arXiv:2603.20856v1 [cs.CV] 21 Mar 2026

用于类别不平衡白细胞分类的小型分类器集成

Siddharth Srivastava[1,⋆], Adam Smith[1], Scott Brooks[2,3], Jack Bacon[1], Till Bretschneider[1]

[1]计算机科学系，华威大学 [2]华威医学院，华威大学 [3]Intelligent Imaging Innovations Ltd ⋆通讯作者：siddharth.srivastava@warwick.ac.uk

摘要

自动化白细胞分类以诊断白血病，是一种有前景的替代方案，可取代由专业病理学家进行的耗时且耗费资源的细胞检查。然而，由于染色、扫描以及患者间异质性的差异，为罕见细胞类型设计稳健的分类算法仍然面临挑战。我们提出了一种轻量级集成方法，用于造血过程中的细胞分类，并重点关注粒细胞生成、单核细胞生成和淋巴细胞生成的生物学特征。通过扩展数据集以缓解部分类别不平衡问题，我们证明，由轻量级预训练 SwinV2-Tiny、DinoBloom-Small 和 ConvNeXT-V2-Tiny 模型组成的简单集成，在这一具有挑战性的数据集上表现出色。我们在分层的 3 折交叉验证框架中，为每种架构训练 3 个实例；对于输入图像，我们将其依次输入全部 9 个模型，并通过 logit 平均汇总结果。我们还分析了模型的不足之处：在粒细胞生成过程中容易混淆外观相似的骨髓细胞，在淋巴细胞生成过程中也容易混淆淋巴细胞。代码：https://gitlab.com/siddharthsrivastava/wbc-bench-2026

1 引言

血液学图像分析对于白血病的早期诊断具有重要的临床意义。白血病是一组起源于骨髓的癌症，会导致异常白细胞（WBC）过量生成。传统诊断通常依赖人类专家进行显微镜检查，这既耗费资源，又耗时。自动白细胞分类有望成为一种资源消耗较低的替代方案，用于检测稀有血细胞。虽然已有一些方法逐渐涌现，但在应用于临床之前仍需进一步验证，尤其是其稳健性方面 [1]。创建 ISBI WBCBench 2026：稳健白细胞分类挑战赛 [2]，旨在对用于将白细胞分为 13 类的机器学习模型进行基准测试。这些类别体现了白细胞的发育过程，经历

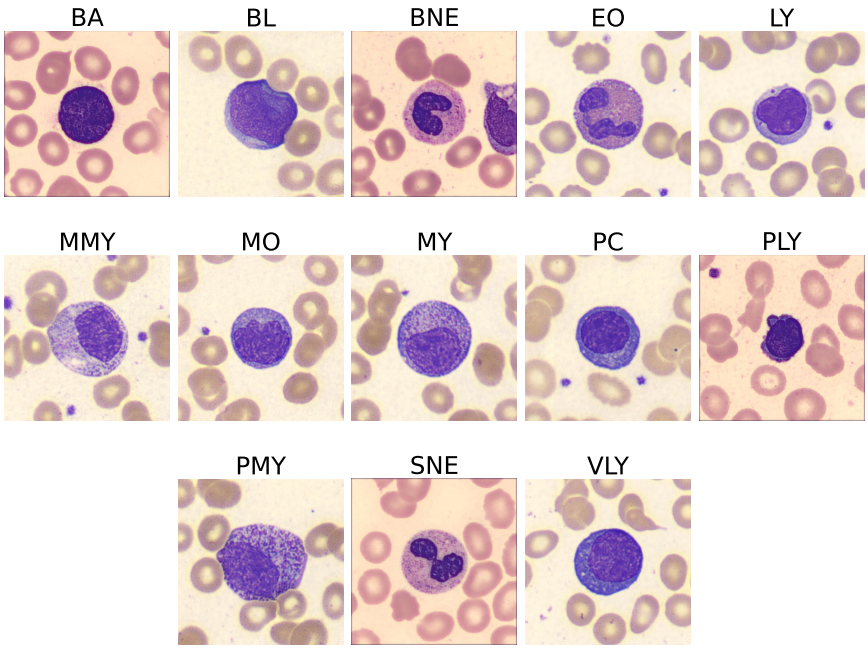


图1：WBCBench竞赛数据集中各类别的第1阶段（未扰动）图像示例。

造血沿着三条主要分支进行，产生粒细胞、单核细胞和淋巴细胞这几大类。 [3]。尽管用于细胞分类 [4], 的深度学习模型不断进步，但由于训练数据存在固有局限，例如噪声、染色变化和类别不平衡，这一问题仍然颇具挑战。为解决这一问题，我们提出一种由轻量级分类器构成的集成方法，旨在以有限的计算资源实现接近最先进的性能。

