用于吉姆萨染色血涂片中疟原虫检测的 COCO 格式实例级数据集

Frauke **Wilm** [1,2], Luis Carlos Rivera Monroy [1, 2], Mathias Öttl [1, 2], Lukas Mürdter [1], Leonid Mill [1, 2], Andreas Maier [2]

1 MIRA Vision Microscopy GmbH, 73037 Göppingen, Germany
2 Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany

Abstract
Accurate detection of *Plasmodium falciparum* in Giemsa-stained blood smears is an essential component of reliable malaria diagnosis, especially in developing countries. Deep learning-based object detection methods have demonstrated strong potential for automated Malaria diagnosis, but their adoption is limited by the scarcity of datasets with detailed instance-level annotations. In this work, we present an enhanced version of the publicly available NIH malaria dataset, with detailed bounding box annotations in COCO format to support object detection training. We validated the revised annotations by training a Faster R-CNN model to detect infected and non-infected red blood cells, as well as white blood cells. Cross-validation on the original dataset yielded F1 scores of up to 0.88 for infected cell detection. These results underscore the importance of annotation volume and consistency, and demonstrate that automated annotation refinement combined with targeted manual correction can produce training data of sufficient quality for robust detection performance. The updated annotations set is publicly available via GitHub: `https://github.com/MIRA-Vision-Microscopy/malaria-thin-smear-coco`.

Keywords
Malaria, Plasmodium Falciparum, Thin Blood Smear, NIH, COCO

Article informations

Corresponding author: fwilm@mira.vision

1. 背景

M疟疾 是一种由属原虫寄生虫引起的热带疾病，这种寄生虫感染红细胞，主要通过雌性按蚊的叮咬传播。在人类中，该疾病主要与四个物种相关：*P. falciparum* ，*P. vivax* ，*P. malariae* 和 *P. ovale* 。近年来，第五种物种 i. e. ，*P. knowlesi* 的传染也在增加。在这五个物种中，*P. falciparum* 和 *P. vivax* 是最普遍的，而 *P. falciparum* 导致了大部分与疟疾相关的死亡 (World Health Organization, 2024) 。

大多数疟疾感染报告发生在热带和亚热带地区，这些感染影响了在医疗资源有限的低收入国家的人口。尽管现代治疗方法可以有效治愈疟疾，但早期诊断仍然至关重要，检测延迟是疟疾相关死亡率的一个主要原因 (Sultani et al., 2022) 。在疟疾的寄生虫学诊断中，通常进行厚血和薄血涂片图像的显微镜检查。除了识别 *Plasmodium* 物种外，光学显微镜还可以进行寄生虫定量和监测治疗反应。因此，通常较分子测试更受青睐 (World Health Organization, 2024) 。尽管如此，血涂片图像的寄生虫学评估需要较高的专业水平，而训练有素的人员在低资源国家或农村地区可能较为稀缺 (Poostchi et al., 2018) 。

最近，基于机器学习的数字化血液涂片图像分析方法在寄生虫定量方面展示了良好的效果 (Poostchi et al., 2018) 。然而，这些方法通常依赖于大规模、标注完善的数据集进行有效训练，因此公开可用的资源尤为珍贵。大多数现有研究集中在将单个细胞贴块分类为感染或未感染 (Kassim et al., 2020) ，这一过程需要事先提取出单个细胞的裁片。在密集分布的血液涂片图像中，这一步骤可能很具有挑战性，并限制了这类方法在实际诊断流程中的应用，在这些流程中，直接定位和准确量化感染细胞是必不可少的。与基于贴块的分类方法相比，目标检测架构需要具有详细实例级别标注的数据集，通常是以标记边界框的形式。然而，获取如此详细的标注工作量大且耗时，限制了其可用性。NIH 数据集包含 965 张图像，是最大的公开可用的 *P. falciparum* 检测资源之一。然而，其中只有 165 张图像附有基于多边形的详细标注，而剩下的 800 则仅限于点标注，标记了细胞中心。这种稀少性限制了其用于训练深度学习对象检测模型的适用性，这些模型通常需要边界框标注。

