arXiv:2305.02593v1 [cs.CV] 4 May 2023

如何为单一书写者微调选择预训练手写识别模型

Vittorio Pippi[1 [0009−0001−7365−6348]]、Silvia Cascianelli[1 [0000−0001−7885−6050]]、Christopher Kermorvant[2 3 [0000−0002−7508−4080]] 和 Rita Cucchiara[1 [0000−0002−2239−283X]]

[1] 摩德纳-雷焦艾米利亚大学，意大利摩德纳[2] TEKLIA，法国巴黎[3] LITIS，鲁昂-诺曼底大学´，法国{vittorio.pippi、silvia.cascianelli、rita.cucchiara}@unimore.it kermorvant@teklia.com

摘要。 近年来，基于深度学习的手写文本识别（HTR）取得了长足进展，使模型在大型基准数据集中的现代和历史手稿上均表现出色。然而，当应用于具有特殊特征的手稿时，例如语言、纸张载体、墨水和作者笔迹，这些模型难以达到同等性能。对于历史档案馆中保存的珍贵但规模较小的文献集而言，这一问题尤为突出，因为获取足够的标注训练数据成本高昂，有时甚至无法实现。为克服这一挑战，一种可行方案是先在大型数据集上预训练手写文本识别模型，再用小规模的单一作者文集对其进行微调。本文考虑了大型真实基准数据集，以及利用风格化手写文本生成模型生成的合成数据集。通过广泛的实验分析，并考量微调行的数量，我们定量指出了此类数据中哪些特征最为关键，从而获得能够有效转录小型文集手稿的手写文本识别模型，而所需的真实微调行数仅为五行。

关键词： 文档合成 · 历史文档分析 · 手写识别 · 合成数据。

1 引言

数字化正成为文化和工业领域中文档高效管理、保存与增值的关键步骤。因此，文档分析（DA）技术，尤其是用于应对手写文本这一复杂场景的技术，正受到研究界的广泛关注。基于大型公开数据集训练的最先进手写文本识别（HTR）模型，在处理与训练数据特征相似的文档时能够取得令人瞩目的成果。然而，当目标领域的数据与训练数据差异过大时，这些模型的性能并不理想。在这方面，

许多档案馆保存的小型但珍贵的历史手稿藏品，给现代手写文本识别模型带来了挑战。事实上，此类档案馆往往仅藏有少量由某位特定但重要的作者书写的样本文页，其视觉特征和语言学特征都颇为独特。因此，要让这些文档也能实现高性能的手写文本识别，采用相应策略至关重要，这有助于高效地将此类文档数字化。应对此类情形的一种常见方法，是先在大型数据集（真实数据或合成语数据集）上预训练手写文本识别模型，再使用目标领域中数量有限的真实数据进行微调。这一策略还有望实现对整个单一作者文集的高质量按需转录，为数字图书馆和档案馆的用户提供一项颇具价值的服务。具体而言，图书馆可以存储预训练的手写文本识别模型，用户只需提供几行文本的标注（例如，$5-15$），即可请求转录自己感兴趣的文集。此时，可以根据文集的特征（例如语言、年代、作者身份、风格）选择最合适的预训练模型；这些特征可能以元数据的形式提供。随后，只需较短时间，便可使用用户提供的标注对模型进行微调。之后，微调后的模型即可低错误率地转录整本文集。这种与文集互动的方式也有助于提升数字档案用户的整体体验。值得注意的是，迄今为止，对交互式转录增强的探索主要集中在语言模型优化方面 [36]。本研究旨在探索一种更全面的方法，同时考虑语言和外观。

需要指出的是，文献中已有研究尝试使用合成数据对手写文本识别模型进行预训练 [25,2,30,52,32]。合成数据越接近真实数据，这些策略就越有效 [12]。在这一方向上，手写文本生成（HTG）技术正逐渐兴起 [5,7,23,31]，尤其是风格化 HTG 技术，它们有望生成具备特定领域手写文本识别所需特征的训练数据。事实上，风格化 HTG 模型可以根据少量风格示例图像，生成包含任意文本且采用指定手写风格的图像。这些模型通常包含一个编码器，用于提取作者特有的风格特征，以及一个生成器；生成器以风格特征和代表字符的内容标记为输入，从而生成具有指定风格和内容的文本图像。鉴于此，本文考虑采用预训练后再微调的方法，并使用自动生成的作者专属合成数据集进行训练；该数据集通过利用最先进的风格化 HTG 网络获得。此外，我们还评估了在现有多种语言、作者数量各异且年代不同的基准数据集上进行预训练的效果。通过这种方式，我们探究能否构建一种有效的流程，以交互式、按需识别单一作者文集中的手写文本。具体而言，我们提出了一套定量指南，同时考虑视觉和语言学方面的因素，以设计出最有效的流程，使手写文本识别模型只需使用目标手稿中的少至 5 行内容进行微调，就能以较低错误率转录特定手稿集。档案管理者可以利用这些关于如何选择最合适预训练模型的指南，也可以将其提供给用户，让用户更深入地参与转录过程。

2 相关工作

手写文本识别策略。由于在工业和文化领域的应用具有实际价值，手写文本识别一直是备受关注的研究课题。尽管如此，它仍是一项具有挑战性的任务。手写文本识别可以针对单个字符进行，这是惯用语种中一种常见的选择 [15]，单词 [48,6],或整行文本 [45,40]，段落和页面 [38,9,8,55,16]。对于非惯用语种，行级识别是最常用的方式之一，既可以独立使用，也可以作为页面级系统的一部分 [56,38,9] 目前最常用的基于学习的手写文本识别解决方案依赖多维长短期记忆网络（MD-LSTM） [26] ，或采用卷积网络与一维长短期记忆网络相结合的架构 [45,40,39,50,10] 来表示文本图像，并使用连接时序分类器（CTC）解码策略输出转录文本 [26,8]。除采用循环模型的方法外，研究者也提出了用于手写文本识别的全卷积网络 [56,18], ， 以及 [30,32,53] 基于 Transformer 编码器-解码器架构的解决方案 [49]。最后值得一提的是，可以利用显式语言模型或词典来优化转录结果。不过，转录图像所用的语言越规范（即没有错误、罕见词和专有名词）且代表性越充分，这种策略就越有效。因此，语言模型并非总是适用，尤其是在处理历史手稿时。