2 数据集

2.1 数据集扩展

挑战数据集（WBCBench [2]）包含 55，012 张大小为 368×368 的苏木精-伊红染色外周血涂片图像，涵盖造血过程中出现的 13 类细胞：分叶中性粒细胞（SNE）、淋巴细胞（LY）、单核细胞（MO）、原始细胞（BL）、嗜酸性粒细胞（EO）、骨髓细胞（MY）、杆状核中性粒细胞（BNE）、变异（非典型）淋巴细胞（VLY）、中幼粒细胞（MMY）、早幼粒细胞（PMY）、浆细胞（PC）和前淋巴细胞（PLY）；图1展示了各类别的示例图像。16，477 张图像未标注，留作测试，因此有 38，535 张已标注图像用于模型训练。该数据集类别极不平衡，SNE 和 LY 两类合计占 28，155 张图像；而数量最少的两类 PC 和 PLY 分别仅有 90 张和 14 张。因此，我们使用外部公开数据集扩充了该数据集：

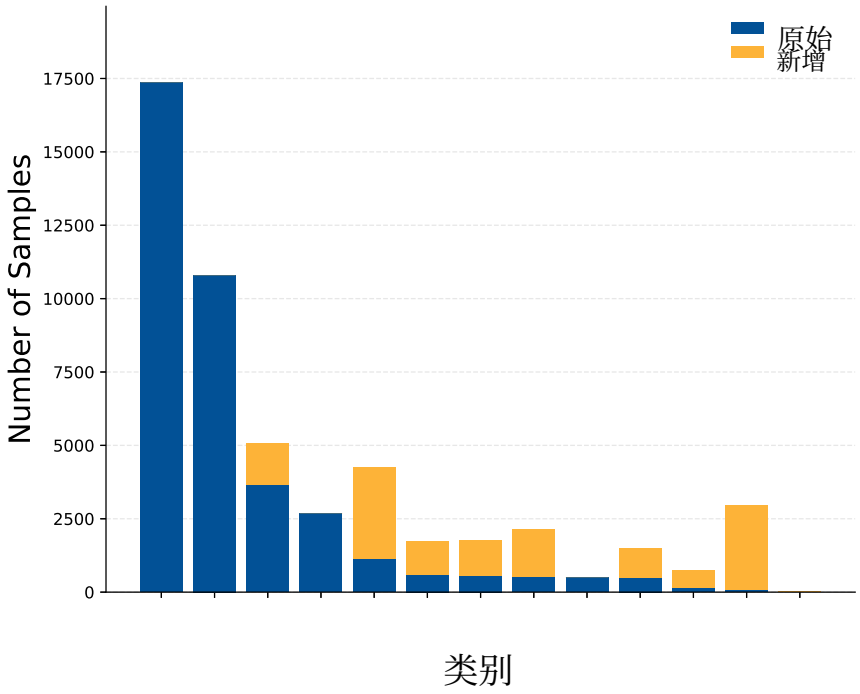


图 2：扩展数据集中每个类别的样本数量。

• *Acevedo-20* [5]：超过 17，000 张图像，其中我们使用了以下类别的 10， 312 张图像：MO（1420）、EO（3117）、BNE（1633）、PMY（592）、BA（1218）、MY（1137）和 MMY（1015）；
• ***Taeyeon Kim*** 的血液数据集 [6], 最初以血液八分类数据集的形式开发 [7]：包含来自 8 个类别的 46,000 多张图像，我们仅使用其中的 2895 张浆细胞图像。

• 来自 CellWiki 的幼淋巴细胞图像 [8]：新增 18 个幼淋巴细胞样本，使 PLY 实例总数达到 32 个。

总计，我们额外纳入了 13,045 个训练样本，缓解了原始数据集中一些最严重的类别不平衡问题（各类别的频次见图 2）。表 1 展示了加入外部公开数据后，模型的鲁棒性有所提升。