在这项工作中，我们提出了 NIH 数据集的修订版本，具有增强的注释。利用 Cellpose 框架 (Pachitariu and Stringer, 2022) 和手动标签修正，我们将原始点注释转换为边界框标签，这更适合于目标检测。为了验证修

订数据集的质量，我们训练了一个 Faster R-CNN（Ren et al., 2015）进行寄生虫检测，在感染细胞识别上获得了高达 0.88 的 F1 得分。更新的注释集通过 GitHub 公开提供：`https://github.com/MIRA-Vision-Microscopy/malaria-thin-smear-coco`

在我们的实验中，我们为 NIH 数据集生成了新的边界框注释，该数据集包含 *P. falciparum* 的吉姆萨染色的薄血涂片图像。我们通过训练一个基于深度学习的物体检测器来识别三种细胞类型：未感染的红细胞、感染的红细胞和白细胞，以此对这些注释进行了技术验证。

1.1 数据详情

NIH 数据集（Kassim et al., 2020）是一个来自孟加拉国吉大港医学院医院的薄涂片疟疾图像数据集，由美国马里兰州贝塞斯达的国家医学图书馆发布。它包括来自 193 名患者（148 名感染和 45 名未感染）的姬姆萨染色的薄血涂片图像，每位患者有五张图像。每张图像是使用安装在显微镜上的智能手机摄像头拍摄的，分辨率为 $5\,312 \times 2\,988$（宽 × 高）像素。注释涵盖三类：未感染的红细胞、感染的红细胞和白细胞。从 965 张总图像中，165 张包括详细的基于多边形的注释，而其余的 800 张仅提供标记细胞中心的点注释。Table 1 对这些子集进行了总结，以下简称为 NIH_{polys} 和 NIH_{points}。Figures 1a and 1b 显示了兴趣区域的示例，带有轮廓和点注释，对应于 NIH_{polys} 和 NIH_{points} 子集。

Table 1: 数据集子集概述。NIH_{polys} 包含详细的基于多边形的标签，而 NIH_{points} 用指示细胞中心的点标注。MIRA_{boxes} 为 NIH_{points} 数据集提供了修订的标签，带有详细的边界框注释。

	NIH_{polys}	NIH_{points}	MIRA_{boxes}
patients	33	160	160
no. of images	165	800	800
annotations	contours	points	boxes
no. of annotations			
non-infected	33 071	155 640	155 201
infected	1 142	6 810	6 805
white blood cell	51	220	220
ambiguous	-	-	19 592

为了使 NIH 数据集能用于训练目标检测模型，我们将点注释转换为详细的边界框注释。为此，我们首先使用 Cellpose 2（Pachitariu and Stringer, 2022）检测细胞实例，这是一个设计用于稳健、通用分割的开源框架。Cellpose 使用超过 70 000 个细胞的多样化数据集进行训练，提供了在广泛的细胞类型和成像模态上的强大性能，使其非常适合用于分割吉姆萨染色的血涂片图像。

在细胞实例分割之后，我们通过叠加原始点注释为检测到的细胞分配标签。如果一个点注释落在一个预测的边界框内，则该框被分配对应的细胞类别。然而，Cellpose 偶尔会检测到在原始数据集中未注释的细胞。这些通常是在视野边缘部分可见的细胞。在更新的注释集中，这些检测被标记为模糊的。Figure 2 显示了在视野边缘有模糊细胞的示例。总体而言，更新后的注释包括了 19 592 个模糊细胞，占了原始 NIH_{points} 子集的大约 10 %。