手写文本生成（HTG）策略。手写文本生成（HTG）是一个日益热门的研究领域，旨在生成逼真的手写文本图像。在本文研究的风格化任务中，目标是仅凭书写者的少量风格示例图像，生成具有该书写者特征的手写文本图像，以模仿其风格 [5,23,31]。早期的手写文本生成（HTG）方法，无论是否采用风格化，都取得了令人瞩目的成果，但代价是需要大量人工干预和手工设计特征 [51,27]。相比之下，近期提出的基于学习的方法完全自动化。通常，这些策略涉及使用生成对抗网络（GAN） [24]。对于非风格化的手写文本生成（HTG），这些网络可以是无条件的 [1,23]。而对于风格化的手写文本生成（HTG），所用的 GAN 则以编码器从手写风格示例图像中提取的风格特征为条件。请注意，风格示例可以是行图像 [20], 、少量词语图像 [31,5],或单张图像 [37]。还值得一提的是，近期提出的一种基于编码器—解码器生成式 Transformer 的方法 [5]。

用于手写文本识别的合成数据。 训练数据不足是手写文本识别面临的一项主要挑战，尤其是在单一作者文档或具有独特特征的古代手稿中。一种可能的解决策略是进行数据增强，包括通用的颜色修改和几何变形 [50,40,54]，或精心设计、旨在匹配目标数据特征的图像修改 [14]。另一种常用策略是先在大型数据集上预训练手写文本识别模型，再使用目标数据对其进行微调 [25,28,47], ； 事实证明，对于历史手稿而言，这种方法比数据增强更有益 [2]。预训练数据集可以是真实数据（例如，

公开可用的基准数据集），也可以是合成数据，通常通过对以书法字体渲染的文本图像进行变换而获得 [44,30]。为生成外观更逼真的文本图像，近期一些研究利用手写文本生成模型（无论是否经过风格化）生成合成数据，用于训练手写文本识别模型并提升其在真实数据上的性能。例如，在 [46], 中，作者采用一种基于贝叶斯程序学习的组合方法，生成加密语料中的符号，再将这些符号组合成逼真的文本行，用于训练手写文本识别模型转录历史加密手稿。研究者还在 [29], 中探讨了使用多种书写风格的生成文本行训练手写文本识别模型的益处，作者应用了风格化的手写文本生成模型来构建预训练数据集。最后，针对单一作者的情形， [12], 中的作者展示了这样一种预训练方式的优势：使用忠实仿照真实数据的合成数据预训练手写文本识别模型，而非使用风格多样的通用图像。不过，他们的方法严重依赖人工工作，才能获得这类高质量合成数据。鉴于此，本文聚焦于使用全自动手写文本生成模型生成的合成数据。

3 提出的方法

在本研究中，我们探索了一种流程，通过利用大型数据集进行预训练，并使用目标文集中的少量样本进行微调，来获得具有独特特征（在手写体、语言和纸张载体方面）的手稿集的高质量行级转录。具体而言，我们考虑在真实数据集和通过少样本风格化手写文本生成模型生成的合成数据集上进行预训练，以更好地反映目标数据的特征。为了构建合成数据集，我们需要少量词语图像（本研究中为 15 张，与 [31,5]中相同），这些图像可以方便地从感兴趣的小型馆藏中的数字化手稿中获取。此外，我们还需要指定要以所需风格呈现的文本。在本研究中，我们考虑数字图书馆和档案馆中的两种典型情形：一种是仅知道目标手稿的语言，另一种是还掌握作者信息。如果已知感兴趣馆藏的作者，并且存在该作者撰写的其他已转录文本，我们建议让 HTG 模型生成这些文本。如果仅知道语言（或者没有该作者撰写的其他文本），则让 HTG 模型生成与目标文集语言相同的文本。请注意，在这两种情况下，HTG 模型都会输出手写词语图像，随后我们将这些图像组合成长度不一的行。下文将介绍用于转录的 HTR 网络，以及用于生成合成预训练数据的 HTG 网络。完整流程的概览如图1所示。

3.1 手写文本识别模型

多年来，将卷积神经网络与循环神经网络结合用于手写文本识别一直是标准选择，许多现有的转录服务也采用了这类模型，因为它们效率很高。在这项工作中，我们

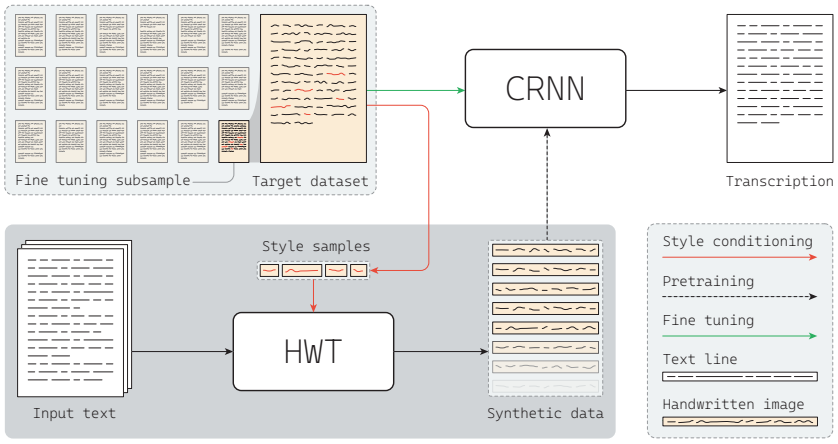

图 1. 我们从特定集合的手写行生成合成数据的流程概览。生成过程根据给定文本，并以目标数据集中的少量风格样本为条件，渲染出手写行图像。随后，使用合成数据集对卷积循环神经网络进行预训练。

考虑采用包含一维 LSTM 的模型，因为事实证明，这类模型的性能可与 MD-LSTM 相媲美，甚至更优 [26] ，而且训练速度更快 [40]。

特别是，在本研究中，我们考虑了 [45] 中提出的一种方法变体（以下简称**卷积循环神经网络**）。该架构的卷积部分包含七个卷积块。对于前六个卷积块，我们采用与 VGG-11 网络相同的架构，不同之处在于，在最后两个最大池化层中使用矩形池化，以更好地反映文本行图像的纵横比。第七个卷积块具有一个 2×2核。在所采用的变体中，卷积组件采用了 [17,12,11], 中提出的可变形卷积 [19], ，从而提升了性能。最后一个卷积层的特征图是一个 $2 \times W \times 512$ 张量，其中W 取决于输入文本行图像的宽度。沿通道维度压缩该张量，可得到由 W 个特征向量组成的序列，每个特征向量包含 1024 个元素，并将其输入架构的循环部分。该部分由两个双向 LSTM 层组成，每层具有 512 个隐藏单元，两层之间由一个 Dropout 概率为 0.5 的层隔开。循环部分输出序列中每个特征向量包含字符集中的各个字符的概率。