为模拟扫描仪和染色差异造成的真实域偏移，WBCBench 引入了高斯噪声、运动模糊和颜色扰动等形式的数据增强。运动模糊和颜色扰动可通过训练时数据增强来处理，而对于噪声退化，我们采用非局部均值去噪进行自适应处理 [9]，并使用稳健的小波估计器估算高斯噪声的标准差 $\hat{}\sigma$ [10]。去噪所用的图像块距离设为 $\sqrt{\hat{}}\sigma$。经验表明，这种设置既能保留干净图像的精细形态细节，也能充分去除噪声图像中的噪声。训练数据集中的类别不平衡极为严重，因此训练时采样批次需要使用加权 采样器，使稀有类别与最常见类别以相同的频率被采样。因此，我们根据有效数量 E_n [11]对每个类别赋予权重：

$$E_n = (1 - \beta^n)/(1 - \beta),$$

n 表示类别规模， $\beta = 0.9999$ 为超参数。

3 方法

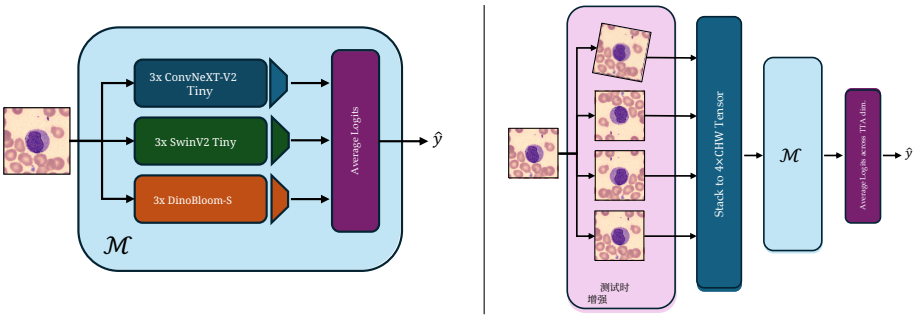


图 3：模型架构。**左侧：**我们的模型由 3 个不同架构组成的 3 模型集成构成：ConvNeXT-V2 Tiny [12], 、SwinV2-Tiny [13], 和 DinoBloom-Small [14] 模型。对于每个模型，我们都将预训练分类头替换为简单的多层感知机（MLP） h。每种架构的实例都在 3 个数据折中的 2 个上独立微调，并在最后一个折上验证，以减少过拟合；因此，每种架构有 3 个模型，总计 9 个模型。最终模型 $\mathcal{M}$ 将输入图像传入全部 9 种架构，并对 logits 取平均，从而生成 13 类逻辑值向量。**右侧**：推理时，我们通过随机翻转和旋转为每张图像生成多个增强视图。这些视图由 $\mathcal{M}$ 独立处理，并对其 logits 取平均，以获得最终类别预测。

3.1 模型架构

我们在图 3 中概述了模型架构，该架构由以下 3 个预训练分类器的实例组成：

- Swin Transformer V2 [13]：一种分层视觉变换器，其表征由互不重叠的移位窗口组成。Swin-V2 通过引入稳定训练的技术、更好地处理不同的图像分辨率，以及采用减少训练数据量的新预训练方法，改进了原始的 Swin 变换器架构 [15] 。我们使用包含 2750 万个参数的 swinv2-tiny，并将分类头替换为一个小型多层感知机，使参数总数达到 28.1M。

- ConvNeXT-v2 [12]：一种纯卷积神经网络，通过使用层归一化层取代批归一化层、使用高斯误差线性单元取代修正线性单元，并减少激活函数的数量，从而改进了残差网络 [16]。ConvNeXT-V2 在原版 ConvNeXT 的基础上，采用全卷积掩码自编码器框架和全局响应归一化层，进一步提升了性能 [12]。我们使用 convnextv2-tiny，并同样将分类头替换为多层感知机，总参数量为 28.4M。

• DinoBloom [14]：一种基于 DINOv2 [17] 构建的单细胞血液学图像基础模型。我们使用 DinoBloom-Small，并搭配自定义分类头，参数总量为 23M。

所选骨干模型具有互补的归纳偏置：SwinV2 是一种基于Transformer的模型，采用层次化注意力机制和移位窗口，使图像中的长程图像块能够相互作用并进行推理。相比之下，ConvNeXT-V2 保留了卷积神经网络固有的局部性和平移等变性偏置 [16]„，这对于捕捉细粒度纹理和形态学模式十分有效。最后，DinoBloom 是一种面向血液学分析的领域专用基础模型，不同于基于自然图像训练的 ConvNeXT 和 Swin [13, 12], ，它所提供的表征更适合细胞形态学和 H&E 染色。结合具有不同偏置和训练数据的多种网络，为图像分类提供了丰富的图像表征。所有MLP分类头均由4个 Linear-ReLU-Dropout 模块构成（$p = 0.1$）

