arXiv:2305.02208v1 [cs.CV] 3 May 2023

DocLangID：改进小样本训练，以识别历史文档的语言

Furkan Simsek
furkan.si
msek@student.hpi.de 哈索·普拉特纳研究所 德国波茨坦

Brian
Pfitzmannbrian.pfitzmann@student.hpi.de哈索·普拉特纳研究所德国波茨坦

Hendrik Raetz
hendrik.raetz@hpi.de
哈索·普拉特纳研究所，德国

Jona Otholt
jona.otholt@hpi.de
哈索·普拉特纳研究所
德国波茨坦

Haojin Yang
haojin.yang@hpi.de
哈索·普拉特纳研究所
德国波茨坦

Christoph Meinel
christoph.meinel@hpi.de
哈索·普拉特纳研究所
德国波茨坦

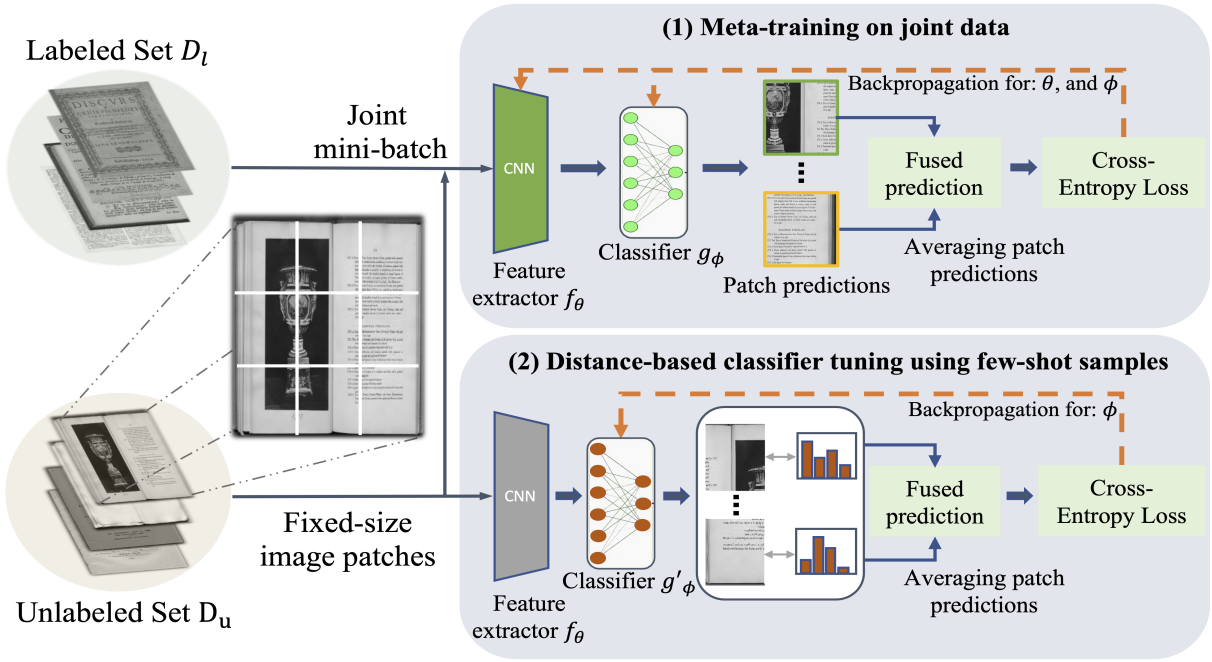


图 1：*DocLangID*概览：这种两阶段方法首先在一个联合数据集上训练卷积神经网络，该数据集包含来自源领域数据集的大量带标签样本 D_l 以及来自未标注目标领域数据集的少量人工标注样本 D_u。在第二阶段，我们固定特征提取器的参数，并用一个新的基于距离的分类器替换原有分类器。这个新分类器在与前一阶段相同的少量样本上进行微调，利用特征提取器对源领域的知识来帮助适应目标领域。