按照 HTR 的惯例，模型经过训练以优化 CTC 损失，因此模型字符集中特别包含一个空白 字符。请注意，为了实现跨语言适应性，我们不会将任何语言模型与 HTR网络结合使用。

3.2 手写文本生成方法

风格化的手写文本生成模型能够高效地生成大量采用目标作者手写体风格的合成文本图像，这些图像随后可用于训练

针对目标作者定制的手写文本识别模型。在这项工作中，我们基于近期提出的基于 Transformer 的少样本风格化手写文本生成模型 [5], ，即手写体变换器（**HWT**）。

所采用的手写文本生成方法使用卷积-Transformer 编码器-解码器架构。首先，将手写体样本输入编码器中的卷积特征提取器（即 ResNet18），其输出经由多层多头 Transformer 编码器（包含 3 层和 8 个注意力头）处理，并通过自注意力机制融入长程依赖。随后，将得到的风格向量用作多层多头 Transformer 解码器（与编码器一样，包含 3 层和 8 个注意力头）中的键和值；该解码器与表示待渲染单词中字符的向量执行交叉注意力。然后，向得到的向量添加正态高斯噪声，以获得一定的多样性，再将其输入由四个残差块和一个 tanh 激活函数组成的卷积解码器，最终输出风格化的单词图像。

HWT 模型通过优化一个多项损失函数，与其他模块一同训练。具体而言，我们遵循对抗范式，采用铰链对抗损失 [33] ，并将 HWT 与卷积判别器共同训练。此外，为了确保生成的词语图像清晰可读，我们加入一项 CTC 损失：让手写文本识别模型预测生成图像中的文本内容。该手写文本识别模型的设计也受到了 [45]中所提出架构的启发。最后，为了促使 HWT 忠实呈现所需风格，我们使用了另外两项损失。一项是卷积分类器的交叉熵损失，该分类器旨在根据 HWT 训练集中的书写者对生成图像进行分类。另一项是循环一致性损失，其值为 HWT 编码器部分对真实图像和生成图像进行编码后所得编码之差的 l_1-范数。

4 实验

本节介绍我们的实验分析。首先，我们进一步说明所采用的手写文本识别和手写文本生成模型的实现细节。然后，我们介绍所考虑的小型单作者目标数据集、真实基准数据集（其详情见表 1，部分样本见图 2），以及合成预训练数据集的构建流程。最后，我们说明所采用的评估方案，并讨论所得结果。

4.1 实现细节

本研究的实验，包括手写文本生成和手写文本识别部分，均在一块 NVIDIA RTX 2080 Ti GPU 上进行。

卷积循环神经网络。 为了训练卷积循环神经网络模型，我们将所有图像缩放至高度为 60 像素，同时保持原始纵横比，然后将其归一化至 -1 到 1 之间。此外，在预训练时，我们会进行以下数据增强：将图像亮度调整为随机选取的

表1。 所考虑的行级数据集特征。

	训练行	字符集	时期	语言	作者
华盛顿 [22]	526	68	1755	英语	One[1]
圣加仑 [21]	468	49	约 890-900 年	拉丁语	One
Leopardi [12]	1303	76	1818-1832	意大利语	One
IAM [35]	6482	79	现代	英语	Many
ICFHR16 [42]	8367	88	1470-1805	德语	Many
Rodrigo [43]	9000	105	1545	西班牙语	One
ICFHR14 [41]	9198	93	约1760—1832年	英语	One
RIMES [3]	10188	95	现代	法语	Many
NorHand [34]	19653	111	1820-1940	挪威语	Many
LAM [13]	19830	89	1691-1750	意大利语	One
华盛顿合成语	23121	78	-	英语	One
圣加仑合成数据集	70494	66	-	拉丁语	One
用于 Leopardi 的合成语	89068	113	-	意大利语	One

1 - 少量行文字出自另一位作者。

0.5 到 5 之间的倍数，对比度调整为随机选取的 0.1 到 10 之间的倍数，饱和度调整为随机选取的 0 到 5 之间的倍数，色调调整为随机选取的 -0.1 到 0.1 之间的倍数。此外，我们还会应用高斯模糊，卷积核大小设为 5，标准差在 0.1 到 2 之间随机选取。最后，我们会从以下几种几何变换中随机选择一种：随机旋转（角度在 $-1°$ 到 $1°$ 之间）、仿射变换（随机旋转角度在 $-1°$ 到 $1°$ 之间，随机剪切幅度在 $-50°$ 到 $30°$ 之间）以及随机单应性变换。预训练时的批次大小为 16，微调和从头开始训练时的批次大小为 8。所有实验的学习率均为 10^{-4}。我们使用 Adam 作为优化器训练所提出的模型，参数为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$；如果模型在验证集上的字符错误率停滞不前，则通过学习率调度器将学习率降低 10%。对于验证集上的字符错误率，我们将模型的耐心值设为 20 个周期。请注意，微调时通常在第二个周期内就能达到最佳字符错误率，耗时不到一小时。

HWT。我们在词级设置下使用 IAM 数据集训练 HWT 模型（参见 4.3）。所有图像样本均为灰度图像，其高度调整为 32 像素，同时保持相同的纵横比，并在 -1 和 1 之间进行归一化。我们使用与原论文 [5] 中相同的设置训练模型 7000 个 epoch。HWT 模型生成单词图像，其中字符平均占据 16 像素。因此，我们将不同图像拼接起来，生成单行文本，并使单词之间的间距为 16 像素。

4.2 目标数据集

Leopardi [12]。 Leopardi数据集包含少量19世纪早期的信件，这些信件由浪漫主义语言学家、作家和诗人Giacomo Leopardi用意大利语书写。该数据集包含1303条训练行、596条验证行，以及

587条测试行。所有样本均为用墨水书写在古纸上的文档RGB扫描件。

华盛顿 [22]。 乔治·华盛顿数据集包含 20 封由美国总统乔治·华盛顿及其一位合作者于 1755年撰写的英文信件。该数据集分为 526 行训练行、65 行验证行和 65 行测试行。所有样本均为二值图像。

