. Vincent、M. Rabbat 和 N. Ballas，“自监督学习中隐藏的均匀聚类先验”，载于《国际学习表征会议》。国际学习表征会议，2023年。[20] H. Zhou、T. Luo 和 Y. He，“通过异构知识迁移实现长尾视觉识别的动态协同学习”，《信息融合》，第 115 卷，第 102734 页，2025年。[21] X. Zhang 和 T. Luo，“基于张量积的语义融合实现不平衡多实例多标签学习”，《计算机科学前沿》，第 19 卷，第 8 期，第 198346 页，2025年。[22] J. Salazar Cavazos、J. A. Fessler 和 L. Balzano，“ALPCAH：样本级异方差数据的子空间学习”，《IEEE信号处理汇刊》，2025年。[23] S. Kullback 和 R. A. Leibler，“关于信息与充分性”，《数学统计年刊》，第 22 卷，第 1 期，第 79–86 页，1951年。[24] H. Zhou、T. Luo、J. Zhang 和 L. Liu，“通过因果特征增强网络探究场景图生成中的关系本质”，《IEEE《模式分析与机器智能汇刊》

，2025年。

[25] A. Clauset、C. R. Shalizi 和 M. E. Newman，《经验数据中的幂律分布》，SIAM评论，第 51 卷，第 4 期，第 661–703 页，2009年。[26] M. Mahajan、P. Nimbhorkar 和 K. Varadarajan，《平面 k 均值问题是 NP 难问题》，《理论计算机科学》，第 442 卷，21，2012 13pp. – 。[27] A. Martins 和 R. Astudillo，《从 softmax 到 sparsemax：一种用于注意力和多标签分类的稀疏模型》，载于《国际机器学习会议》。PMLR，2016，第 1614–1623 页。[28] J. Duchi、S. Shalev-Shwartz、Y. Singer 和 T. Chandra，《用于高维学习的 l 1-球高效投影》，载于第 25 届国际机器学习会议论文集，2008，第 272–279 页。[29] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，《用于视觉表征对比学习的简单框架》，载于《国际机器学习会议》。PMLR，2020，第 1597–1607 页。[30] A. Krizhevsky，《从微型图像中学习多层特征》，硕士论文，多伦多大学，2009年。[31] L. Bossard、M. Guillaumin 和 L. Van Gool，《Food-101——利用随机森林挖掘判别性成分》，载于《欧洲计算机视觉会议》。施普林格，2014，第 446–461 页。[32] H. W. Kuhn，《用于指派问题的匈牙利算法》，《海军研究后勤季刊》，第 2 卷，第 1-2 期，第 83–97 页，1955。[33] L. Fei-Fei、R. Fergus 和 P. Perona，《物体类别的单样本学习》，IEEE《模式分析与机器智能汇刊》，第 28 卷，第 4 期，第 594–611 页，2006。[34] M. Cimpoi、S. Maji、I. Kokkinos、S. Mohamed 和 A. Vedaldi，《野外环境中的纹理描述》，载于 IEEE计算机视觉与模式识别会议论文集，2014，第 3606–3613 页。[35] P. Helber、B. Bischke、A. Dengel 和 D. Borth，《EuroSAT：用于土地利用和土地覆盖分类的新型数据集与深度学习基准》，IEEE《应用地球观测与遥感精选主题期刊》，第 12 卷，第 7 期，第 2217–2226 页，2019年。[36] A. Acevedo、A. Merino、S. Alf´erez、 ´A. Molina、L. Bold´u 和 J. Rodellar，《用于开发自动识别系统的外周血细胞显微图像数据集》，《数据简报》，第 30 卷，105474，2020 p. 。[37] P. Tschandl、C. Rosendahl 和 H. Kittler，《HAM10000 数据集：常见色素性皮肤病变的多源皮肤镜图像大型集合》，《科学数据》，第 5 卷，第 1 期，第 1–9 页，2018年。[38] G. Van Horn、O. Mac Aodha、Y. Song、Y. Cui、C. Sun、A. Shepard、H. Adam、P. Perona 和 S. Belongie，《iNaturalist 物种分类与检测数据集》，载于 IEEE计算机视觉与模式识别会议论文集，2018，第 8769–8778 页。[39] D. S. Kermany、M. Goldbaum、W. Cai、C. C. Valentim、H. Liang、S. L. Baxter、A. McKeown、G. Yang、X. Wu、F. Yan 等，《通过基于图像的深度学习识别医学诊断和可治疗疾病》，《细胞》，第 172 卷，第 5 期，第 1122–1131 页，2018年。[40] P. Bilic、P. Christ、H. B. Li、E. Vorontsov、A. Ben-Cohen、G. Kaissis、A. Szeskin、C. Jacobs、G. E. H. Mamani、G. Chartrand 等，《肝肿瘤分割基准（LiTS）》，《医学图像分析》，第 84 卷，文章编号 102680，2023年。[41] V. Ljosa、K. L. Sokolnicki 和 A. E. Carpenter，《用于验证的带注释高通量显微图像集》，《自然·方法》，第 9 卷，第 7 期，第 637 页，2012。[42] J. Yang、R. Shi、D. Wei、Z. Liu、L. Zhao、B. Ke、H. Pfister 和 B. Ni，《MedMNIST v2——用于二维和三维生物医学图像分类的大规模轻量级基准》，《科学数据》，第 10 卷，41，2023 1，no. p. 。[43] W. F. Wiggins 和 A. S. Tejani，《放射学自然语言处理中基础模型的机遇与风险》，《放射学：人工智能》，第 4 卷，第 4 期，文章编号 e220119，2022年。