由于依赖于平均细胞大小，Cellpose 有时会将较大的细胞，特别是白细胞，分割成多个实例。为了解决这个问题，我们手动审查并合并了这些碎片化的检测。此外，Cellpose 有时会错误地将伪影或血小板分类为细胞。这些误检在手动后处理过程中也被去除。Figure 1c 显示了在边界框检测后具有代表性的重要区域，其中模糊细胞以橙色突出显示，而 Table 1 的最后一列总结了这一标注修订后细胞实例的数量。

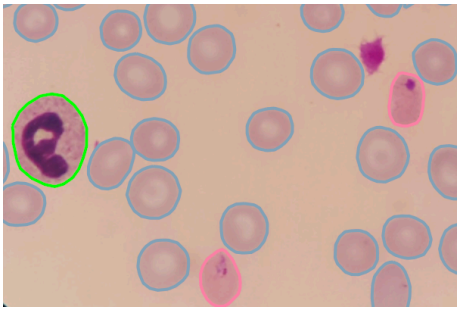
为了验证修订后的注释，我们训练了一个 Faster R-CNN 模型（Ren et al., 2015）来检测三种细胞类别：未感染的红细胞、感染的红细胞和白细胞。我们通过在 NIH_{polys} 或修订后的 MIRA_{boxes} 子集上训练模型，并分别评估其在另一个子集上的检测性能，进行了交叉验证实验。

我们采用了一个 Faster R-CNN 模型（Ren et al., 2015），基于 ResNet34（He et al., 2016）骨干网络，并在 ImageNet（Russakovsky et al., 2015）上进行了预训练。数据集被划分为 70 % 用于训练和 30 % 用于验证。在 NIH_{polys} 数据集上，该模型使用余弦退火学习率调度在前 50 个 epoch 进行线性预热，并使用最大学习率为 10^{-4} 的条件下训练 1 000 个 epoch。对于 MIRA_{boxes} 数据集，训练时间缩短到 200 个 epoch，以匹配数据子集近五倍的大小。为优化，采用 Adam 优化器和标准 Faster R-CNN 损失函数。从原始 $5\,312 \times 2\,988$ 像素图像中采样了 $1\,280 \times 960$ 像素的训练补丁。选择这个分辨率是为了匹配显微图像的典型 4:3 长宽比，同时确保每个补丁包含足够的细胞数量以进行有效训练。然后将补丁按 2 倍缩小到最终大小为 640×480 像素，以使批量大小达到 32 而不超过内存限制。为解决类别不平衡问题，我们应用了一个自定义补丁采样策略，对包含代表性不足类别（如白细胞）的区域进行过采样。通过在验证集上的 mean average precision (mAP) 进行监控，并且根据最佳验证 mAP 选择最终模型。

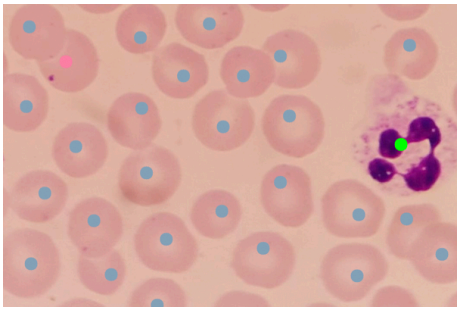
对于全分辨率的 $5\,312 \times 2\,988$ 像素图像推断，我们使用了 SAHI 框架（Akyon et al., 2021, 2022），该框架执行滑动窗口预测，并应用 non-maximum suppression (NMS) 来消除重叠块中的重复检测。作为后处理步骤，我们移除了所有面积小于 2 500 像素或大于 140 000 像素的预测边界框。这些阈值是根据原始 NIH 数据集中观察到的最小和最大注释尺寸确定的。

训练是在 NVIDIA A100 GPU 上进行的。实验使用 torchvision Faster R-CNN 模型实现，采用 PyTorch Lightning（Falcon and The PyTorch Lightning team, 2019）以简化训练，并使用 Hydra（Yadan, 2019）进行配置管理。

