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

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
PSNR	Peak Signal Noise Ratio	峰值信噪比
SSIM	Structural Similarity Index Measure	结构一致性指标

第一章 绪论

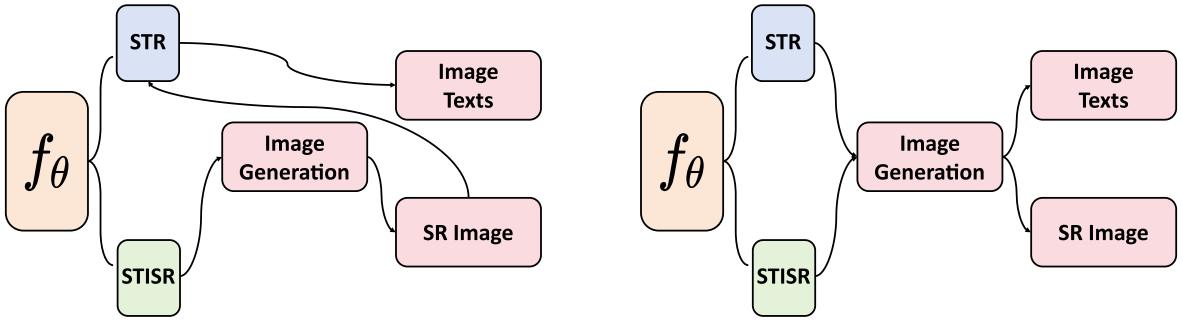
1.1 课题背景和意义

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）模型和场景文本图像超分辨率（STISR）模型，并讨论其与本文模型的关系。除此之外，本章还将介绍本工作将要使用的 TextZoom 数据集，以便后文能顺利地引入本文模型。

2.1 场景文本识别

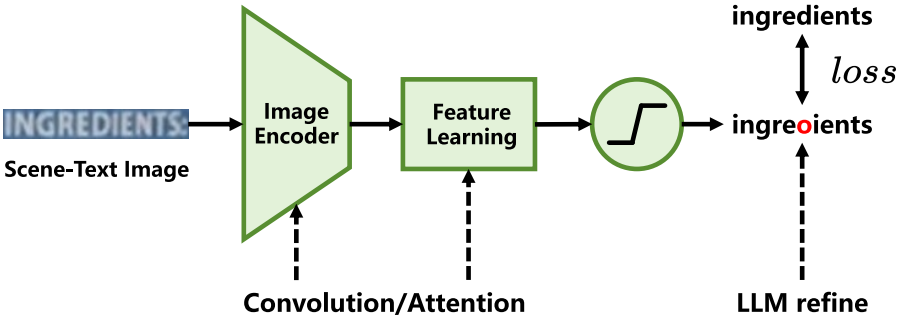


图 2-1　STR 识别器基本架构。其中虚线表示可选项

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，大部分模型的基本架构如图 2-1 所示，它们通常包含图像编码器用于将图像特征嵌入模型，再通过一系列的图像特征学习实现图像特征和文本特征的映射，最后使用激活函数将特征映射为标签，再经过损失函数将损失传播到模型参数，实现模型的训练。目前有研究人员在编码器和特征学习时使用卷积、循环神经网络以及注意力机制以便提取出更好的特征，除此之外，还有研究人员使用预训练的大语言模型对输出结果进行优化。基于相关工作中使用的不同特征提取方式，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。 尽管注意力机制的特征表示能力比卷积和循环模块强，但注意力机制可以让模型在全局特征的角度上进行计算，适合长度适中的特征序列，但

其计算复杂度随着序列长度的增加而成平方及增长，不仅如此，其模型参数规模也远大于卷积和循环模块。

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

场景文本图像超分辨率（STISR）的目的是提高低分辨率文本图像的质量，STISR 能大大提升前文提到的 STR 任务对低分辨率文本的识别率，因此对该领域的研究对自动驾驶等下游任务的发展有着重要意义。与单张图像的超分辨率略有不同，STISR 需要弥补图像特征和文本语义信息之间的模态差异，STISR 的工作包含了文本的语义信息，在增