3.2 模型训练

我们将扩充后的训练数据集（第2.1节）划分为3个分层折，并为每种架构训练3个独立实例；每个实例使用其中两折进行训练，剩余一折用于验证，以限制过拟合。对于每个模型，我们使用 α平衡焦点损失 [18], ，并将 α 和 γ 分别设为0.25和2。焦点损失会降低已正确分类样本的相对损失（其 p_t 较高），并将重点转向难以分类的错误样本（其 p_t较低）。我们使用批量大小32进行训练，采用 AdamW优化器 [19], ，其参数为 $\beta =$ (0.9,0.999)，并使用从 5×10^{-4} 开始、衰减至 5×10^{-6}的余弦衰减学习率，不进行学习率预热；在验证集宏平均F1连续10个 epoch 未改善时停止训练。我们对模型权重使用指数移动平均（EMA），并将 β 设为0.999。最后，我们采用图像增强来模拟训练数据中的损坏并提高鲁棒性：随机翻转、旋转、高斯噪声与模糊、运动模糊，以及随机调整亮度和对比度。此外，每个批次中 CutMix [20] 和 MixUp [21], 的使用概率均为0.15。在全参数微调时，每个模型在3张 NVIDIA A10 显卡上并行训练，每个 epoch 耗时5分钟。

3.3 模型推理

在推理过程中，我们执行测试时增强，如图 3 右侧所示；我们将一张测试图像依次进行一系列增强（随机翻转和随机旋转），再将这些图像输入模型，并对输出的 logits 求平均，以获得原始输入图像的“平滑”logit 表示。使用我们的 9 个模型在单台 NVIDIA A40 上完整遍历包含 16,447 张图像的测试数据集需要 30 分钟。

4 结果

我们在表 1 中报告宏平均 F1 分数、平衡准确率、宏平均精确率和宏平均特异度。这些指标根据折外预测计算得出：对于每个折 $k \in \{1, 2, 3\}$，模型在 2 个折上训练，并在

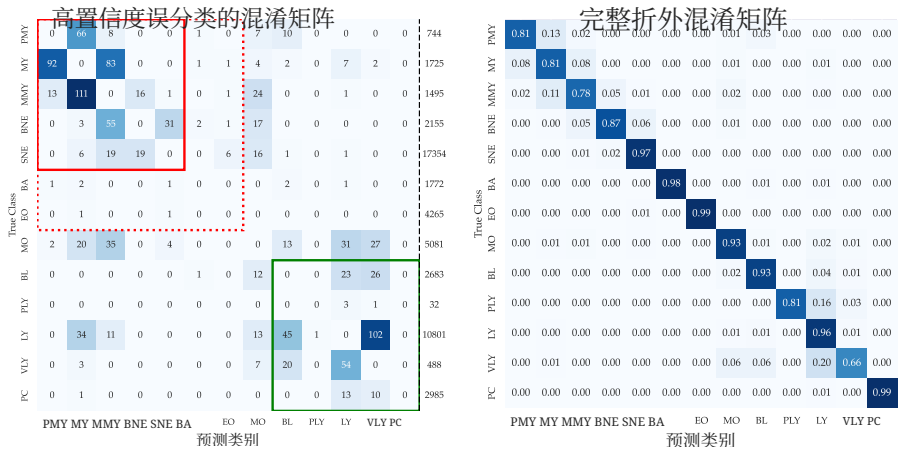


图4：在扩展数据集上训练的集成模型的混淆矩阵。**左侧**：1149 个高置信度误分类样本的混淆矩阵，其中 841 个来自 WBCBench，其余来自 Acevedo20；通过可信学习 [22]发现。红色虚线框标出粒细胞生成中的类别，绿色框标出淋巴细胞生成中的类别；MO 属于单核细胞生成。**右侧**：使用各组模型的折外样本生成的混淆矩阵。