圣加仑 [21]。 圣加仑数据集包含一份由单一作者于 9th 世纪末以拉丁语撰写的手写历史手稿，共 60 页。该数据集分为 468 行训练行、235 行验证行和 707 行测试行。所有样本均为二值图像。

4.3 真实数据集预训练

Rodrigo [43]。 Rodrigo数据集包含一位作者用西班牙语撰写的 853 页内容。这些页面来自一份名为“Historia de Espa˜na del ar¸cobispo Don Rodrigo”的手稿，写于 1545年。所有样本均为在古纸上用墨水书写的文档灰度扫描件。

NorHand [34]. NorHand数据集由4144页组成，主要是15位挪威作者于1820年至1940年间书写的日记和信件。训练集包含19653行。所有样本都是用墨水书写在泛黄纸张上的文档灰度扫描件。

LAM [13]. LAM数据集包含历史学家 Ludovico Antonio Muratori 于1691年至1750年间用意大利语书写的1171封信件，因此由于这一较长的时间跨度而呈现出一定程度的可变性。训练集包含19830行。所有样本均为使用墨水书写在古纸上的文档RGB扫描件。

ICFHR14 [41]。 ICFHR14 数据集包含 433 页关于法律和道德哲学的文档，这些文档由英国哲学家杰里米·边沁于1760年至1832年间撰写。训练集包含 9198 行。所有样本均为用墨水书写在泛黄纸张上的文档灰度扫描件。

ICFHR16 [42]。 ICFHR16数据集由Ratsprotokolle文集中400页文档组成，这些文档写于1470年至1805年，使用早期新高地德语。作者人数未知。训练集包含8367行。所有样本均为墨水书写在泛黄纸张上的文档灰度扫描件。

RIMES [3]。 RIMES数据库（Reconnaissance et Indexation de donn´ees Manuscrites et de fac simil´ES / 手写文档和传真件的识别与索引）由12723页扫描信件组成，这些信件由1300位不同的作者用法语书写。训练集包含10188行。所有样本均为二值图像。

IAM [35]。 IAM手写数据库3.0是一个现代文集，包含由657位不同作者用英语书写的1539页扫描页面。IAM 数据集包含不同的设置：我们使用词级设置训练手写文本生成网络，并使用行级设置预训练HTR网络。具体而言，我们使用339位不同作者提供的所有可用词语来训练手写文本生成网络



图 2. 所考虑的真实数据集中的示例行图像。这些数据集使用不同的语言，来自不同的时期，且作者数量各不相同。

在词级训练集中。对于行级设置，训练集包含 6482 行，验证集包含 976 行，测试集包含 2915 行。所有样本均为文档的灰度扫描件，这些文档以墨水书写在白纸上，并经过数字化清理。

4.4 合成数据

我们使用基于词级 IAM 数据集训练的 HWT 模型，为每个目标数据集生成专属的合成数据。具体而言，对于每个数据集，我们提取若干词语图像，并反复随机选择其中的 15 张作为风格示例。至于文本内容，我们分别使用 Giacomo Leopardi 的部分散文作为 Leopardi 数据集的文本，使用 George Washington 的部分日记作为 Washington 数据集的文本，以及一本中世纪拉丁语《圣经》作为.....

圣加仑数据集。这样一来，合成数据集的语言就更接近目标数据集的语言。下文将这些合成数据集称为 **HWT生成**。请注意，HTW 输出的词语中，字符的平均宽度为 16 像素。在实验中，我们还考虑了一种合成数据集变体，其中生成的图像会在宽度方向上调整大小，以匹配目标数据集的平均字符宽度。下文将所得数据集称为 **HWT生成+宽度调整** （宽度调整 - WA）。最后一种变体是生成与 HWT生成版本文本内容相同、但采用 IAM 数据集书写风格图像的合成数据集。下文将此变体称为**HWT生成+宽度调整+多样化风格**（多样化风格 - VS）。此外，对于 Leopardi 数据集，我们还利用了随数据集一同发布的合成文本行。这些文本行是使用手动构建的随机字体渲染而成，渲染文本与我们用于 HWT生成 Leopardi 的文本相同，该字体模仿了作者的手写体。下文将这一合成数据集称为**人工合成**。

4.5 评估方案

为了评估在目标数据集中只有少量行完成标注的情况下，预训练和微调策略的效果，我们逐步减少训练行的数量，分别采用 100%（作为参考）、50%、5%、2.5% 和 1.25% 的训练行进行微调。此外，作为额外的对比，我们使用相同数量的训练行从头开始训练。同时，我们还考虑将预训练模型直接迁移至目标数据集。转录性能采用常用的字符错误率（CER）和词错误率（WER）进行报告。

为了从语言学角度进一步深入了解训练数据的特征，我们计算了每个目标数据集和预训练数据集中的字符一元语法、二元语法和三元语法分布之间的KullbackLeiber散度。此外，我们报告了所考虑数据集的语言之间的词汇相似度 [4] 。该指标根据两种语言中的词语是否具有共同词源以及相似的发音和含义，量化了语言对之间的词汇相似度。

4.6 结果

作为基准实验，我们将预训练于真实数据和合成数据上的卷积循环神经网络模型进行直接迁移。结果见表 2。可以发现，平均而言，在合成数据上预训练的模型表现不如在真实数据上训练的模型，但在 Leopardi数据集上，人工合成数据集的情况除外。在表 3-5中，我们报告了使用真实数据集中的少量行进行微调的结果。这些实验旨在反映按需转录应用场景。作为参考，我们还报告了从头开始训练的卷积循环神经网络模型在相同行数数据上的结果。请注意，当用于训练的行数少于 230 行时，模型无法收敛，因此更凸显了预训练的必要性。

表 2. 所考虑的模型在真实数据集或通过不同方式获得的合成数据集（WA 表示“宽度调整”，VS 表示“多样化风格”）上预训练后，直接应用于所考虑的目标数据集测试集时的性能。

	Leopardi		**华盛顿**		**圣加仑**	
	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率
IAM	57.0	95.6	50.5	**88.7**	**49.9**	**97.7**
RIMES	68.0	97.2	**48.2**	96.2	51.2	98.5
Rodrigo	88.6	99.5	84.3	104.2	55.2	111.4
ICFHR14	75.0	105.6	83.6	128.8	83.7	106.9
ICFHR16	80.5	104.5	87.8	127.4	80.1	106.3
NorHand	49.1	95.4	64.9	101.4	76.0	114.4
LAM	**23.4**	**57.3**	78.8	103.8	78.0	99.5
人工合成	56.9	95.3	-	-	-	-
HWT生成	93.2	99.4	90.9	100.0	83.7	106.9
HWT生成+宽度调整	87.5	99.2	87.5	100.0	76.7	100.2
HWT生成+宽度调整+多样化风格	87.9	99.3	83.0	98.6	75.3	99.9