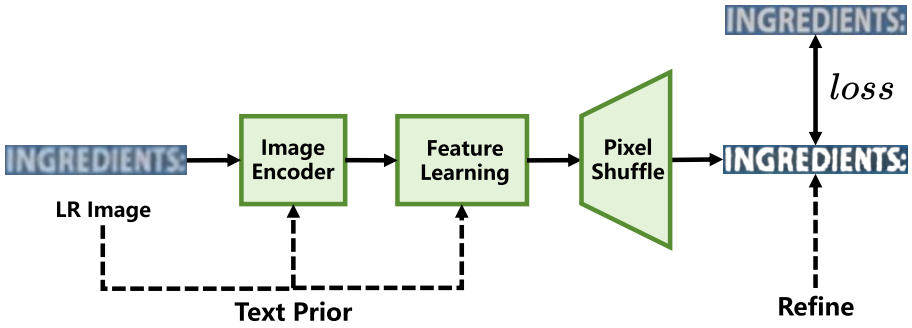


图 2-2　STISR 超分器基本架构。其中虚线表示可选项

强图像分辨率的同时不能破坏文本结构，因此 STISR 的约束条件比普通单张图像超分辨率更强。如图 2-2 所示，目前常用的模型框架通常包含一系列的编码器和特征学习模块，末尾通常使用像素重排模块 [22] 提高图像分辨率。目前，已经有很多研究人员在该领域做出了巨大贡献，例如有研究人员将文本先验知识引入图像的特征学习模块中，用先验特征引导超分辨率，还有研究人员在 SR 图像的基础上进一步做结果的优化。本节先介绍传统插值的图像超分辨率方法，接着以该方法为基础，引出当前该领域的相关工作。

插值。插值主要目的是根据给定数据点，估计给定点附近的值。用于图像超分辨率的插值方式包括最近邻插值、双线性插值和双三次插值等方式。 如图 2-3 假设原始图像大小为 (h, w)，以将图片分辨率放大到原来的**两倍**为例，即超分辨后的图像大小为 $(2h, 2w)$，分别讨论三种超分辨率方式的效果。对于一个像素点，最近邻插值忽略其周围像素值，直接在周围扩展该像素的值，其效果如图 2-3 （c）所示，可以发现其锯齿化比较严重。而双线性插值会考虑到以该像素点为顶点的周围像素值，假设四个顶点分别为 $A(x_1, y_1), B(x_1, y_2), C(x_2, y_2), D(x_2, y_1)$，这些点所代表的像素值分别由 $f(A), f(B), f(C), f(D)$ 给出，则该四个顶点间任意一个点的像素值 $P(x, y)$ 由公式公式 (2.1) 公式 (2.2) 和公式 (2.3) 给出。其插值效果如图 2-3 （d）所示。相较于最近邻插值，双线性插值在一定程度上减轻了锯齿化。而双三次插值插值一次使用的像素点个数是双线性插值的 4 倍，即使用 16 个像素值估计一个像素值，类似于双线性插值其基本思想是使用三次多项式函数拟合给定的四个像素值，并通过一阶导数和二阶混合偏导保证连续性，从而得到 16 个方程组成的方程组，解出 16 个顶点对应的权重。双三次插值的效果如图 2-3 （e）所示，可以看出其效果比双线性插值多了更多的图像细节。

$$f\left(P_{y_1}\right) = \frac{x_2 - x}{x_2 - x_1} f(A) + \frac{x - x_1}{x_2 - x_1} f(B) \tag{2.1}$$