摘要

语言识别是指识别文档中书写文本所用语言的任务。这一信息至关重要，因为它可用于支持对文档词汇和上下文的分析。近年来，监督学习方法推动了语言识别任务的发展。然而，这些方法通常需要大量带标签的数据集，而且往往需要针对文档或场景图像等不同领域提供相应的数据集。在这项工作中，我们提出了*DocLangID*，一种迁移一种用于识别无标签历史文档语言的学习方法。我们首先利用来自不同但相关的历史文档领域的有标签数据。其次，我们实现了一种基于距离的少样本学习方法，使卷积神经网络能够适应无标签数据集中的新语言。通过从无标签图像中抽取少量人工标注的样本，我们的特征提取器增强了对历史文档中新数据分布的适应能力。我们展示了，只需重复使用相同的少样本示例，就能对图像无标签集合有效地微调此类模型。我们在 10 种语言上展示了这项工作，

HIP'23，*2023*年*8*月*25*日至*26*日，圣何塞，加利福尼亚州
2023年。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$
15.00 https://doi.org/10.1145/nnnnnnn.nnnnnnn

这些语言大多使用拉丁字母。我们在历史文档上的实验表明，组合方法提升了语言识别性能，在无标签数据集的四种未见语言上达到了 74% 的识别准确率。

CCS概念

• **应用计算** → **文档分析**；光学字符识别；• **计算方法** → 兴趣点和显著区域检测；**基于分类的监督学习**；**迁移学习**；**神经网络**。

关键词

语言识别、小样本训练、卷积神经网络。

ACM 参考格式：

Furkan Simsek、Brian Pfitzmann、Hendrik Raetz、Jona Otholt、Haojin Yang 和 Christoph Meinel。2023年。DocLangID：改进小样本训练，以识别历史文档的语言。载于第七届历史文档成像与处理国际研讨会论文集（*HIP'23*）。ACM，纽约，美国纽约州，6页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

语言识别是图像分类的一个子领域，旨在识别图像中印刷或手写文本所使用的语言。一种令人振奋但研究较少的图像类型是历史文档的扫描件。例如，多个欧洲图书馆编纂的、内容丰富且多样的文本类历史文档合集，可通过 IMPACT [13] 数据集获取。此类数据集的创建，以及为较小子集配套制作的真实标注数据，通常与改进历史文档的分析和数字化处理这一目标有关。尽管许多历史文档主要包含文本，例如书籍、文章或通讯中的文本，但也有一些文档的主要内容是大幅图片。艺术史文献就是一个具体例子。这些文献是关于艺术作品及其创作者——艺术家——的信息的书面记录。它们包括手稿、信件、日记及其他书面材料，可帮助人们了解特定艺术作品的创作过程、历史和意义。

作为讨论较为广泛的计算机视觉任务之一，光学字符识别（OCR）方法旨在将图像中的文本转换为数字格式。文档的语言是字符识别过程中一项有价值的上下文信息。例如，如果事先知道文档的语言，就可以选择特定语言的 OCR 引擎。这有助于减少字符识别中的歧义，进而可能提高识别质量和速度。

近年来，深度学习方法在多种文档分析问题上取得了显著的性能提升，例如图像分类和 OCR [9, 11, 21]。然而，文档输入图像的多个方面仍会削弱上述方法的识别性能。图像分辨率低、字体大小小，以及输入图像本身或扫描过程引入的噪声，都是常见问题 [9, 11, 21]，同时也是艺术史文献的常见特征。已有多种监督学习方法被提出用于解决这些问题并提升识别性能 [11, 21]。这些方法通常会在神经网络训练过程中使用大量带标签的数据。然而，如今的大多数数据都来自多样且非结构化的来源，因此缺乏额外的真实标注信息。因此，要准确地赋予数据有意义的标签，往往需要具备上下文理解能力和专业知识，这通常既耗时，也难以应用于大量数据。

受这一问题的启发，本文提出 *DocLangID*，一种用于解决语言识别领域适应问题的两阶段训练方法。给定两个图像域，其中一个有标签，另一个无标签，任务是将一个在有标签域上训练的模型调整为能够识别无标签域中的语言。在这种设定下，两个域中的语言不必重叠。

DocLangID 是一种简单而有效的少样本学习方法，采用基于距离的分类器，如 Chen 等人所示[2]，，以改进无标签域中的语言识别。我们只需使用少量人工标注的样本，就能实现这一改进。

我们的贡献可总结如下：

(1) 我们在监督式设置下实现了对艺术类历史文档的高水平语言识别，这一成果此前在拉丁字母语言中尚未实现。(2) 我们提出了一种简单而有效的领域自适应方法，利用大量标注数据和少样本示例，将模型的知识迁移到新的历史文档领域。(3) 我们分析了图像数据训练中常用的图像块提取流程，并评估其对我们方法的影响。(4) 我们实现并评估了主要*DocLangID* 方法的监督式变体，并在 Github [1]上发布了我们的研究成果。