从结果中可以看出，尤其是在使用 1.25% 和 2.5% 的训练行进行微调时，性能差异最为明显。结果表明，对于每个目标数据集，都可以确定最佳的预训练数据集，这说明准确选择预训练数据是提升识别性能的关键。具体而言，在Leopardi数据集上，LAM数据集总体上最为适用（见表 3）；在华盛顿数据集上，ICFHR14 最有帮助（见表 4）；而在圣加仑数据集上，Rodrigo 最为适用（见表 5）。

请注意，在与目标数据集不同的数据集上对网络进行预训练会引入偏差，而这种偏差取决于所用数据集的某些特征。具体而言，预训练数据集与目标数据集在语言学和视觉方面越相似，这种偏差对最终性能的提升就越有帮助。下文将探讨影响这种性能的主要因素，旨在为选择最合适的预训练数据集提供一些指导。

外观。选择用于网络预训练的数据集时，需要考虑的一个方面是整体视觉外观（例如纸张载体、墨水颜色和粗细，以及字符宽度）。为分析相似性，我们参考图2，其中展示了各种数据集的一些示例。可以看出，LAM和Leopardi数据集中的样本看起来相似，这一点也反映在卷积循环神经网络在LAM上预训练、再在Leopardi上微调时所达到的性能上。相反，Leopardi和Rodrigo中的样本存在显著差异。因此，在后一个数据集上预训练，再将模型用于Leopardi上的手写文本识别，会导致性能不佳。圣加仑和华盛顿这另外两个数据集也有类似的情况。例如，圣加仑图像中的手写体较为规整，与Rodrigo中的手写体相似；Rodrigo是能带来最佳性能的数据集之一。此外，Washington数据集中的图像在视觉上与RIMES或IAM相似。事实上，这两个数据集正是进行

表3. 所考虑模型在使用真实数据集或以不同方式生成的合成数据集（WA 表示“宽度调整”，VS 表示“多样化风格”）进行预训练，并在 Leopardi 训练集的不同比例数据上微调后的性能。

	在以下数据上微调/训练									
	1.25%（15 行）		2.5%（32 行）		5%（65 行）		50%（652 行）		100%（1303 行）	
	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率
Leopardi	-	-	-	-	-	-	4.9	18.4	2.8	10.8
IAM	21.7	63.2	17.1	54.1	12.0	41.0	4.9	19.4	3.3	13.5
RIMES	25.3	68.7	18.2	54.6	13.1	42.3	4.3	15.7	2.7	10.5
Rodrigo	34.9	81.1	23.0	63.7	15.8	48.8	5.2	18.6	2.9	11.0
ICFHR14	23.6	67.7	16.6	53.6	11.7	40.7	4.0	15.9	2.7	10.8
ICFHR16	38.7	85.8	24.4	68.1	16.7	53.4	6.3	24.2	4.3	17.8
NorHand	21.0	63.1	15.2	50.3	11.5	40.2	**3.4**	13.2	**2.3**	8.7
LAM	**12.7**	**42.0**	**12.1**	**39.6**	**8.2**	**28.8**	3.5	**12.7**	**2.3**	**8.6**
人工合成	25.3	70.1	17.5	53.8	11.7	38.5	4.3	16.1	2.5	9.6
HWT生成	66.3	97.8	43.8	84.8	22.5	59.6	5.0	18.3	2.8	10.9
HWT生成+宽度调整	43.4	89.2	27.9	70.2	17.4	51.4	5.1	18.6	2.7	10.3
HWT生成+宽度调整+多样化风格	35.6	80.9	23.6	63.8	14.8	45.6	4.5	16.9	2.6	10.2

预训练能在Washington上取得最佳性能。从这些观察中，我们可以得出结论：视觉上的某些相似性有助于从预训练数据集向目标数据集进行迁移学习。不过，正如我们将在下文看到的，这只是需要考虑的诸多方面之一。另一个需要考虑的视觉方面是平均字符宽度。如第4.1节所述，HWT生成的图像高度为32像素，平均字符宽度为16像素。这个纵横比与目标数据集中的图像大不相同；目标数据集的平均字符宽度约为8像素。因此，我们利用有限的可用示例估算了一个形状因子，以缩小HWT生成图像的宽度，使其与目标数据集相匹配。表2-5中报告的、在宽度调整（WA）合成数据集变体上预训练的CRNN模型结果表明，与原始HWT生成版本相比，平均字符宽度较小且接近目标数据集字符宽度时，性能更好。

手写体。 结合表 2-5 中的结果以及表 1 中的数据集信息（尤其是作者数量和时间跨度），我们可以发现性能与相应数据集中不同书法风格之间存在相关性。尤其是后者，往往使数据集图像中的手写风格具有很高的多样性。如果预训练数据集包含多位作者书写的文本，网络就会接触到较大的方差，并具备处理不同书写风格和书法风格的能力。另一方面，如果使用仅包含一位作者手写内容的图像对网络进行预训练，则会降低方差，使网络专注于该作者的手写风格。例如，LAM数据集由于仅包含一位作者，方差较低。尽管如此，由于语言相似且历史时期相近，它与Leopardi数据集相似。因此，使用LAM进行预训练会引入一种偏差，使手写文本识别模型能够有效地泛化到Leopardi。一个例子

表4. 所考察模型在使用真实数据集或以不同方式获得的合成数据集（WA 指“宽度调整”，VS 指“多样化风格”）进行预训练，并使用华盛顿数据集训练集的不同比例进行微调时的性能。