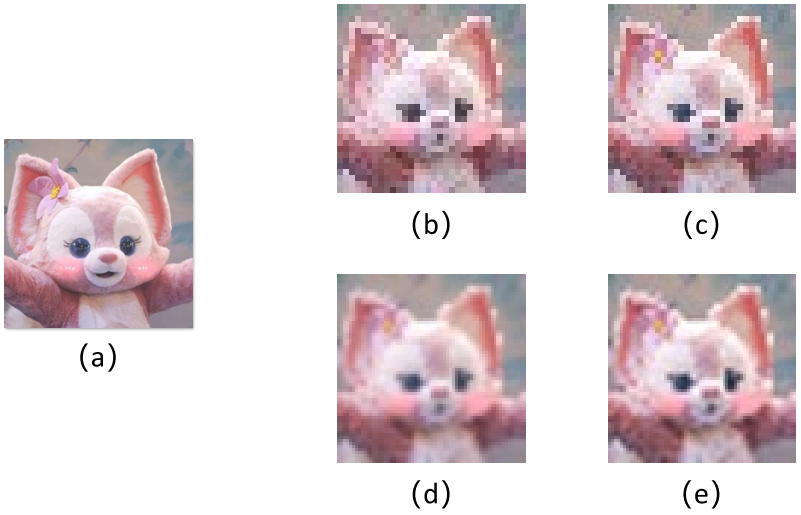


图 2-3 （a）表示原始高分辨率图像 $(2h, 2w)$，（b）表示压缩后大小为 (h, w) 的图像等比放大后的低分辨率图像 $(2h, 2w)$，（c）表示使用最近邻插值方式的超分图像 $(2h, 2w)$，（d）表示使用双二次插值的超分图像 $(2h, 2w)$，（e）表示使用双三次插值的超分图像 $(2h, 2w)$

$$f\left(P_{y_2}\right) = \frac{x_2 - x}{x_2 - x_1} f(D) + \frac{x - x_1}{x_2 - x_1} f(C) \tag{2.2}$$

$$f(P) = \frac{x_2 - x}{x_2 - x_1} f\left(P_{y_1}\right) + \frac{x - x_1}{x_2 - x_1} f\left(P_{y_2}\right) \tag{2.3}$$

使用插值的方式进行图像超分辨率，是在图像的宽和高两个维度上进行像素扩展，在使用深度学习模型进行超分辨率时，需要先将 LR 图像插值到 SR 图像，再使用卷积、注意力机制等模块增强其分辨率。而近年来提出的像素重排（PixelShuffle）模块 [22] 可以直接在 LR 图像上提取特征，它的主要思想是将通道维度特征重组到宽和高维度上，该方式既能提高图像分辨率，又能在通道维度上提取特征，是当前网络增强图像分辨率的主要方式。如图 2-4 所示，像素重排主要基于网络提取的特征图，特征图的通道维度上是具有超分辨像素信息的特征，通过重排后，通道维度的像素信息补全到宽和高两个维度，从而增强图像分辨率。与传统的插值方式相比，这种在通道维度上提取特征的方式可以减少网络传播中特征图的大小，同时也更适合具有卷积模块的网络在通道层面上提取特征的方式。基于该模块，有很多研究人员提出了新的超分辨率模型，并在基准数据集上取得了较好的结果。

模型。使用卷积和循环网络搭建模型的代表性工作是 TSRN [5]，同时该模型的作者也是 STISR 任务的提出者，并创建了新的数据集用于训练和测试。该模型的主要思路是残