2 相关工作

语言识别近年来，已有多种基于机器学习的方法致力于解决从图像中进行语言识别的问题 [1, 16, 19]。Vatsal 等人 [19] 开发了一种卷积神经网络（CNN），通过识别图像中是否存在附加符号字符来确定语言。Shah 等人 [16] 采用层次化的多个卷积神经网络，识别属于两个不同语族的六种印度语言。他们首先使用一个卷积神经网络作为二元分类模型，以确定正确的语族；随后使用另一个卷积神经网络识别具体语言。此外，Chakraborty 等人 [1] 提出了一种集成方法，该方法基于 RGB 颜色模型，从场景文本中生成五种不同的样本，并结合每个样本的预测结果，以提升识别性能。此外，识别图像中所使用的文字体系也是一个重要的相关领域，通常被视为文本理解系统不可或缺的初步步骤 [10, 12]。Lu 等人 [10]提出结合局部卷积神经网络以利用局部图像特征，以及全局卷积神经网络以提取整体图像特征。最近，Mahajan

[1]https://github.com/caesarea38/DocLangID

等人 [12] 采用卷积神经网络开展词级文字体系识别，并将其应用于包含印度文字文本的风景图像。

少样本学习 少样本学习是一种机器学习类型，其通过使用极少量（带标签的）样本训练模型，旨在提高模型对未见数据的泛化能力。为克服大规模未标注数据集成本高昂的标注过程，研究人员提出了多种适用于神经网络的少样本学习算法，下面将对此进行简要回顾。Finn 等人 [3] 提出了一种元学习方法，旨在通过多种任务训练模型，从而能够利用少量带标签的样本和少量梯度更新步骤，快速适应新任务。此外，Rusu 等人 [15]提出了一种潜在嵌入优化方法，该方法在模型参数的嵌入空间中执行基于梯度的元学习。此外，基于距离度量学习的方法[4, 14] 通过"学习比较"来解决少样本学习问题。这些模型经过训练，能够通过确定未见输入图像与训练期间少量带标签样本之间的相似度，对该图像进行分类。训练期间常用的距离度量包括余弦相似度 [20] 和到类别均值表示的欧几里得距离 [18]。然而，以往的方法通常应用于 CIFAR [8], 等标准基准数据集并在其上进行评估，而忽略了艺术史文献中的现实约束和挑战，例如噪声或颜色失真对文本可理解性的影响。据我们所知，尚无先前研究探索通过监督学习和领域自适应进行语言识别的可行性，尤其是在包含拉丁字母语言的艺术史文献这一特定背景下。

3 方法

在本研究中，我们探讨如何将基于带标签的源领域数据集（D_l）训练的语言识别模型，适配到一个相关但未标注的目标领域数据集（D_u）的问题。我们将其定义为一个多类别分类问题，其中语言代表不同的类别。我们考虑用于 D_l 的标签集 Y_l 和用于 D_u 的标签集 Y_u，这两个标签集可能包含共同的语言。

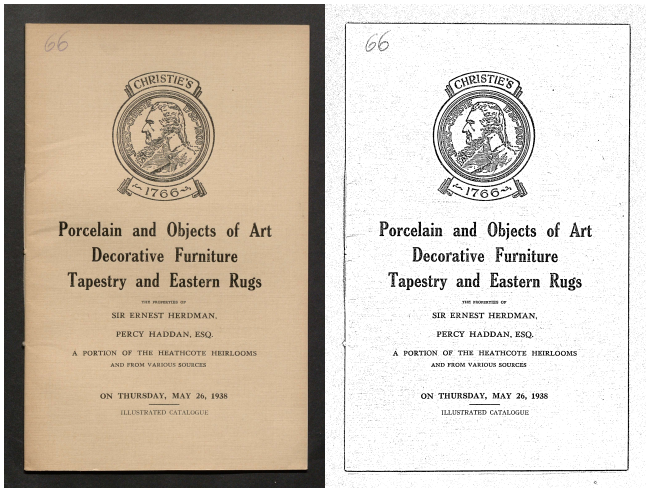
3.1 数据集

在我们的整个研究过程中，我们使用了两个不同的数据集，下面将对它们进行简要介绍。

我们的第一个数据集是 IMPACT [13] 数据集的一个子集，其中包含来自欧洲主要图书馆的超过五十万张文本文档图像。该数据集包含各种图像，涵盖多种语言的报纸、书籍、小册子和打字笔记等材料，包括荷兰语、法语、西班牙语、捷克语和保加利亚语等 [13]。

