東南大學

本科毕业设计（论文）报告

题 目：面向低分辨率场景文本图像的增强与识别技术研究

学　　号：61520324

姓　　名：许睿

学　　院：吴健雄学院

专　　业：计算机科学与技术

指导教师：薛晖

起止日期：2023.12~2024.6

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：_____ 日期：_____年_____月_____日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：_____ 导师签名：_____

日期：_____年_____月_____日 日期：_____年_____月_____日

摘　　要

中文摘要

关键词：关键字 1，关键字 2

ABSTRACT

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et aut officiis debitis aut rerum necessitatibus saepe eveniet, ut et voluptates repudiandae sint et molestiae non recusandae. Itaque earum rerum defuturum, quas natura non depravata desiderat. Et quem ad me accedis, saluto: 'chaere,' inquam, 'Tite!' lictores, turma omnis chorusque: 'chaere, Tite!' hinc hostis mi Albucius, hinc inimicus. Sed iure Mucius. Ego autem mirari satis non queo unde hoc sit tam insolens domesticarum rerum fastidium. Non est omnino hic docendi locus; sed ita prorsus existimo, neque eum Torquatum, qui hoc primus cognomen invenerit, aut torquem illum hosti detraxisse, ut aliquam ex eo est consecutus? – Laudem et caritatem, quae sunt vitae sine metu degendae praesidia firmissima. – Filium morte multavit. – Si sine causa, nollem me ab eo delectari, quod ista Platonis, Aristoteli, Theophrasti orationis ornamenta neglexerit. Nam illud quidem physici, credere aliquid esse minimum, quod profecto numquam putavisset, si a Polyaeno, familiari suo, geometrica discere maluisset quam illum etiam ipsum.

KEY WORDS: Keywords1, Keywords2

目　　录

摘　　要 I
ABSTRACT II
目　　录 III
第一章 绪论 2
1.1 课题背景和意义 2
1.2 本文研究内容 3
1.3 论文各章节安排 3
第二章 相关工作 5
2.1 场景文本识别 5
2.2 场景文本图像超分辨率 7
2.3 扩散模型 7
2.4 多任务损失函数 7
2.5 数据集 7
第三章 模型原理与设计 8
3.1 模型总体架构 8
3.2 编码器模块设计 8
3.3 特征增强模块设计 8
3.4 识别器模块设计 8
3.5 超分辨模块设计 8
第四章 实验结果与分析 9
4.1 模型总体性能 9
4.2 与目前工作的对比 9
第五章 总结与展望 10
参考文献 11
附录 A 扩散模型推导 13
附录 B 模型推理结果 14
致　　谢 15

表 0.1 中英术语对照表

术语	英文	中文
STR	Scene-Text Recognition	场景文本识别
STISR	Scene-Text Image Super-Resolution	场景文本图片超分
NLP	Natural Language Processing	自然语言处理
OCR	Optical Character Recognition	光学字符识别
LR	Low Resolution	低分辨率
HR	High Resolution	高分辨率
SR	Super-Resolution	超分辨率