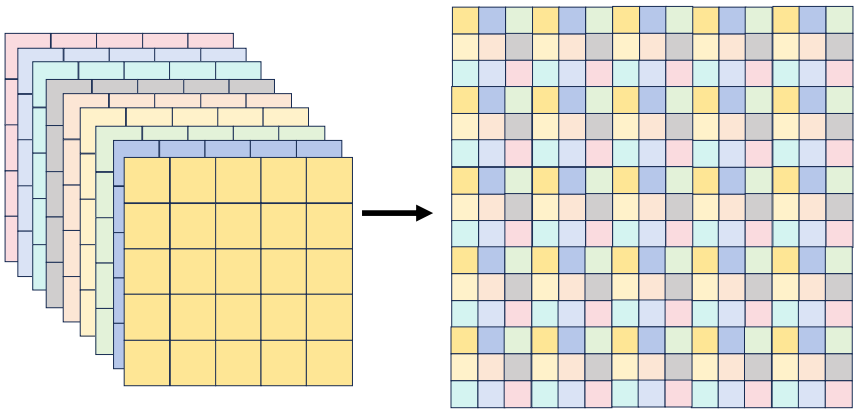


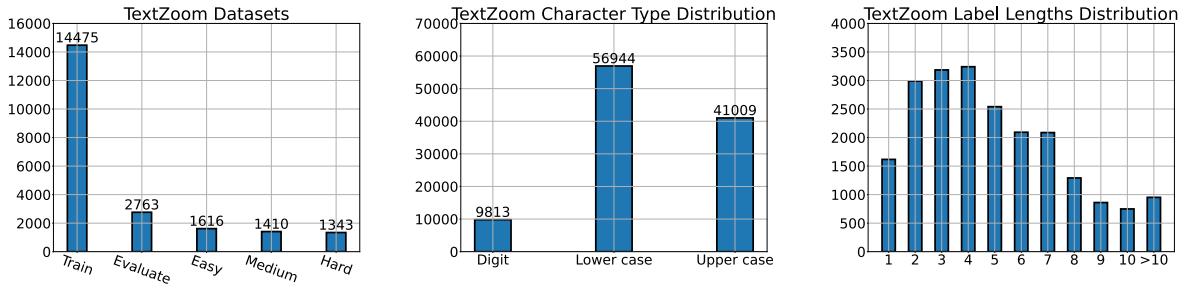
图 2-4　像素重排（PixelShuffle）模块的主要操作原理。原特征图的通道维度被重排至宽和高两个维度，从而提高图像的分辨率

差学习，模型使用了重复的超分子模块 SRB [5] 对超分图像的残差进行学习后，将残差与浅层特征进行加和，最后使用像素重排输出较高分辨率的图像。在整个网络的特征传播中，特征图的大小保持一致且与 LR 图像大小相同，模型仅在通道维度上进行特征学习，与前文介绍的像素重排相互印证。TSRN 是 STISR 领域的开山之作，因而有研究人员延续其特点，提出了新的架构。例如 TATT 模型 [23] 在此基础上增加了注意力机制，目的是混合先验文本信息和图像特征，在一定程度上用文本的先验特征引导超分辨率的进行。TATT 模型在 LR 图像输入时，通过预训练识别器对 LR 图像先进行一次识别得到预测的文本标签，同时用多个卷积模块提取 LR 图像的图像特征，再将二者输入到带有交叉注意力的特征融合模块，最后通过 TSRN 的主干网络输出超分辨率图像。与此思想类似的工作还包括 TPGSR [24]，该模型同样使用了文本先验模块对超分辨率模块进行引导，并同时使用由高分辨率图像产生的文本先验约束 LR 图像的文本先验，以便生成更可靠的文本先验。由于大部分超分网络具有一定的缺陷，有研究人员基于这些超分网络，设计了即插即用的超分辨率模块用于提高预训练超分模型的性能。DPMN [25] 则从优化的角度出发，在原有的超分辨率模型的基础上添加即插即用的性能提升模块，将现有的超分辨率模型输出的 SR 图像作为 DPMN 模块的输入，输入 PGRM（Prior-Guided Refinement Module）后使用预训练的 ViT 提取先验信息，并对其做像素重排，将多个 PGRM 的输出再进行组合，最终得到分辨率更高的 SR 图像。PCAN [26] 则训练出用于提取超分语义信息的注意力模块，并提出了新的损失函数用于模型的监督训练。STT [27] 模型则使用注意力机制关

注单个文本的字形的细节特征，并在特征图层面上用损失函数进行监督。而 C3-STISR [28] 则从不同特征理解的角度出发，利用注意力机制权衡识别、图像和语义三种理解，在基准数据集上超过同期模型。而随着 DDPM [29] 等扩散模型（Diffusion Model）的提出，研究人员发现该模型在图像恢复上面有一定的效果，因此部分研究人员将 DDPM 及其衍生模型引入 STISR 领域，扩充了文本图像的超分辨率模型类型。DDPM 在原图上进行多步加噪和多步去噪过程，用模型预测加的噪声，最终推理时用纯噪声逐步减去预测的噪声，实现图像的生成。由于原始的 DDPM 生成效果带有随机性，因此 SR3 [30] 基于隐式扩散模型（Stable Diffusion）将 LR 图像作为条件概率中的条件，引导生成图像的效果，最终在单张图像超分上取得了很好的效果。在 STISR 任务上，TextDiff [31] 首次使用扩散模型进行残差学习，在基准数据集上取得良好效果。