我们从*Wildenstein PlattnerInstitute(WPI)* 2公开在线可访问的资源中创建了第二个数据集。该资源包含 21,000 多份扫描版历史拍卖目录，以及来自 17 至 20 世纪各单页的 100 万至 200 万张未标注图像。我们从中随机选取一部分图像并下载，然后为其手动标注正确的语言。最终，我们为每种语言（英语、法语、德语和荷兰语）各整理出 1,000 张已标注图像，构成了我们的目标领域数据集。

[2]https://wpi.art



(a) 原始图像 **(b) 二值化图像版本**

图 2：WPI 数据集示例图像，未经过（a）预处理及经过预处理（b）的位置。

我们的 IMPACT 标注子集恰好包含 3,000 张图像，每种语言 500 张。训练时，我们使用其中 2,100 张作为训练数据（每种语言 350 张），其余 900 张作为评估数据。在少样本学习部分，我们为 WPI 的每种语言手动选取了 50 个样本。其余 WPI 数据用于评估。

在训练模型之前，我们对 IMPACT [13] 和 WPI 数据集的原始图像进行两步预处理。由于旧的历史文档往往包含噪声，且颜色也会失真，这些预处理步骤旨在降低噪声，并使实际文本在图像中更加突出。因此，我们首先将所有图像转换为灰度格式，然后应用自适应二值化阈值处理，最终获得二值化图像。这一系列图像转换的输出示例见图 2。

3.2 特征学习

在整个研究过程中，我们使用未采用预训练权重的 ResNet-18 卷积神经网络 [5] 作为特征提取器。首先，选择它是因为 ResNet-18 是一种稳健且广为人知的视觉任务模型，近年来在 ImageNet [5] 等具有挑战性的图像分类基准测试中展现出了出色的性能。其次，从实际应用角度来看，ResNet-18 是一种相对小型的神经网络，约有 1100 万个可学习参数。凭借其架构设计和较少的层数，ResNet-18 能够实现更快的训练和推理。

3.3 图块提取与预测

给定一张训练期间的输入图像，*DocLangID* 预测过程中的一个关键步骤是将该图像划分为固定大小的图像块。与单张大图像相比，多个较小的图像块能够实现更高效的并行处理，同时通过固定的图像块大小保留局部空间信息，从而有效地

确保每个图像块对最终预测作出同等贡献。我们使用（高度，宽度）尺寸表示图像块大小。例如，在训练期间，我们将每个图像块的大小设为（256，256），图像大小设为（1024，1024），由此最多可得到 16 个图像块。训练期间使用的图像块数量可通过超参数进行控制。从输入图像的左上部分开始，我们依次提取（高度，宽度）个像素以获得图像块。提取图像块后，我们对所有图像块进行一次前向传播。随后，我们对每个图像块计算得到的概率分布取平均，以获得融合后的概率分布。最后，我们选择融合分布中概率最高的类别，作为整张图像的预测类别。预测计算所涉及的组件和完整过程如图 1 所示。

3.4 我们的两阶段训练方法

我们的主要 *DocLangID* 方法分为两个依次进行的训练阶段。在第一个元训练阶段，我们训练 CNN 特征提取器 f_θ（由其网络参数 θ 参数化）以及一个分类头 g_ϕ。这一部分被设计为一个简单的线性层，并将特征投影到一个输出向量 $\hat{y} = [c_1, c_2, ..., c_k]$，其中 k 等于 Y_l 和 Y_u 中语言的总数，即 $k = |Y_l| + |Y_u|$。随后，我们对线性层的输出应用 softmax 函数，并使用标准交叉熵损失函数来获得最终损失。