	微调/训练于									
	1.25%（6 行）		2.5%（13 行）		5%（26 页）		50%（263 页）		100%（526 行）	
	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率
华盛顿	-	-	-	-	-	-	5.3	24.3	3.4	15.9
IAM	**18.8**	**52.7**	14.4	45.9	12.5	40.0	4.9	20.7	3.9	16.1
RIMES	27.1	78.9	20.8	65.2	16.8	55.7	4.6	20.7	3.7	17.1
Rodrigo	48.9	92.6	39.6	84.3	27.3	69.0	6.4	25.8	4.5	19.9
ICFHR14	26.1	64.6	**14.2**	**43.9**	**11.1**	**36.2**	**3.9**	**17.7**	**2.8**	**12.9**
ICFHR16	58.0	97.8	49.1	90.1	30.3	74.6	7.1	28.0	5.2	21.3
NorHand	31.4	76.7	21.7	61.2	16.1	52.9	5.8	23.7	3.7	15.3
LAM	37.2	81.7	27.6	72.6	20.7	60.7	5.9	23.1	4.8	20.3
HWT生成	54.9	93.8	43.4	84.7	28.8	70.8	5.3	20.9	3.7	16.5
HWT生成+宽度调整	47.4	90.9	34.0	79.7	27.9	70.8	5.5	22.1	3.6	17.3
HWT生成+宽度调整+多样化风格	31.3	77.7	23.9	63.4	19.1	56.9	4.8	20.9	3.3	16.1

表 5. 在真实数据集或通过不同方式获得的合成数据集（WA 表示“宽度调整”，VS 表示“多样化风格”）上进行预训练，并在圣加仑训练集的不同比例数据上进行微调时，所考虑模型的性能。

	微调/训练于									
	1.25%（5升）。		2.5%（11 行）		5%（23 行）		50%（234 行）		100%（468 行）	
	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率	字符错误率	词错误率
圣加仑	-	-	-	-	-	-	5.8	38.6	4.5	32.5
IAM	16.5	68.3	13.3	61.0	10.6	54.2	5.4	36.0	4.6	31.4
RIMES	28.2	94.2	19.7	79.7	14.3	66.8	6.5	39.9	5.8	36.9
Rodrigo	**14.4**	**66.4**	**11.3**	**58.7**	**8.8**	**50.8**	**5.3**	35.8	4.6	31.9
ICFHR14	20.4	77.8	16.6	70.5	12.1	54.2	**5.3**	35.4	**4.5**	**30.9**
ICFHR16	32.0	94.2	22.8	83.5	16.1	71.1	6.8	41.9	5.7	36.0
NorHand	27.0	87.9	19.5	75.0	12.8	61.3	**5.3**	**35.1**	4.6	31.5
LAM	20.8	77.8	15.8	68.1	12.2	60.0	**5.3**	35.5	4.6	31.4
HWT生成	20.4	77.8	16.6	70.5	12.1	58.8	5.5	37.5	4.8	33.3
HWT生成+宽度调整	19.7	80.1	14.3	66.6	11.4	58.6	5.4	36.8	**4.5**	31.2
HWT生成+宽度调整+多样化风格	18.8	76.1	13.3	61.2	11.0	55.8	5.4	35.9	**4.5**	31.6

相反的情况可以从使用ICFHR16数据集进行预训练、再在圣加仑数据集上微调的例子中看到（表 7）。ICFHR16是德语数据集，与使用拉丁语的圣加仑数据集存在显著差异。此外，由于ICFHR16包含多位作者书写的文本，其方差高于仅包含一位作者的圣加仑数据集。因此，在微调过程中，网络需要进行更多校正，以调整偏差并降低方差，从而专注于圣加仑文本，因此需要更多样本才能取得良好结果。总体而言，表 2、 3、 5和 4 的结果表明，平均而言，所有单作者数据集都能提升模型在Leopardi和圣加仑数据集上的性能；而在包含两位作者文本的华盛顿数据集中，具有许多不同书写风格且方差较大的多作者数据集（例如IAM 数据集）则有助于取得更好的结果。

表 6. Leopardi数据集与所考虑的相关数据集之间的语言比较，按 n 元语法的平均 Kullback-Leibler 散度排序。

	词汇相似性	Kullback-Leibler 散度			在 1.25% 上进行微调	
		一元语法	二元语法	三元语法	CER	WER
LAM	-	0.02	0.09	0.23	12.7	42.0
Leopardi 合成语	-	0.05	0.19	0.54	35.6	80.9
RIMES	9.54	0.11	0.84	1.89	25.3	68.7
IAM	6.76	0.17	0.85	1.62	21.7	63.2
ICFHR14	6.76	0.17	0.91	1.83	23.6	67.7
Rodrigo	10.45	0.20	0.71	1.48	34.9	81.1
NorHand	3.99	0.34	1.15	2.06	21.0	63.1
ICFHR16	4.19	0.40	1.50	2.50	39.0	86.0

表 7. 圣加仑数据集与所考虑的相关数据集之间的语言比较，按 n 元语法的平均 Kullback-Leibler 散度排序。

	词汇相似性	Kullback-Leibler 散度			在 1.25% 的数据上进行微调	
		一元语法	二元语法	三元语法	CER	WER
LAM	5.81	0.17	0.87	1.59	20.8	77.8
Rodrigo	6.08	0.18	0.89	1.74	14.4	66.4
RIMES	5.39	0.19	0.87	1.74	28.2	94.2
ICFHR14	3.50	0.20	0.79	1.43	20.4	77.8
IAM	3.50	0.21	0.74	1.31	16.5	68.3
圣加仑合成数据集	-	0.23	0.60	1.08	18.8	76.1
NorHand	2.39	0.42	1.22	1.78	27.0	87.9
ICFHR16	2.73	0.58	1.60	2.10	32.0	94.2

语言。表6、7和8 比较了预训练数据集与目标数据集之间的语言相似度和 Kullback-Leibler 散度（KL）。为突出文本信息与网络性能之间的相关性，我们按 KL 散度对表格排序，并列出使用目标数据集 1.25% 的数据进行微调后所得的结果。这样，我们强调了一个趋势：语言差异较小的数据集比差异较大的数据集更适合用于预训练。具体而言，从 KL 的角度看，德语 ICFHR16 平均而言与我们比较的所有目标数据集差异最大。因此，在该数据集上进行预训练，会导致三个目标数据集上的识别错误最多。另一方面，在英语数据集 IAM 和 ICFHR14 上进行预训练，由于它们与所有目标数据集的语言相似，因此能够取得良好性能。值得注意的是，从表6可以看出，从词汇角度而言，Leopardi数据集比包含该作者散文的合成数据集更接近LAM数据集。这三个数据集中的文本均为意大利语，创作年代相同，彼此相差不到70年。可以认为，LAM与Leopardi数据集更接近的原因是二者均为信件文集，具有许多相似的结构特征（例如日期、开头和称呼语）。综合考虑这一特点、语言和年代，在LAM数据集上预训练的卷积循环神经网络在Leopardi测试集上取得了出色的性能，尤其是在直接迁移的情况下（见表2）。