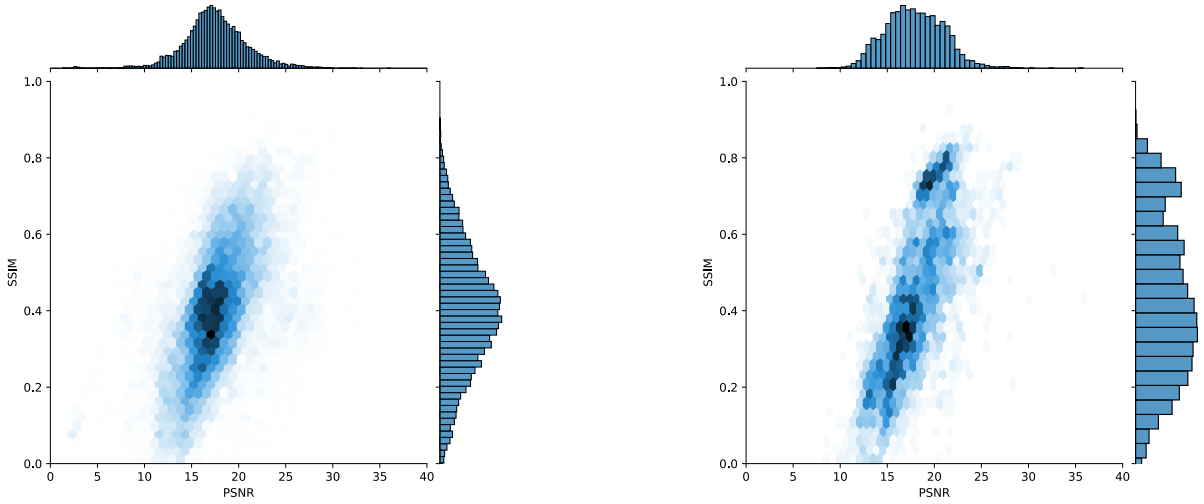
2.3 数据集

对于单独的 STR 任务而言，目前绝大部分识别器需要在大规模的人造数据集和真实数据集上进行训练，得到的模型通常包含了与文本相关的语义信息，可以在诸多下游任务的数据集上进行微调。用于大规模训练的人造数据集有包含了 900K 样本量的 MJSynth [32] 和包含了 800K 样本量的 SynthText [33]。当前已经有很多 STR 工作在这些数据集上进行训练，用于后续模型的微调。而用于 STISR 和 STR 双任务的数据集目前常用的是 TextZoom [5]，该数据集由两个单张图片超分辨率的基准数据集 RealSR [34] 和 SR-RAW [35] 筛选和截取而成，其中标签的分布如图 2-5 所示，图像的分布如图 2-6 所示。TextZoom 分为训练集（包含验证集）、简单难度的测试集、中等难度的测试集和困难难度的测试集，各个数据集的样本量如图 2-5 (a) 所示。如图 2-5 (b) 所示，在文本字符角度，数据集的标签文本类型包含了数字和大小写英文字母，每个图像的标签不仅包含数字，



(a) 各个数据集样本量　　(b) 标签文本类型　　(c) 标签文本长度

图 2-5 TextZoom 数据集标签特征统计



(a) 训练集 PSNR 与 SSIM 分布图　　(b) 测试集 PSNR 与 SSIM 分布图

图 2-6 TextZoom 数据集图像特征统计。PSNR 和 SSIM 越高表示 LR 图像质量越高

也有可能同时包含大小写的英文字母。如图 2-5 (c) 所示，数据集大部分的标签长度在10以内，且各个长度的分布比较均匀。在图像方面，图像的质量由 PSNR（Peak Signal Noise Ratio）和 SSIM（Structural Similarity Index Measure）反映，指标的值越高，则两张图像越接近。在图 2-6 中，本文统计了 LR 图像相对于 HR 图像的 PSNR 和 SSIM，可以发现训练集和测试集在分布上有一定的相似性，因此使用 TextZoom 训练集作为模型的训练数据。而由于测试集图像质量比较分散，因此使用不同难度的 LR 图像作为模型性能的评估数据集也是合理的。