对于我们方法的第二阶段，我们首先冻结网络参数 θ，这在迁移学习中是常见做法。其次，我们舍弃线性分类头 g_ϕ，并重新训练一个新的分类器 $g_{\hat{\phi}}$。此外，我们现在遵循 [2, 14]，为分类器采用基于距离的设计。对于输入图像，我们计算其与分类器权重矩阵之间的余弦距离。然后，针对应用于余弦距离输出和少样本样本标签的交叉熵损失，优化新分类器的参数。在我们的问题中采用这种设计的直觉源于目标领域数据集丰富多样的图像特征。即使一组图像属于同一类别，它们在外观方面也可能存在巨大差异（是否存在标题页、图像中是否包含大幅图片以及不同的字体大小），从而可能导致较高且因此具有问题的类内方差。这种基于距离的分类器设计的优势在于，它使用距离值来衡量图像与特定类别的接近程度，从而显式地降低类内方差。

考虑到*DocLangID*的各种变体，我们使用*ResNet-FewShot* 一词来指仅使用 WPI 少样本样本训练整个模型（特征提取器 + 分类器）。此外，我们将*ResNet-Meta* 称为在*DocLangID* 方法第一阶段后评估的模型；在这一阶段，训练时仅使用联合数据集（见图 1 中的第一个训练状态）。联合数据集指合并后的数据集，其中包含来自 IMPACT 子集的带标签样本和来自 WPI 子集的少样本样本。据此，我们的关键见解是，即使只有少量带标签样本，也足以让模型更好地理解未见过的领域；而来自相关领域的大量带标签样本则能在训练过程中加深这种理解。

3.5 Tesseract OCR

如引言中所述，我们采用开源 Tesseract OCR 算法 [17] 和 Python 库 fasttext-langdetect [6, 7] 作为基线，与*DocLangID*进行比较。

Tesseract 能够识别 100 多种语言，支持 30 多种文字系统，并可同时识别多种语言，从而允许运行时输入多种语言。使用 OCR 算法来改进 OCR，乍看之下似乎与通过语言识别改进 OCR 的目标相矛盾。然而，这种方法在 IMPACT 和 WPI 数据集上都取得了令人瞩目的结果。此外，语言识别并不要求完美提取文本，只要能正确识别大多数单词，就足以确定语言。

我们使用数据集中与源图像具有相同文字体系的语言作为输入，对源图像运行一次或多次 Tesseract OCR。例如，对于 IMPACT 数据集，我们可以在第一次运行时使用拉丁字母语言，在第二次运行时使用西里尔字母语言。多次运行 Tesseract 可以确保至少有一次输出采用正确的文字体系，从而避免因使用错误的文字体系而导致语言识别错误。

我们比较 Tesseract 对每个识别词给出的 OCR 结果置信度值，预期使用正确文字体系对应的输入语言时，置信度最高。因此，我们选用置信度最高的 Tesseract 输出结果。

我们采用 Python 库 fasttext-langdetect，使用由 Facebook 开发、基于维基百科及其他来源训练的模型，从提取的文本中识别 176 种语言。

然而，这种方法的主要缺点是识别过程耗时较长。根据文本长度的不同，Tesseract 的执行时间最长可达 30 秒，而多次执行会进一步延长这一时间。因此，该方法不适用于包含数万张图像的大规模数据集，主要用作推理时间方面的对比基线。

4 实验评估

下面我们展示并讨论在 IMPACT 和 WPI 数据集的评估子集上的结果。关于实现细节，请参阅附录。

4.1 识别质量

WPI 目标领域的核心分类指标值见表 1。可以看出，我们的组合方法比单一的少样本学习变体*ResNet-FewShot* 高出 20 多个百分点。图 3 展示了更多分类结果。

表 1：WPI 评估数据集的结果（50 个少样本样本）

方法	准确率	精确率	召回率	F1
ResNet-FewShot	0.52	0.49	0.54	0.51
ResNet-Meta	0.64	0.68	0.69	0.68
DocLangID（我们的方法）	**0.74**	**0.77**	**0.74**	**0.74**