		训练数据集：WBCBench 训练集 +评估集					训练数据集：扩展数据集				
		评估集：WBCBench 训练集 +评估集 折外				WBCBench Test	评估集：WBCBench 训练 +Eval 折外				WBCBench Test
	模型	宏平均 F1	平衡 准确率	宏平均 精确率	宏平均 灵敏度	宏平均 F1	宏平均 F1	平衡 准确率	宏平均 精确率	宏平均 灵敏度	宏平均 F1
No TTA	3×SwinV2	0.7351	0.7759	0.7126	0.9932		0.7410	0.7523	0.7412	0.9936	
	3×ConvNeXt	0.7523	0.7669	0.7488	0.9938	-	0.7662	0.7660	0.7741	0.9942	-
	3×DinoBloom	0.7276	0.8095	0.6863	0.9925		0.7417	**0.7739**	0.7296	0.9936	
	集成	0.7666	0.789	0.7586	0.9941	0.6638	0.7791	0.7736	0.7977	**0.9945**	0.6737
TTA	3×SwinV2	0.7477	0.7806	0.7306	0.9935		0.7499	0.7503	0.7666	0.9939	
	3×ConvNeXt	0.7557	0.7688	0.7535	0.9939	-	0.7724	0.7696	0.7815	0.9944	-
	3×DinoBloom	0.7323	**0.8166**	0.6901	0.9928		0.7419	0.7718	0.7321	0.9939	
	集成	**0.7741**	0.7987	**0.7634**	**0.9942**	**0.6674**	**0.7798**	0.7734	**0.7989**	**0.9945**	**0.6772**

表 1：对 SwinV2-Tiny、DinoBloom-Small、ConvNeXt-Tiny 及其集成模型进行定量比较，这些模型分别在两个数据集上训练：WBCBench Train+ Eval 和扩展数据集。结果分别报告了使用和未使用测试时增强（TTA）的情况。指标根据折外预测计算，以提供无偏评估；测试集宏平均F1（以粉色突出显示）作为留出基准。数值越高表示性能越好；**粗体** 表示最佳性能，下划线表示次佳性能。我们的最佳模型是在扩展数据集上训练并使用 TTA 的集成模型。

留出折 k上评估，因此每个样本都恰好由训练过程中未接触过它的模型评估一次，从而得到无数据泄漏的逻辑值，确保评估无偏。结果显示，在扩展数据集上训练的全部三个骨干网络组成的集成，在所有指标上表现最佳。我们还在图4右侧展示了集成模型的折外混淆矩阵，其中揭示了一个有趣的生物学发现：矩阵左上方的骨髓细胞常常彼此混淆，而 VLY 和 PLY 类别最常与淋巴细胞混淆。为突出这一点，我们在图4左侧展示了通过置信学习 [22] 发现的高置信度错误分类（即模型赋予真实类别的概率极低）的混淆矩阵。

左上方的红色虚线框显示了粒细胞生成类别；排除嗜碱性粒细胞和嗜酸性粒细胞后，红色实线框显示这些类别最容易混淆，因为从生物学角度看，它们代表了细胞从早幼粒细胞逐渐成熟为分叶核中性粒细胞的过程。绿色框显示了淋巴细胞生成类别，这些类别之间也常常相互混淆。最后，混淆矩阵显示，标记为“原始细胞”的细胞很可能是淋巴母细胞而非髓母细胞，因为它们会与 LY 和 VLY 一同混淆。

5 结论与讨论

我们展示了，对一组小型预训练分类器组成的流程进行微调，能够在白细胞图像这一高度类别不平衡的分类数据集上取得出色表现。我们的方法在分层三折交叉验证框架下，集成了三种不同架构——自监督 DiNO、卷积神经网络和分层视觉变换器。我们在最终竞赛测试集上的最佳得分为 0.67726。尽管表现强劲，我们仍认为，错误最常出现在骨髓细胞和淋巴细胞中生物学特征相似的类别之间；我们还推测，未来的工作若能针对粒细胞生成、淋巴细胞生成和单核细胞生成设计专门的专家模型，以学习更强的类间表征并提高分类性能，将会有所助益。

6 伦理标准合规性

这是一项数值模拟研究，无需伦理审批。

7 致谢

我们感谢 [2]提供 WBCBench 数据集。作者感谢在完成本研究过程中，得以使用华威大学计算机科学系的批处理计算系统及相关支持服务。T.B. 获英国工程与自然科学研究理事会资助，项目编号 EP/V062522/1。J.B. 获英国工程与自然科学研究理事会资助，项目参考编号 2882348。T.B. 和 A.S. 获英国工程与自然科学研究理事会/美国国家科学基金会联合资助，项目编号 EP/X026663/1。S.B. 获英国创新署通过华威大学与 Intelligent Imaging Innovations Ltd 之间的知识转移合作伙伴关系资助（资助编号 10159795）。

参考文献

[1] Ario Sadafi、Raheleh Salehi、Armin Gruber、Sayedali Shetab Boushehri、Pascal Giehr、Nassir Navab 和 Carsten Marr，“一种用于跨域白细胞分类的持续学习方法”，2023年。