同前文所述，使用在大规模人造数据集和真实数据集上预训练的识别器进行微调可以给模型带来一定的先验语义特征，TextZoom 数据集和 MJSynth 数据集在标签和图像分布上存在一定差异，适合用作微调数据集。因此本文识别器采用一种在 MJSynth 数据集上预训练后的识别器作为主干，并研究了相应的微调方法，最终取得较好的识别性能。

第三章 模型原理与设计

本章首先介绍模型的总体架构及其形式化描述，再分别详细介绍模型各个模块的设计和实现效果，最后给出模型训练所使用的损失函数以及训练和推理算法。

3.1 模型总体架构

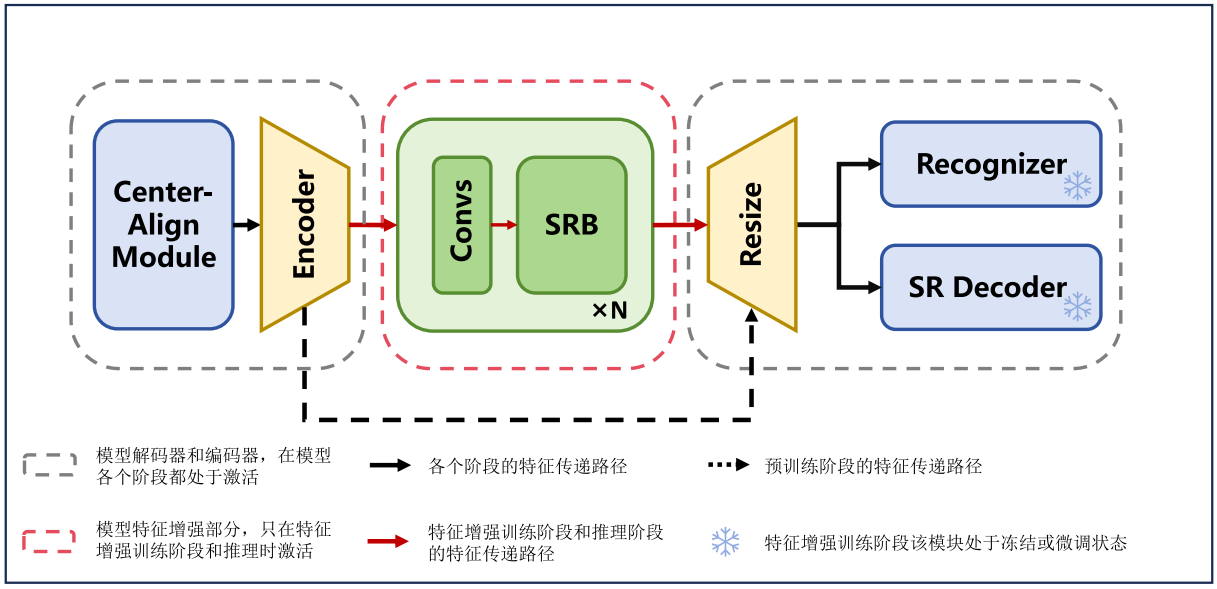


图 3-1 模型整体架构。模型编码器和解码器（灰色虚线框内部）在各个阶段都处于激活状态，特征增强部分（红色虚线框内部）只在特征增强阶段被激活和训练

不失一般性，模型的输入和输出都是以一个批量大小进行，超分辨率的倍数为 2，假设输入的 LR 低分辨率图像用 I_l 表示，对应的 HR 高分辨率图像用 I_h 表示，而超分辨率模型的输出 SR 为 I_s，图像对应的真实标签和识别器预测标签在嵌入模型后，分别为 S_t, S_p。图像和嵌入标签维度分别满足：$I_l \in \mathbb{R}^{B\times C\times H\times W}, I_h \in \mathbb{R}^{B\times C\times 2H\times 2W}, I_s \in \mathbb{R}^{B\times C\times 2H\times 2W}, S_t \in \mathbb{R}^{B\times L_t}, S_p \in \mathbb{R}^{B\times L_p}$ 其中 B 表示输入和输出数据的批量大小，C 表示输入和输出图像的通道数，通常为 3，而 H, W 分别表示 LR 图像的高和宽，由于放大倍数是 2，因此 SR 图像和 HR 图像的高和宽分别是 $2H, 2W$，对于标签，L_t, L_p 分别表示了真实标签的长度和预测标签的长度。