各目标领域语言的指标结果。同样，我们使用每个目标领域 50 个少样本样本评估了我们的 *DocLangID* 方法。

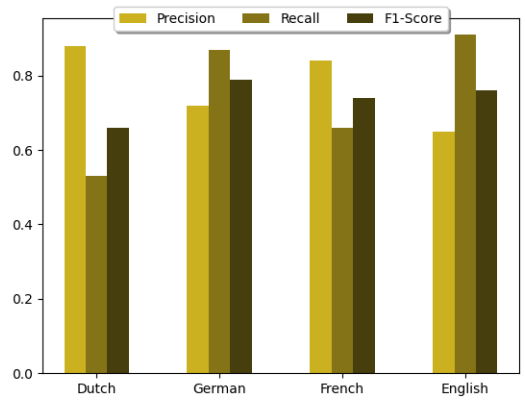


图 3：每种 WPI 语言的分类指标结果（50 个少样本样本）。总体而言，尽管语言之间存在细微差异，*DocLangID* 的分类指标结果依然很高。

针对领域语言而言。总体来看，尽管语言类别之间存在一些细微差异，我们的模型仍取得了较高的分类指标。不过，我们观察到，模型区分德语和荷兰语语言类别的难度略高，导致荷兰语的召回率较低。这可能与德语和荷兰语之间较高的语言相似性有关。在识别准确率方面，Tesseract 通常取得了非常好的结果（> 90%），总体而言优于本次比较中的*DocLangID*。不过，当然，要达到这一性能，我们需要事先输入正确的语言集合。影响*DocLangID* 性能的一个关键因素，是两阶段训练中使用的少样本样本数量。图 4 展示了 50 个少样本样本的不同子集对识别性能的影响。总体而言，可以看出，随着少样本样本数量增加，识别性能也有所提升，并且*DocLangID* 在所有配置下均优于其他变体。有趣的是，即使每种语言使用的样本数量较少，也能显著提升所有变体的识别性能。我们方法的另一个重要组成部分是

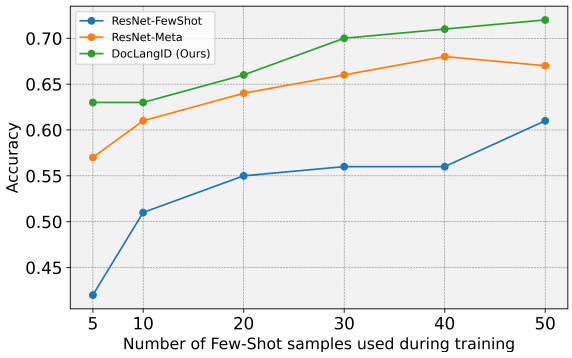


图 4：不同数量的少样本样本对识别性能的影响。

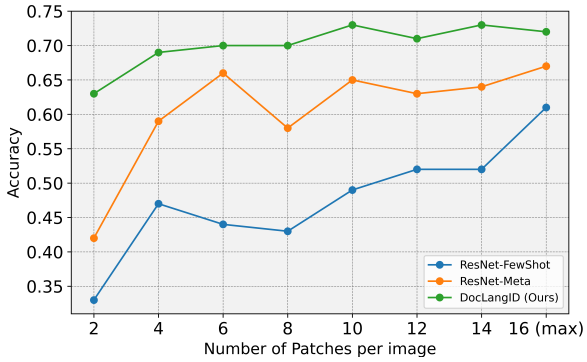


图5：不同数量的图像块对识别性能的影响（使用50个少样本样本的准确率得分）。

训练过程中考虑的图块数量。图5展示了不同图块数量对识别性能的影响。仅使用四个图块，我们的方法就实现了近70%的识别准确率。尽管增加图块数量（≥ 8）仍能提高准确率，但性能提升似乎并不明显。相比之下，较少的图块数量（例如两个）会显著降低*ResNet-FewShot* 和*ResNet-Meta*的性能。在其他实验中，我们采用了最大图块数量16，这一设置使所有方法都取得了最佳结果。尤其对于*ResNet-FewShot*而言，该数量带来了显著的识别性能提升。一种可能的解释是，与*ResNet-Meta*和*ResNet-FewShot*相比，*ResNet-FewShot*仅使用少样本样本进行训练，因此训练数据集规模要小得多。虽然这可能使模型在目标领域上的学习过程更加直接，但也可能增加过拟合的风险。相比之下，我们发现，尽管图像结构和语言类别的差异会造成领域差异，但IMPACT数据集中的大量标注样本仍能对模型的适应能力产生积极影响。此外，需要注意的是，图块的最大数量在很大程度上取决于单个图块的尺寸，而本文中的数值属于一种设计选择。因此，使用较少数量但尺寸更大的图块，也可能产生类似的观察结果。此外，我们仅展示了所提出方法在两个领域之间应对领域偏移的能力，对于领域自适应任务而言，这可能被认为是领域数量较少。然而，我们考虑了源领域和目标领域之间两个基本的领域特征：语言和文档类型。两个领域的语言集合不仅互不重叠（荷兰语除外），其文档类型也存在较大差异。我们的WPI目标领域主要包含拍卖目录的图像，这些图像通常只展示拍卖物品的大幅图片，而文字内容相对较少且字号较小。相比之下，我们的IMPACT [13] 源领域则包含各种文档的文本丰富型图像，例如文章、报纸和小册子。在此背景下，我们发现，源领域本身涵盖了足够多样化的特征，可以用于证明我们的方法在适应具有挑战性的WPI领域时的有效性。