第一章 绪论

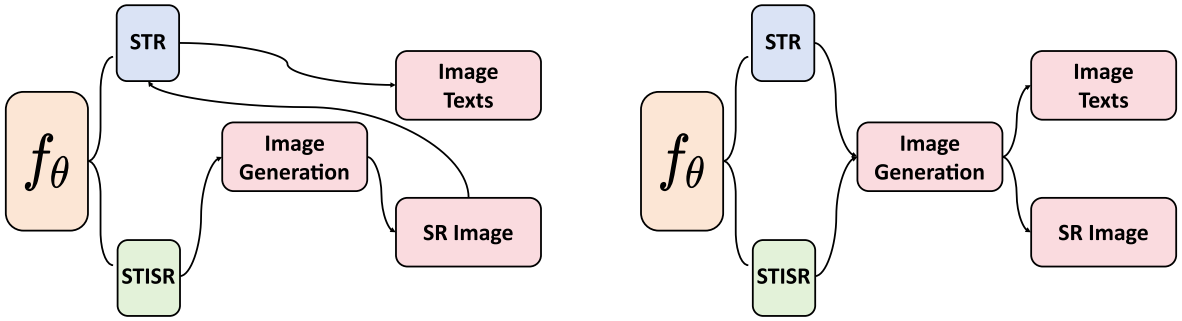
1.1 课题背景和意义

从一般性的场景图片中识别文本信息不仅能帮助深度学习模型在训练时理解场景逻辑，还能在推理时给予使用者更多场景相关的信息，该类型的任务被称为场景文本识别（Scene-Text Recognition, STR）。场景文本识别能在诸多实际应用中得到重要应用：例如在辅助驾驶中，汽车可以动态识别路上的交通符号，以便根据路况和交通规则做出安全的决策。近年来有许多研究人员躬身场景文本识别领域并取得了很多优秀成果。这些模型在目前广泛使用的基准数据集 [1] [2] [3] 中对于高分辨率（High-Resolution, HR）的图像取得了非常高的准确率。现实中的 STR 任务比基准数据集的形式更为复杂，被识别的字形可能因设计要求等因素，单个字符可以有多种字体和不同程度的扭曲，因此也有很多研究人员提出可以使用矫正模块[4] 对场景文本进行空间上的矫正，使其便于识别。因此给一张比较清晰的场景文本图像，当前有许多方式可以以较高置信度对其进行识别。

然而，在场景文本图像中的特征受到大量影响时，识别器的性能也大幅下降。例如，受到拍摄环境或拍摄器材的制约时，拍摄的场景文本图像可能带有一定的模糊，这对识别器的鲁棒性带来了巨大的挑战。因此有研究人员[5] 提出了更加一般的 STR 任务，即对于带有模糊等不良因素的低分辨率（Low-Resolution, LR）图像，可以先进行场景文本图像超分辨率（Scene-Text Image Super-Resolution, STISR）处理，得到较为清晰的超分辨率（Super-Resolution, SR）图像后，再进行常规的 STR。该研究人员构建了专门的模糊场景文本数据集[5] 用于评判模型在不良条件下的超分辨率和识别能力。与传统的图像超分辨率（Image Super-Resolution）任务不同，STISR 任务更加注重于文本的恢复，由于不少文本包含一定的语义信息，这不仅给图像超分辨率模型带来了挑战，也使 STISR 模型得到更多新的思路。

对于 LR 场景文本的识别，当前广泛使用的模型架构是如图 1-1 (a) 所示的串行架构。这种架构虽然直观，但同时存在不少问题。串行架构中，两个任务的耦合程度太高，STR 任务非常依赖于 STISR 任务的性能，如果 STISR 得到的 SR 场景文本图像存在一定的问题，则会对识别任务造成很大的麻烦；同时，STISR 任务模型的监督信号含有一定的识别损失而不是专门的超分辨率损失，因此并不会生成比较完整的 SR 场景文本图像。而本文实现的模型是如图 1-1 (b) 所示的并行架构，该架构首先使用统一的编码器将输入的图片映射到一个统一的特征空间，再通过不同任务的解码器输出相应的预测目标。该架构

的优势在于多种任务使用之间的耦合度低，识别效果受到SR质量影响的程度较低，同时SR的质量也不受识别损失的干扰。



(a) STISR 和 STR 模型的串行架构　　(b) STISR 和 STR 模型的并行架构

图 1-1　两种 STISR 和 STR 结合的架构

1.2 本文研究内容

本文主要使用如图 1-1 (b) 所示的多任务模型架构，分别设计了特征编码器、特征增强模块、识别解码器和超分辨率解码器四个部分，并使用了两阶段训练和多任务损失函数进行模型监督训练，最终实现了从LR场景文本图像到文本内容和SR图像的端到端任务。本文的主要贡献如下：

1. 提出了一般性的 LR 场景文本图像识别和超分辨率的模型架构；
2. 本文的模型使用了基于隐式扩散模型的特征增强模块，并在推理时对扩散模型进行有效的控制，实现在特征层面上增强 LR 图像；
3. 模型采用两阶段的训练模型，在第一阶段进行预训练，目的是让模型保存高分辨率特征信息，在第二阶段进行特征增强，目的是提升 LR 图像特征信息，提升多任务模型的泛化能力；
4. 本文使用基于不确定性权重的损失函数进行多任务模型训练，实现STR任务和STISR任务之间的平衡；

1.3 论文各章节安排

本文主要讨论STR任务和STISR任务的多任务模型和特征增强方法，以及本文所实现的模型在数据集上的性能。第二章介绍了目前现有的场景文本识别模型和场景文本图像超分辨率模型以及用于特征增强的扩散模型。第三章将会着重介绍本文提出的模型架构、模型设计思路的实现方式，并会从各个模块出发，分别介绍该模型实现特征增强的方