表 8. Washington数据集与所考虑的预训练数据集之间的语言比较，按 n 元语法的平均 Kullback-Leibler 散度排序。

	词汇相似性	Kullback-Leibler 散度			使用 1.25% 进行微调	
		一元语法	二元语法	三元语法	CER	WER
ICFHR14	-	0.05	0.30	0.66	26.1	64.6
用于华盛顿的合成语	-	0.07	0.31	0.69	23.9	63.4
IAM	-	0.08	0.30	0.59	18.8	52.7
NorHand	4.30	0.29	1.03	1.64	31.4	76.7
RIMES	9.67	0.31	1.36	2.22	27.1	78.9
Rodrigo	7.91	0.35	1.38	2.32	48.9	92.6
ICFHR16	4.72	0.36	1.24	1.83	58.0	97.8
LAM	6.76	0.36	1.52	2.31	37.2	81.7

5 结论

本文探讨了训练数据有限时历史手稿的行级手写文本识别。为此，我们提出在现有基准数据集上，或在大量合成数据上预训练专用的手写文本识别模型；这些合成数据反映了手稿目标作者的手写体特征，并通过全自动流程生成。随后，我们使用感兴趣文集中一部分真实数据进行微调。具体而言，我们对预训练数据集应具备的主要特征进行了广泛的定量分析，以便仅使用目标文集中的五行数据，也能获得性能强劲的手写文本识别模型。

实验结果表明，在选择真实数据集或生成合成数据集用于预训练时，应同时考虑整体外观（由纸张载体、书写工具和平均字符宽度决定）以及语言。此外，结果还表明，与仅使用单一手写风格训练的模型相比，使用手写风格变化较大的文本图像训练的手写文本识别模型更加稳健，也更容易适配。尽管如此，只要合成数据在手写体方面忠实地再现真实数据，仍能取得令人满意的性能。

鉴于这些结论，本研究有助于指导如何选择最合适的预训练数据集，以提升手写文本识别模型在小规模领域专用文档上的性能，并让人们进一步了解手写文本生成领域的成熟度及其对手写文本识别的潜在益处。最后，本研究也揭示了在单一作者文集上实现交互式、按需手写文本识别的可行性；这一任务值得进一步研究，以应用于数字档案的使用与完善。

致谢

本研究得到“面向数字人文的人工智能”项目（Pratica Sime n.2018.0390）的资助，该项目由摩德纳基金会资助；同时还得到意大利国家复苏与韧性计划（PNRR）项目“意大利加强欧洲科研基础设施韧性”（ITSERR）的资助，该项目由欧盟“下一代欧盟”计划提供资金（CUP: B53C22001770006）。

参考文献

1. Alonso, E., Moysset, B., Messina, R.：基于序列条件的手写文本图像对抗生成。载于：ICDAR（2019 3 2）。Aradillas, J.C., Murillo-Fuentes, J.J., Olmos, P.M.：利用少量标注行提升历史文档中的离线手写文本识别。IEEE Access（2021）2, 3 3. Augustin, E., Carr´e, M., Grosicki, E., Brodin, J.M., Geoffrois, E., Prˆeteux, F.：手写邮件处理的 RIMES评测活动。载于：IWFHR（2006）7, 8 4. Bella, G., Batsuren, K., Giunchiglia, F.：当代词典相似性的数据库与可视化（2021 10 5）。Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Khan, F.S., Shah, M.：手写体 Transformer。载于：ICCV（2021 2, 3, 4, 6, 7 6）。Bhunia, A.K., Das, A., Bhunia, A.K., Kishore, P.S.R., Roy, P.P.：基于对抗学习的低资源文字手写识别。载于：CVPR（2019）3 7. Bhunia, A.K., Ghose, S., Kumar, A., Chowdhury, P.N., Sain, A., Song, Y.Z.：MetaHTR：迈向书写者自适应手写文本识别。载于：CVPR（2021 2 8）。Bluche, T.：联合行分割与转录的端到端手写段落识别。载于：NeurIPS（2016 3 9）。Bluche, T., Louradour, J., Messina, R.：扫描、关注并阅读：利用 MDLSTM注意力的端到端手写段落识别。载于：ICDAR（2017 3 10）。Bluche, T., Messina, R.：用于多语言手写识别的门控卷积循环神经网络。载于：ICDAR（2017 3 11）。Cascianelli, S., Cornia, M., Baraldi, L., Cucchiara, R.：利用可变形卷积提升现代与历史手写文本识别。IJDAR，第1–11 页（2022 5 12）。Cascianelli, S., Cornia, M., Baraldi, L., Piazzi, M.L., Schiuma, R., Cucchiara, R.：学习阅读《L'Infinito》：利用合成语训练数据进行手写文本识别。载于：ICPR（2021 2, 4, 5, 7 13）。Cascianelli, S., Pippi, V., Martin, M., Cornia, M., Baraldi, L., Christopher, K., Cucchiara, R.：LAM数据集：一个用于行级手写文本识别的新型基准。载于：ICPR（2022 7, 8 14）。Chammas, E., Mokbel, C., Likforman-Sulem, L.：利用少量标注数据进行历史文档手写识别。载于：DAS（2018 3 15）。Cilia, N.D., De Stefano, C., Fontanella, F., di Freca, A.S.：一种基于排序的手写字符识别特征选择方法。Pattern Recognition Letters 121，77 3 16 86 2019–（）。Clanuwat, T., Lamb, A., Kitamoto, A.：KuroNet：基于深度学习的近代以前日本古文字字符识别。载于：ICDAR（2019 3 17）。Cojocaru, I., Cascianelli, S., Baraldi, L., Corsini, M., Cucchiara, R.：关注笔画：利用可变形卷积改进手写文本识别。载于：ICPR（2020 5 18）。Coquenet, D., Chatelain, C., Paquet, T.：使用门控全卷积网络进行无循环、无约束的手写文本识别。载于：ICFHR（2020）3 19. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.：可变形卷积网络。载于：CVPR（2017）5