4.2 识别速度

除了预测的正确性，另一个重要的定性方面是计算预测所需的时间。在将*DocLangID* 与使用 Tesseract 的基线进行比较时，我们评估了推理阶段的平均处理时间，如表 2 所示。数值是对 100 张原始图像取平均得出的，其中每种语言所占比例相同。该表区分了分类所需的处理时间和第 3 节所述的预处理步骤。

Tesseract 方法所需的时间几乎是 IMPACT 的十倍。当使用 WPI 数据集并扣除预处理时间后，这一趋势变得更加明显。在这种情况下，我们的 *DocLangID* 模型只需几毫秒即可计算图像的语言分类，而 Tesseract 仍需要超过一秒钟。一种对这一巨大差异的解释是，尽管两个数据集中的图像平均文件大小相近，但与 WPI 图像相比，许多 IMPACT 图像包含更复杂的细节、文本和图案。这反过来可能导致更高的图像复杂度，从而增加分析单张图像的计算成本。因此，采用像*DocLangID* 这样更高效的基于深度学习的解决方案，有助于克服这一潜在的 OCR 瓶颈，同时仍能提供相当准确的语言信息。随后可以利用这些信息选择特定语言的 OCR 引擎，从而有效减少耗时较长的 OCR 运行次数。有关两个数据集以及本研究所涉及全部语言的示例图像，请参阅附录。

表 2：两个数据集所有语言类别的平均推理时间（以秒计），分别为包含（w）和不包含（w/o）输入预处理时间。

	w预处理时间		无 预处理时间	
数据集	Tesseract	DocLangID	Tesseract	DocLangID
IMPACT	21.2	**2.95**	19.9	**1.18**
WPI	7.5	**0.63**	6.5	**0.003**

5 结论与未来工作

在这项工作中，我们提出并全面评估了*Do-cLangID*，这是一种两阶段深度学习方法，结合元学习与少样本学习，用于历史文档的语言识别。我们通过研究主*DocLangID* 方法的不同变体，为这项任务建立了基线；这些变体不使用源领域中大量的已标注样本。对识别质量的评估表明，我们结合的方法在 WPI 目标领域上取得了最高的预测准确率。我们还比较了*DocLangID* 的推理速度与 Tesseract 以及 Python 库 fasttext-langdetect 执行语言识别的速度。对于 IMPACT 和 WPI 数据集，*Do-cLangID* 的推理时间比 Tesseract 快十倍以上。