该模型的目标是通过训练集 $\mathcal{S} = \left\{(L_t, I_h, I_l)_i\right\}_{i=0}^{N}$ 训练模型 $\mathcal{F}_\theta$，使其能够以 I_l 为输入，预测对应的标签 L_p 和重建 SR 图像 I_s，即

$$\mathcal{F}(I_l) = \left(L_p, I_s\right) \tag{3.1}$$

从目前研究中发现，现有的模型多是基于串行结构搭建，即先对 LR 图像进行超分辨率，提高文本图像的分辨率，再用识别器进行文本预测。类似这一类的串行结构理解简单、训练更为方便，同时有很多研究人员达到了部分预期效果。然而这种框架经常会带来一些问题：

1. 首先，由于识别结果的准确率大多基于 SR 图像的质量，因此如果超分辨率模块的输出结果并不可靠，从而生成错误的超分辨率图像，就极有可能导致后续识别器预测出错误的文本。
2. 其次，两个任务是串行关系，因此如果不使用梯度截止等方法，后一个模块的损失在进行传播时会影响到前一个模块参数的更新，从而导致超分辨率模块被识别损失监督，原有的超分损失受到影响，从而生成的 SR 图像并不完全有着较高的保真度。

为解决以上问题，本文使用了一种新型的多任务并行架构，如图 3-1 所示，该模型的识别模块和超分模块处于并行关系，二者共享图像的编码特征，但解码特征相互独立。在使用多个损失函数进行监督训练时，两个解码器会更加注重于更新参数使得与其对应的损失降低，而不会受到另一个损失的影响。其次由于识别器从特征层抽取特征，受到超分辨率效果的影响较小，两个任务的耦合度大大降低。为了 进一步提升图像特征的质量，该模型还设计了特征增强模块，提高 LR 图像的特征质量。设编码器为 $\mathcal{E}_\theta$，识别器为 $\mathcal{D}_\theta^r$，超分器为 $\mathcal{D}_\theta^s$，特征增强模块为 $\mathcal{A}_\theta$，编码器输出特征为 $F_e \in \mathbb{R}^{B\times C_e\times H_e\times W_e}$，特征增强模块输出为 $F_a \in \mathcal{R}^{B\times C_a\times H_a\times W_a}$ （其中 C_x, H_x, W_x 分别表示中间特征的通道数、高度和宽度）该架构可以由公式 (3.2) 表示：在推理阶段，模型的输入为 LR 低分辨率图像，先经过编码器进行初步的特征提取，再由特征增强模块提取 LR 特征潜在的超分辨率图像特征和文本特征，最后将含有超分辨率和文本识别混合特征的张量图输入不同任务的解码器，从而输出识别文本和 SR 超分辨率图像，实现模型对 LR 图像的识别和超分辨率。

$$\begin{aligned} \mathcal{F}_\theta(I_l) &= (\mathcal{D}_\theta^r(F_a), \mathcal{D}_\theta^s(F_a)), \\ F_a &= \mathcal{A}_\theta(F_e), \\ F_e &= \mathcal{E}_\theta(I_l) \end{aligned} \tag{3.2}$$

整个模型的训练方式可以分为两个阶段：预训练阶段和特征增强阶段。

预训练阶段。该阶段特征增强模块不接入模型。模型主要使用 HR 图像训练识别器分支和超分辨率分支。该阶段目的是将 HR 图像的编码特征保留在编码器中，让编码器

学习到HR图像的特征解码方式，以便后续能以类似的方式提取LR图像的特征。其次是预训练超分器和识别器，提高超分器的重建效果和识别器对编码特征的识别性能。

特征增强阶段。该阶段主要训练特征增强模块，需要将其接入模型。特征增强思路是使用LR图像和HR图像共同进行训练，使得该模块能输出与HR编码后特征相近的超分辨率特征。同时，该阶段需要微调解码器的两个分支，以便弥补SR特征和HR特征之间的分布差异。