20. Davis, B., Tensmeyer, C., Price, B., Wigington, C., Morse, B., Jain, R.：用于生成离线手写文本行的文本与风格条件生成对抗网络。收录于：BMVC（2020）3 21。Fischer, A., Frinken, V., Forn´es, A., Bunke, H.：使用隐马尔可夫模型对拉丁文手稿进行转录对齐。收录于：HIP（2011）7, 8 22。Fischer, A., Keller, A., Frinken, V., Bunke, H.：使用字符隐马尔可夫模型进行无词典手写单词检索。Pattern Recognition Letters 33(7), 934–942（2012）7, 8 23. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.：ScrabbleGAN：半监督可变长度手写文本生成。收录于：CVPR（2020）2, 3 24. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.：生成对抗网络。收录于：NeurIPS（ 2014 3 ）。25 Granet, A., Morin, E., Mouchere, H., Quiniou, S., Viard-Gaudin, C.：面向历史文档手写识别的迁移学习。收录于：ICPRAM（2018）2, 3 26. Graves, A., Schmidhuber, J.：使用多维循环神经网络进行离线手写识别。收录于：NeurIPS（2009）3, 5 27。Haines, T., Mac Aodha, O., Brostow, G.：我的文本在你的笔迹中。ACM Trans. Graphics 35 3 28 3 2016（）（）。Jaramillo, J.C.A., Murillo-Fuentes, J.J., Olmos, P.M.：利用迁移学习提升小型数据库中的手写文本识别。收录于：ICFHR（2018）3 29。Kang, L., Riba, P., Rusinol, M., Fornes, A., Villegas, M.：为手写识别生成内容与风格感知的文本行图像。IEEE Trans. PAMI pp. 1 4 30 1 2021–（）。Kang, L., Riba, P., Rusi˜nol, M., Forn´es, A., Villegas, M.：关注所读内容：非循环手写文本行识别。Pattern Recognition 129, 108766 2, 3, 4 31 2022（）。Kang, L., Riba, P., Wang, Y., Rusi˜nol, M., Forn´es, A., Villegas, M.：GANwriting：内容条件生成风格化手写单词图像。收录于：ECCV（2020）2, 3, 4 32。Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.：TrOCR：基于预训练模型的基于Transformer的光学字符识别。arXiv预印本 arXiv:2109 2, 3 33 10282 2021.（）。Lim, J.H., Ye, J.C.：几何生成对抗网络。arXiv预印本 arXiv: 1705 02894 2017 6 。34 （）。Maarand, M., Beyer, Y., Kasen, A., Fosseide, K.T., Kermorvant, C.：挪威语手写文本识别开源库的全面比较。收录于：DAS（2022）7, 8 35。Marti, U.V., Bunke, H.：IAM数据库：用于离线手写识别的英语句子数据库。IJDAR **5** 1 39 46 2002 7, 8 ( ), – ( )。Martın-Albo 36 Simon, D., Romero Gomez, V., Toselli, A.H., Vidal Ruiz, E.：文本图像的多模态计算机辅助转录及字符级交互。Int. J. Pattern Recognit. Artif. Intell. pp. 1263003 2 37 1 2012–（）。Mattick, A., Mayr, M., Seuret, M., Maier, A., Christlein, V.：SmartPatch：使用图像块判别器改进手写单词模仿。收录于：ICDAR（2021）3 38。Moysset, B., Kermorvant, C., Wolf, C.：整页文本识别：学习从何处开始以及何时停止。收录于：ICDAR（2017）3 39。Pham, V., Bluche, T., Kermorvant, C., Louradour, J.：Dropout可改进用于手写识别的循环神经网络。收录于：ICFHR（2014）3

40. Puigcerver, J.：多维循环层对于手写文本识别是否真的必要？载于：ICDAR (2017 3, 5)。Sanchez, J.A., Romero, V., 41 Toselli, A.H., Vidal, E.：ICFHR2014 手写文本识别竞赛：基于 Transcriptorium 数据集的手写文本识别（HTRtS）。载于：ICFHR (2014 7, 8 42)。Sanchez, J.A., Romero, V., Toselli, A.H., Vidal, E.：ICFHR2016 手写文本识别竞赛：基于 READ 数据集的手写文本识别。载于：ICFHR (2016 7, 8 43)。Serrano, N., Castro, F., Juan, A.：RODRIGO 数据库。载于：LREC (2010) 7, 8 44。Shen, X., Messina, R.：一种用于数据增强的中文手写图像合成方法。载于：ICFHR (2016 4 45)。Shi, B., Bai, X., Yao, C.：一种用于基于图像的序列识别的端到端可训练神经网络及其在场景文本识别中的应用。IEEE Trans. PAMI 39 3, 5, 6 46 11 2298 2304 2016 ( ), – ( )。Souibgui, M.A., Biten, A.F., Dey, S., Fornes, A., Kessentini, Y., Gomez, L., Karatzas, D., Llados, J.：用于低资源手写文本识别的一次性组合式数据生成。载于：WACV (2022 4 47)。Soullard, Y., Swaileh, W., Tranouez, P., Paquet, T., Chatelain, C.：利用光学模型和语言模型的书写者自适应提升文本识别效果。载于：ICDAR (2019) 3 48。Such, F.P., Peri, D., Brockler, F., Paul, H., Ptucha, R.：用于手写识别的全卷积网络。载于：ICFHR (2018 3 49)。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, 2017 3 L., 50 Polosukhin, I.：注意力就是你所需要的一切。载于：NeurIPS ( )。Voigtlaender, P., Doetsch, P., Ney, H.：利用大型多维长短期记忆循环神经网络进行手写识别。载于：ICFHR (2016) 3 51。Wang, J., Wu, C., Xu, Y.Q., Shum, H.Y.：结合形状模型和物理模型的在线草书手写合成。IJDAR **7** 4 219 227 2005 3 ( ), – ( )。52 Wick, C., Zollner, J., Gruning, T.：利用 CTC 前缀对文本行识别的序列到序列模型进行重新评分。arXiv 预印本 arXiv: 2110 05909 2021 2 。53 ( )。Wick, C., Zollner, J., Gruning, T.：基于双向后解码的手写文本识别 Transformer。载于：ICDAR (2021 3 54)。Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., Cohen, S.：使用 CNN-LSTM 网络进行手写单词和文本行识别的数据增强。载于：ICDAR (2017 3 55)。Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., Cohen, S.：开始、跟随、阅读：端到端整页手写识别。载于：ECCV (2018) 3 56。Yousef, M., Bishop, T.E.：OrigamiNet：通过学习展开，实现弱监督、无分割、单步式整页文本识别。载于：CVPR (2020) 3。