(a) 从 NIH$_{polys}$ 中采样



(b) 从 NIH$_{points}$ 中抽样



(c) 边界框标注

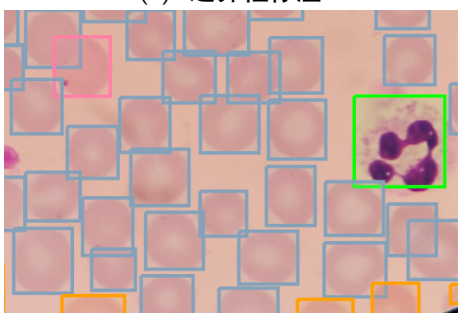


Figure 1: National Institutes of Health (NIH) 数据集提供的不同标注类型。(a)：轮廓标注，(b)：仅点标注，(c) 使用 Cellpose (Pachitariu and Stringer, 2022) 创建的边界框标注。蓝色：未感染的红细胞，粉红色：感染细胞，绿色：白细胞，橙色：不确定细胞。

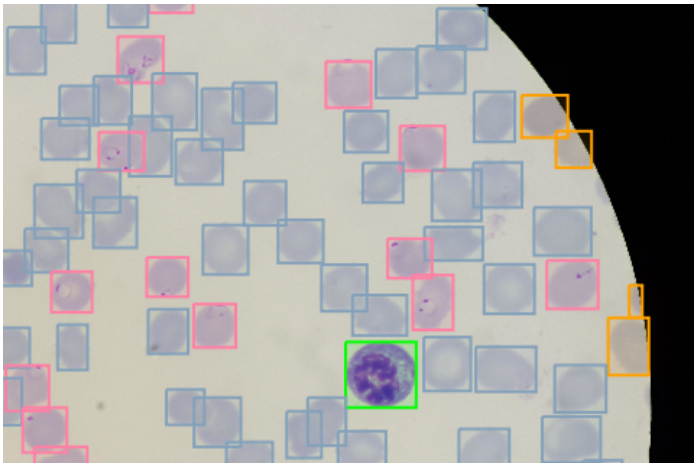


Figure 2: 在标签清理过程中，视野边界处的未标注细胞被标记为不确定 (橙色)。蓝色：未感染的红细胞，粉红色：感染的细胞，绿色：白细胞。

1.2 评估

为了进行评估，我们从实例级别的混淆矩阵中计算了按类别划分的 F1 分数。被 Cellpose 检测到但未被人工注释者标记的细胞（i. e. ，模糊细胞）被排除在评估之外。未被模型检测到的注释细胞被认为是 false negatives due to detection failure (FN_{det})，而未注释且未标记为模糊的模型预测被认为是 false positives due to detection failure (FP_{det})。类别 c 的类别 F1 分数计算如下：

$$F1(c) = 2 \cdot \frac{\text{Prec}(c) \cdot \text{Rec}(c)}{\text{Prec}(c) + \text{Rec}(c)} \quad \text{, with} \tag{1}$$

$$\begin{aligned} \text{Prec}(c) &= \frac{TP(c)}{TP(c) + FP_{\text{cls}}(c) + FP_{\text{det}}(c)} \\ &= \frac{M_{cc}}{\sum_{i=1}^{N+1} M_{ic}} \quad \text{, and} \end{aligned} \tag{2}$$

$$\begin{aligned} \text{Rec}(c) &= \frac{TP(c)}{TP(c) + FN_{\text{cls}}(c) + FN_{\text{det}}(c)} \\ &= \frac{M_{cc}}{\sum_{i=1}^{N+1} M_{ci}} \quad . \end{aligned} \tag{3}$$