式和多任务解码方式。 第四章主要评估第三章中模型的性能，并展示了多种对比结果。最后，第五章将会对本文提出的方法进行讨论和总结。

第二章 相关工作

本章将分别介绍场景文本识别（STR）模型、场景文本图像超分辨率（STISR）模型和隐式扩散模型（Stable Diffusion Model），并讨论其与本文模型的关系。

2.1 场景文本识别

与传统的光学字符识别（OCR）不同，场景文本识别（STR）主要针对场景中的字符进行识别，OCR 面向的字符常常比较规整，而 STR 的识别对象经常由于拍摄问题，出现不同程度的透视或光影效果，有的甚至因为聚焦问题而出现低分辨率（LR）图像。对此问题，目前有很多研究人员提出了许多有效的方法，对高分辨率的场景文本进行识别，有的模型还通过额外设计的模块增强图片中的文本特征。基于不同架构，本节从基于卷积和循环网络、基于注意力机制和基于大语言模型三个角度介绍目前常用的 STR 模型，以及他们在基准数据集上的性能。

卷积和循环网络。卷积和循环网络在计算时拥有绝佳的性能，用该模块设计的网络常常具有训练和推理速度快、模型体积小等特点，因此仍有大量研究人员使用其作为图片特征提取和文本信息处理的主要模块。例如 SVTR [6] 模型使用不同步长的卷积运算模拟 ViT [7] 将图片分割为多个小块（patch）的操作，再对其进行后续处理，实验证明，这种图片处理方式能较为高效地将图片嵌入模型。CRNN [8] 模型则使用多个卷积层深度提取图像特征，并使用双向的 LSTM [9] 作为图片和文本之间的映射模块，最终使用 CTC 损失 [10] 监督模型训练，其识别结果在当时取得了较高性能。而对于含有透视、扭曲结构等场景文本中，ASTER [4] 模型在其主干网络之前添加了矫正模块，将弯曲文本矫正为水平排布的文本，并最终使用基于循环网络 LSTM [9] 和 GRU [11] 的识别模块，最终在弯曲文本的识别中取得非常好的性能。目前，大部分的模型都使用卷积提取图片特征，这种提取方式的优势在于能够尽量保存原有的图像特征，对于文本识别有很大帮助。本文实现的模型同样使用卷积作为编码器的主干。

注意力网络。注意力机制在 NLP 领域中得到了广泛应用，使用注意力机制能大幅提升模型对较长序列处理能力，不会出现类似 LSTM 等 RNN 模型在使用较长序列进行训练时出现的梯度消失或梯度爆炸的现象。对于 STR 任务，存在两种使用注意力机制的情况，一是针对文本进行注意力学习，再与图像特征进行交叉注意力学习，二是先将图像的特征提取后嵌入模型，再直接对图像进行注意力学习，最后用损失约束输出结果。例如 PARSeq [12] 使用了多种注意力机制进行训练。该模型的创新点在于使用排列数作为注

意力机制的掩码，在文本特征层面使用了自注意力机制，并与图像特征共同使用交叉注意力，学习文本特征和图像特征中的映射关系，在当时的数据集上取得了较高性能。由于 PARSeq 模型在多个合成数据集上进行了大规模预训练，因此最终得到的网络具有很强的先验知识，可以根据预训练网络进行微调。前文提到 SVTR [6] 模型利用带有不同步长的卷积运算模拟 ViT 将图像嵌入网络，同时，该模型还使用多个注意力机制用于提取局部和全局特征，该注意力机制使用不同大小的掩码块引导模型对图像注意力的学习。除此之外，MGP-STR [13] 以 ViT 为基础架构，将输入的文本图像分割为互不重叠小块。相较于 ViT 将不同的图像小块（即每个块的特征）聚合在一起作为整个图像的特征表示，在图像文本识别钟，MGP-STR 使用了一种基于注意力机制模块对识别器进行多重约束，用于提取更加注重于文本的特征，将图像的小块有意义地聚合在一起，使得该模型可以预测字母、词组和字母个数等不同粒度的文本信息。 DPAN [14] 模型在架构上进行了研究，并基于并行解耦的编码器-解码器架构（Parallel-Decoupled Encoder Decoder, PDED），改进了其注意力的查询输入矩阵（query），从而弥补了查询矩阵和键值矩阵（key）之间的图像信息，提高了模型的鲁棒性。