未来的工作可以包括利用文本定位技术改进图块提取流程，并结合无监督学习技术，从大量未标注数据中学习更好的图像表征。

参考文献

[1] Neelotpal Chakraborty、Soumyadeep Kundu、Sayantan Paul、Ayatullah Faruk Mollah、Subhadip Basu 和 Ram Sarkar。2020年。基于卷积神经网络分类器集成方法，从多语言场景文本图像中识别语言。《环境智能与人性化计算期刊》12（2020年），7997–8008。[2] Wei-Yu Chen、Yen-Cheng Liu、Zsolt Kira、Yu-Chiang Frank Wang 和 Jia-Bin Huang。2020年。深入探究少样本分类。载于《国际学习表征会议》。[3] Chelsea Finn、P. Abbeel 和 Sergey Levine。2017年。用于深度网络快速适应的模型无关元学习。载于《国际机器学习会议》。[4] Spyros Gidaris 和 Nikos Komodakis。2018 年。不遗忘的动态少样本视觉学习。载于 2018 年 I EEE/CVF计算机视觉与模式识别会议。 4367 4375 10 1109 2018 00459 –。https://doi.org/。/ [5] 2016 CVPR.。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。。用于图像识别的深度残差学习。载于2016年IEEE计算机视觉与模式识别会议（CVPR）。770–778。https://doi.org/10.1109/CVPR.2016.90[6] Armand Joulin、Edouard Grave、Piotr Bojanowski、Matthijs Douze、Hérve Jégou 和 Tomas Mikolov。2016年。FastText.zip：压缩文本分类模型。ArXiv预印本 arXiv:1612.03651（2016年）。[7] Armand Joulin、Edouard Grave、Piotr Bojanowski 和 Tomas Mikolov。2017年。高效文本分类技巧集。载于《欧洲计算语言学协会分会第15届会议论文集：第2卷，短篇论文》。计算语言学协会，427–431。[8] Alex Krizhevsky。2009年。从微型图像中学习多层特征。[9] Ankit Lat 和 C. V. Jawahar。2018年。通过超分辨率提升OCR准确率。载于2018年第24届国际模式识别会议（ICPR）。3162– 3167 10 1109 2018 8545609 。https://doi.org/。/ICPR.。 [10] Liqiong Lu、Yaohua Yi、Faliang Huang、Kaili Wang 和 Qi Wang。2019年。结合局部卷积神经网络和全局卷积神经网络，对自然场景图像进行文字体系识别。《IEEE开放获取》7（2019年），52669–52679。[11] Pengyuan Lyu、Minghui Liao、Cong Yao、Wenhao Wu 和 Xiang Bai。2018年。Mask TextSpotter：用于任意形状文本定位的端到端可训练神经网络。《IEEE模式分析与机器智能汇刊》43（2018年），532–548。[12] Shilpa Mahajan 和 Rajneesh Rani。2022年。利用卷积神经网络增强风景图像的词级文字体系识别。《亚洲及低资源语言信息处理汇刊》21（2022年），1–29。[13] Christos Papadopoulos、Stefan Pletschacher、Christian Clausner 和 Apostolos Antonacopoulos。2013年。历史文档图像IMPACT数据集。载于《第二届历史文档成像与处理国际研讨会论文集》（美国哥伦比亚特区华盛顿）（HIP '13）。计算机协会，美国纽约州纽约，123–130。https://doi.org/10.1145/ 2501115.2501130 [14] Hang Qi、Matthew Brown 和 David G. Lowe。2018年。基于印刻权重的低样本学习。载于2018年IEEE/CVF计算机视觉与模式识别会议。 5822 5830 10 1109 2018 00610 –。https://doi.org/。/ [15] CVPR.。Andrei A. Rusu、Dushyant Rao、Jakub Sygnowski、Oriol Vinyals、Razvan Pascanu、Simon Osindero 和 Raia Hadsell。2019年。基于潜在嵌入优化的元学习。载于《国际学习表征会议》。https: //openreview.net/forum?id = 7 BJgklhAcK Sarathi Shah 和 M. V. Joshi。。 [16] 2021 文档语言分类：深度学习方法的层次模型。载于《图像与模式计算机分析》，Nicolas Tsapatsoulis、Andreas Panayides、Theo Theocharides、Andreas Lanitis、Constantinos Pattichis 和 Mario Vento（编）。施普林格国际出版，Cham，372–381。[17] R. Smith。2007年。Tesseract OCR引擎概述。载于《第九届国际文档分析与识别会议》（ICDAR 2007），第2卷。629–633。https://doi.org/ 10 1109 2007 4376991 。/ICDAR.。Jake Snell、Kevin [18] Swersky 和 Richard Zemel。2017年。用于少样本学习的原型网络。载于《第31届神经信息处理系统国际会议论文集》（美国加利福尼亚州长滩）（NIPS'17）。Curran Associates Inc.，美国纽约州雷德胡克，4080–4090。[19] Shubham Vatsal、Nikhil Arora、Gopi Ramena、Sukumar Moharana、Dhruval Jain、Naresh Purre 和 Rachit S Munjal。2020年。利用附加符号字符，对图像中的文本进行端侧语言识别。载于《国际计算机视觉与图像处理会议》。[20] Oriol Vinyals、Charles Blundell、Timothy Lillicrap、Koray Kavukcuoglu 和 Daan Wierstra。2016年。用于单样本学习的匹配网络。载于《第30届神经信息处理系统国际会议论文集》（西班牙巴塞罗那）（NIPS'16）。Curran Associates Inc.，美国纽约州雷德胡克，3637–3645。[21] Fei Yin、Yi-Chao Wu、Xu-Yao Zhang 和 Cheng-Lin Liu。2017年。利用滑动卷积字符模型进行场景文本识别。ArXiv abs/1709.01727（2017年）。