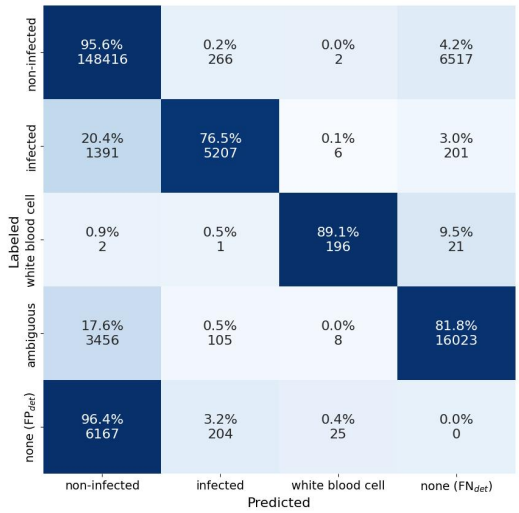
这里，M_{ij} 表示混淆矩阵 i 行 j 列的元素 i. e. ，即被标记为类别 i 并预测为类别 j 的单元数。N 是单元类别的数量，$(N+1)$ 行和列分别表示假阳性 (FP_{det}) 和假阴性 (FN_{det}) 的检测。

1.3 结果

Figure 3a 展示了在 NIH$_{polys}$ 子集上训练并在 MIRA$_{boxes}$ 子集上评估的 Faster R-CNN 模型的混淆矩阵，反之亦然。结果以行正则化的百分比形式显示，并附有绝对单元格计数。

总体上，当在 MIRA$_{boxes}$ 上训练并在 NIH$_{polys}$ 子集上评估模型时，模型表现得更好，这反映在混淆矩阵

(a) 训练：NIH$_{polys}$，测试：MIRA$_{boxes}$



(b) 训练：MIRA$_{boxes}$，测试：NIH$_{polys}$

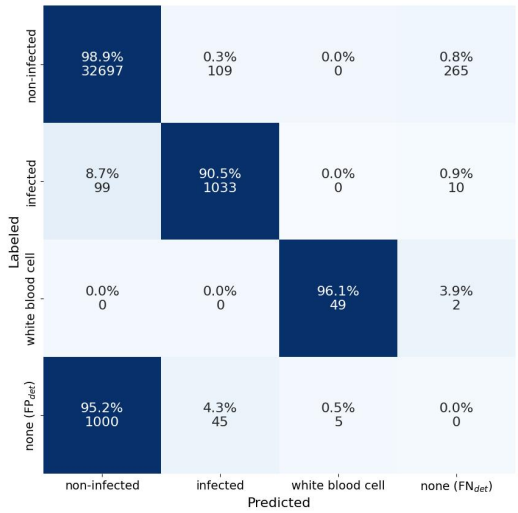


Figure 3: 是 NIH 子集上 Faster R-CNN 预测的混淆矩阵。每个矩阵显示行归一化的百分比以及绝对单元格计数。最后一行表示由模型检测但未在数据集中标注的 false positives (FPs) 、i. e. 个单元实例。最后一列表示 false negatives (FNs) 、i. e. 个未被模型检测到的标注单元实例。

中非对角线项的比例较低。当在 NIH$_{polys}$ 上训练并在 MIRA$_{boxes}$ 上测试时，感染细胞被误分类为非感染细胞的比例（$> 20\,\%$）相对较高，这表明疟疾检测的召回率降低。此外，模型未能检测到 $81.8\,\%$ 被标记为模糊的细胞。对这些细胞的深入检查显示，模糊细胞通常位于视野边缘附近，在这些地方注释应用不一致。具体来说，这些边缘细胞在 NIH$_{polys}$ 和 NIH$_{points}$ 数据集中经常未标注。这表明可能存在标记偏差，这在 Fig. 4 中得到了进一步的说明，其中白色箭头指示了未标注但清晰可见的细胞。

?? 总结了检测性能，报告为根据 Eqs. (1) to (3) 从混淆矩阵计算得出的精度、召回率和 F1 分数。对于每个数据集，训练使用三个不同的随机种子重复进行，我们报告平均性能为均值 ± 标准差（$\mu \pm \sigma$）。