[2] Xin Tian、Xudong Ma、Tianqi Yang、Alin Achim、Bartek Papiez、Phandee Watanaboonyongcharoen 和 Nantheera Anantrasirichai，《WBCBench 2026：针对类别不平衡的鲁棒白细胞分类挑战》，发表于国际生物医学成像研讨会，2026年。

[3] Madhumita Jagannathan-Bogdan 和 Leonard I Zon, 《造血作用》, *Development*, 2013年。[4] Mohammad Shifat-E-Rabbi、Xuwang Yin、Cailey E Fitzgerald 和 Gus- tavo K Rohde, 《细胞图像分类：比较概述》, *Cy-tometry Part A*, 2020年。[5] Andrea Acevedo、Anna Merino、Santiago Alf´erez、 ´Angel Molina、Laura Bold´u 和 Jos´e Rodellar, 《用于开发自动识别系统的外周血细胞显微图像数据集》, *Data in Brief*, 2020年。[6] Taeyoon Kim, 《ehottl/blood 数据集》, 2023年。[7] Falah G. Salieh, 《血液八分类数据集》, 2023年。

[8] CellWiki, "前淋巴细胞"。

[9] A. Buades、B. Coll 和 J.-M. Morel, 《一种用于图像去噪的非局部算法》, 发表于计算机视觉与模式识别会议, 2005年。[10]David L Donoho 和 Iain M Johnstone, 《通过小波收缩实现理想的空间自适应》, 发表于*Biometrika*, 1994年。[11]Yin Cui、Menglin Jia、Tsung-Yi Lin、Yang Song 和 Serge Belongie, 《基于样本有效数量的类别平衡损失》, 发表于计算机视觉与模式识别会议, 2019年。[12] Sanghyun Woo、Shoubhik Debnath、Ronghang Hu、Xinlei Chen、Zhuang Liu、In So Kweon 和 Saining Xie, 《ConvNeXt V2：利用掩码自编码器协同设计并扩展卷积网络》, 发表于计算机视觉与模式识别会议, 2023年。[13] Ze Liu、Han Hu、Yutong Lin、Zhuliang Yao、Zhenda Xie、Yixuan Wei、Jia Ning、Yue Cao、Zheng Zhang、Li Dong、Furu Wei 和 Baining Guo, 《Swin Transformer V2：扩展模型容量与分辨率》, 发表于计算机视觉与模式识别会议, 2022年6月。[14] Valentin Koch、Sophia J. Wagner、Salome Kazeminia、Ece Sancar、Matthias Hehr、Julia A. Schnabel、Tingying Peng 和 Carsten Marr, 《DinoBloom：用于血液学中通用细胞嵌入的基础模型》, 发表于医学图像计算与计算机辅助干预会议, 2024年。[15] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo, 《Swin Transformer：采用移位窗口的分层视觉变换器》, 发表于*ICCV*, 2021年。[16] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feichtenhofer、Trevor Darrell 和 Saining Xie, 《面向2020年代的卷积网络》, 发表于计算机视觉与模式识别会议, 2022年。[17] Maxime Oquab、Timoth´ee Darcet、Theo Moutakanni、Huy V. Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby、Russell Howes、Po-Yao Huang、Hu Xu、Vasu Sharma、Shang-Wen Li、Wojciech Galuba、Mike Rabbat、Mido Assran、Nicolas Ballas、Gabriel Synnaeve、Ishan Misra、Herve Jegou、Julien Mairal、Patrick Labatut、Armand Joulin 和 Piotr Bojanowski, 《DINOv2：无需监督学习，学习稳健的视觉特征》, 2023年。

[18] Tsung-Yi Lin、Priya Goyal、Ross Girshick、何恺明和 Piotr Doll´ar, 《用于密集目标检测的焦点损失》，发表于*ICCV*，2017。[19] Ilya Loshchilov 和 Frank Hutter, 《解耦权重衰减正则化》，发表于*ICLR*, 2019。[20] Sangdoo Yun、Dongyoon Han、Seong Joon Oh、Sanghyuk Chun、Junsuk Choe 和 Youngjoon Yoo, 《CutMix：利用可定位特征训练强分类器的正则化策略》，2019年。[21] Hongyi Zhang、Moustapha Cisse、Yann N. Dauphin 和 David Lopez-Paz, 《MixUp：超越经验风险最小化》，2018年。[22] Curtis G. Northcutt、Lu Jiang 和 Isaac L. Chuang, 《置信学习：估计数据集标签中的不确定性》，发表于*JAIR*，2021。