大语言模型辅助网络。在大语言模型（Large-Language Model, LLM）兴起后，很多与文本相关的任务都可以借助大语言模型实现。由于大语言模型使用大量的数据集进行训练，而模型也拥有巨大的参数量，因此训练出来的大语言模型有着很强的特征表示能力。早期的大语言模型如 BERT [15] 基于 Transformer [16] 编码器设计了双向编码的无监督训练，其参数量达到了 1.1 亿（110 million [15] ）。后续的 GPT 系列 [17] 模型则使用自监督的方式进行训练，其中 GPT-3 [18] 的参数量达到了 1.75 千亿（175 billion [18]）。在研究图像和文本之间的关系中，CLIP [19] 多模态模型使用对比学习，同时得到了文本和图像的特征表示编码器。大语言模型中拥有的大量参数使其具有很强的特征的表示能力，因此目前有研究人员直接使用大语言模型的编码器对下游任务中的数据集进行编码表示，并设计模型直接学习编码后的特征。 使用预训练的大模型可以在文本识别任务中取得较好性能。例如在 CLIP4STR [20] 工作中借助了视觉语言模型（Vision-Language Model, VLM），认为当前大部分识别器是基于单模态进行训练的，而 VLM 给予了文本和对应图像的信息，该模型通过视觉解码器和混合模态解码器的预测结果优化，在 11 个基准测试集中达到了较高的识别准确率。与此同时，由于大语言模型含有丰富的上下文语义信息，因此有研究人员使用大模型进行文本预测结果的优化。例如 ABINet [21] 设计了两个分支处理场景

文本识别，其一是以残差网络和 Transformer 为基础架构，并且带有位置注意力的视觉分支，其二则是以 CLIP 为基础的，用于优化预测文本的语言模型分支。实验发现使用语言模型进行结果的迭代优化可以在基准数据集上达到较高识别准确率。正如前文所述，使用预训练的大语言模型作为特征编码器或预测优化器，可以达到较高的识别准确率，同时由于大语言模型的强语义性，识别的结果绝大多数依然带有语义性，即识别结果是正确的单词。但在场景文本中，存在许多无上下文语义的文本，因此在使用大语言模型时需要权衡语义信息和图像信息之间的分配。

2.2 场景文本图像超分辨率

2.3 扩散模型

2.4 多任务损失函数

2.5 数据集

第三章 模型原理与设计

3.1 模型总体架构

3.2 编码器模块设计

3.3 特征增强模块设计

3.4 识别器模块设计

3.5 超分辨模块设计

第四章 实验结果与分析

4.1 模型总体性能

4.2 与目前工作的对比

第五章 总结与展望

参考文献

[1] RISNUMAWAN A, SHIVAKUMARA P, CHAN C S, et al. A robust arbitrary text detection system for natural scene images[J]. Expert Systems with Applications, 2014, 41(18): 8027-8048.

[2] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]//2013 12th international conference on document analysis and recognition. 2013: 1484-1493.

[3] VEIT A, MATERA T, NEUMANN L, et al. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images[Z]. 2016.

[4] SHI B, YANG M, WANG X, et al. Aster: An attentional scene text recognizer with flexible rectification[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(9): 2035-2048.

[5] WANG W, XIE E, LIU X, et al. Scene text image super-resolution in the wild[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. 2020: 650-666.

[6] DU Y, CHEN Z, JIA C, et al. SVTR: Scene Text Recognition with a Single Visual Model[Z]. 2022.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[Z]. 2021.

[8] SHI B, BAI X, YAO C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition[Z]. 2015.

[9] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J/OL]. Neural Comput., 1997, 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735. DOI:10.1162/neco.1997.9.8.1735.

[10] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C/OL]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006: 369-376. https://doi.org/10.1145/1143844.1143891. DOI:10.1145/1143844.1143891.

[11] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[12] BAUTISTA D, ATIENZA R. Scene text recognition with permuted autoregressive sequence models[C]//European conference on computer vision. 2022: 178-196.

[13] WANG P, DA C, YAO C. Multi-granularity prediction for scene text recognition[C]//European Conference on Computer Vision. 2022: 339-355.

[14] FU Z, XIE H, JIN G, et al. Look back again: Dual parallel attention network for accurate and robust scene text recognition[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. 2021: 638-644.

[15] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[17] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.

[18] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[19] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.

[20] ZHAO S, WANG X, ZHU L, et al. Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model[J]. arXiv preprint arXiv:2305.14014, 2023.

[21] FANG S, XIE H, WANG Y, et al. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7098-7107.

附录 A 扩散模型推导

附录 B 模型推理结果

致　　谢