Table 2: 在 NIH$_{polys}$ 子集上训练并在 MIRA$_{boxes}$ 子集上评估的检测模型的逐类 F1 分数（$\mu \pm \sigma$），反之亦然。

	NIH$_{polys}$ → MIRA$_{boxes}$	MIRA$_{boxes}$ → NIH$_{polys}$
Precision		
non-infected cells	0.96 ± 0.01	0.97 ± 0.00
infected cells	0.91 ± 0.01	0.86 ± 0.01
white blood cells	0.90 ± 0.04	0.88 ± 0.03
Recall		
non-infected cells	0.97 ± 0.01	0.99 ± 0.00
infected cells	0.77 ± 0.01	0.91 ± 0.01
white blood cells	0.92 ± 0.03	0.96 ± 0.00
F1 score		
non-infected cells	0.96 ± 0.01	0.98 ± 0.00
infected cells	0.84 ± 0.01	0.88 ± 0.00
white blood cells	0.91 ± 0.04	0.92 ± 0.02

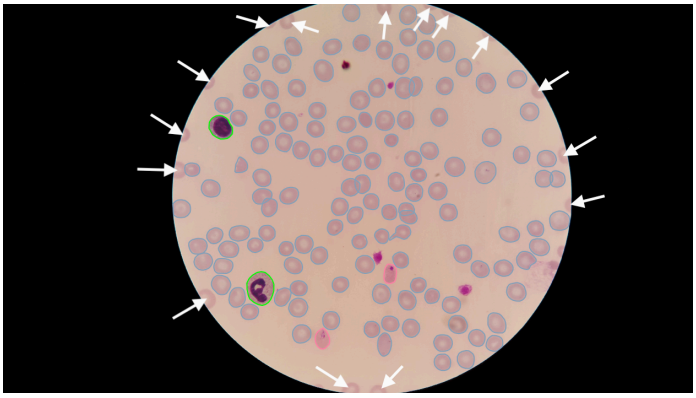
结果表明在检测非感染红细胞和白细胞方面具有高性能，F1 分数高于 $90\,\%$。感染细胞在使用 NIH$_{polys}$ 进行训练时的平均 F1 分数为 0.84，在使用 MIRA$_{boxes}$ 进行训练时为 0.88，这表明性能良好但相对较低。重复的训练运行显示出低变异性，表现结果的标准偏差较低。

性能指标再次证明了在 MIRA$_{boxes}$ 上训练的模型具有优越的性能。这尤其表现在感染细胞的召回率方面，训练在 NIH$_{polys}$ 上的平均值为 0.77，而训练在 MIRA$_{boxes}$ 上的平均值为 0.91。这一观察结果可能归因于标记一致性的差异，也可能归因于注释实例的数量较多（$6\,810$ 对比 $1\,142$），这为模型提供了更多样化的训练样本。

本研究提出了一个修订版的 NIH 疟疾数据集，其中包含以 COCO 格式提供的实例级标注，促进了基于深度学习的对象检测模型用于自动检测感染细胞的发展。我们通过训练一个 Faster R-CNN 来验证这些标注，以检测感染和未感染的红细胞以及白细胞，在检测感染细胞时实现了高达 0.88 的 F1 分数。对于训练有素的显微镜专家，World Health Organization (WHO) 的指导方针建议感染疟疾样本的最低召回率为 0.90 (World Health Organization, 2009)，我们的系统在细胞层面上达到了这个标准。因此，该系统达到了诊断环境中所需的最低能力水平。然而，我们对模糊细胞的分析揭示了原始注释中的不一致之处，尤其是在图像边缘处，病理学家未对细胞进行标注。然而，很难判断这些细胞是被简单忽略还是故意未标注，因为在部分可见的细胞上可能无法进行可靠的疟疾诊断。这引发了对生物医学数据集中的基准数据质量更广泛的担忧，可能是由于标注精度和时间投入之间的权衡所造成的。据我们所知，原始数据集由单一专家进行标注，这可能引入了相当大的标注偏差。未来的工作可以通过增加多位专家共同标注共识的额外手动注释回合来解决这一问题，或为部分可见的细胞引入一个单独的类别。为了评估机器学习模型的性能，我们建议在评估中排除这些细胞。

尽管部分标注数据存在挑战，我们的结果显示，通

(a) 从 NIH$_{polys}$ 中抽样



(b) 从 NIH$_{points}$ 抽样

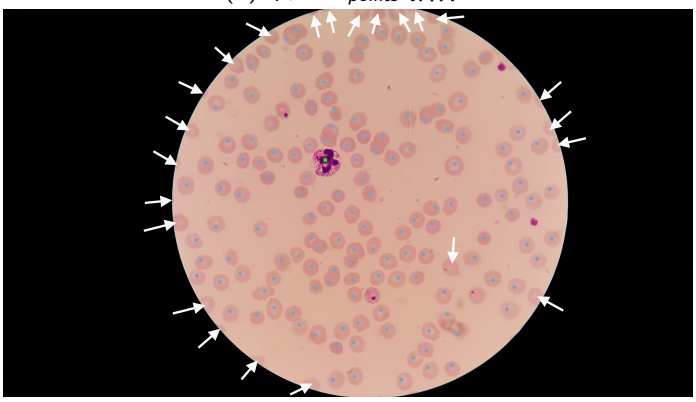


Figure 4: 来自 NIH 子集的代表性样本，白色箭头指示视野边界未注释的细胞：(a) 来自多边形子集的样本，具有详细的轮廓注释，(b) 来自点子集的样本，细胞中心有斑点注释。

过现有工具如 Cellpose 进行注释转换，随后进行有针对性的人工整理，可以产生足够高质量的训练数据，以支持模型的稳健性能。这一发现对于资源有限的环境尤为重要，因为在这些环境中，详细的注释代价高昂或难以实现。

除了标注一致性之外，我们还观察到 NIH 数据集的两个子集之间模型性能的差异，这可能是由于用于训练的标注实例数量不同造成的。这凸显了数据集大小和多样性对于学习微妙形态特征（例如环状阶段寄生虫的存在）的重要性。此外，我们的初步数据集评估显示健康细胞和感染细胞的类别高度不平衡。我们通过采用定制的补片采样策略在一定程度上弥补了这一点，但在未来的工作中，可以整合专门的增强策略或类平衡损失函数。

总体而言，我们的工作提供了一个增强的数据集和一个稳健的管道，用于显微镜下的寄生虫检测，支持对自动化疟疾诊断的进一步研究。

References

Fatih Cagatay Akyon, Cemil Cengiz, Sinan Onur Altinuc, Devrim Cavusoglu, Kadir Sahin, and Ogulcan Eryuksel. SAHI: A lightweight vision library for performing large scale object detection and instance segmentation. *Zenodo*, November 2021.

Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 966–970, 2022.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. URL `https://github.com/Lightning-AI/lightning`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Yasmin M Kassim, Kannappan Palaniappan, Feng Yang, Mahdieh Poostchi, Nila Palaniappan, Richard J Maude, Sameer Antani, and Stefan Jaeger. Clustering-based dual deep learning architecture for detecting red blood cells in malaria diagnostic smears. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1735–1746, 2020.

Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12): 1634–1641, 2022.

Mahdieh Poostchi, Kamolrat Silamut, Richard J. Maude, Stefan Jaeger, and George Thoma. Image analysis and machine learning for detecting malaria. *Translational Research*, 194:36–55, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20655–20664. IEEE, 2022.

World Health Organization. *Malaria microscopy quality assurance manual: v.1*. WHO Regional Office for the Western Pacific, 2009.

World Health Organization. *WHO guidelines for malaria*. WHO, 2024.

Omry Yadan. Hydra - a framework for elegantly configuring complex applications, 2019. URL `https://github.com/facebookresearch/hydra`.