3.2 编码器模块设计

编码器主要负责对齐图像和混合特征。

图像中心对齐。由于数据集通过失焦采样真实的场景文本，LR图像中心和HR图像中心会出现一定的偏移，因此模型在编码器部分设计了空间变换网络（Spatial Transformer Network, STN）[36]用于对图像做空间尺度的变换，将LR图像与HR图像中心对齐。同时编码器还使用薄板样条插值（Thin Plate Spline, TPS）[37]扩大STN变换的灵活性。STN模块实现对图像变换的预测，增强卷积对图像剪切、旋转和仿射的鲁棒性，TPS模块使用插值的方式，通过关注不同图像控制点的变换，实现对图像除控制点以外其余部分的估计。该部分的基本流程为：先通过STN预测输入图像变换后控制点的位置，再将控制点和图像输入TPS实现图像的变换。该部分数学推导见附录。

混合特征。编码器的定位特殊，作为不同任务的前置特征提取模块之一，编码器的设计和训练需要多次实验测试。由第二章图2-1可知，通常的识别器包含图像特征编码器和后续的序列特征学习模块，而由第二章图2-2可知，通常的超分器同样包含图像特征编码器和后续的超分图像特征学习模块。因此本文设想通过构造一种以卷积为主的特征提取编码器，同时实现文本特征的提取和超分特征的提取。预训练阶段，编码器输入的是HR高分辨率图像，超分模块的重建任务较为简单，该阶段需要注重文本特征的提取；特征增强阶段，需要在编码器后接入特征增强模块，此时模型的输入是LR低分辨率图像，此时预训练完成的编码器会先对LR低分辨率图像进行特征提取，接着将LR低分辨率特征进行增强，输出适应预训练解码器的超分特征，该超分特征含有文本信息。因而在编码器设计部分，本文主要利用编码器实现图像文本信息的浅层提取，而超分变率特征主要由后续的特征增强模块实现。后续实验证明，只需要实验两层卷积层和一个激活函数，即可实现文本特征和超分特征的浅提取。

3.3 特征增强模块设计

特征增强模块主要负责提取编码后特征的超分辨率信息和文本信息。

3.4 识别器模块设计

识别器主要负责将输入的特征解码为文本。

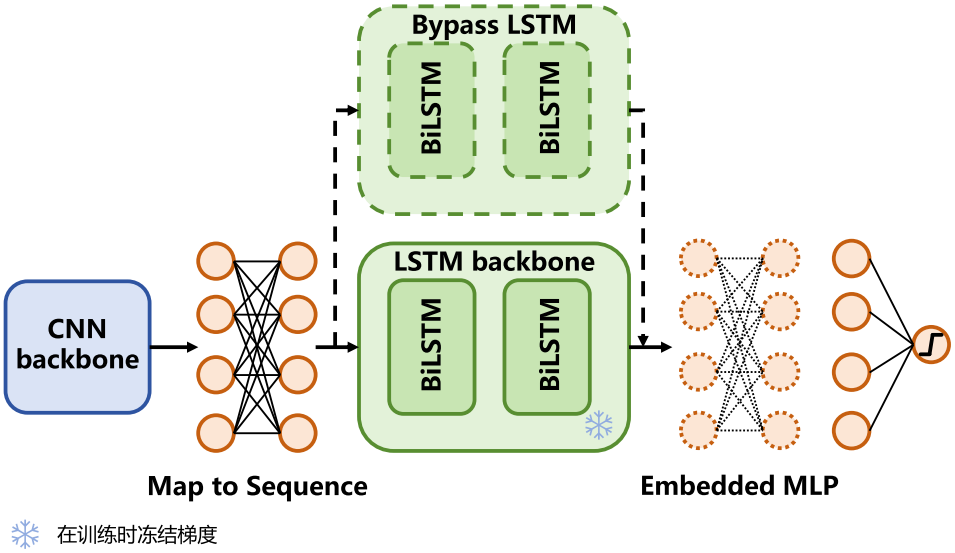


图 3-2　识别器模块。其中实线部分的模块和路线为预训练的 CRNN 主干，虚线部分的模块和路线为在微调时加入的额外的模块和路线

目前绝大多数识别器需要在大规模人造数据集或真实数据集上进行训练，其中常用的人造数据集有 MJSynth [32] 和 SynthText [33]。在大规模数据集上进行预训练的 STR 模型通常有较好的语义特征，泛化能力较强，而 TextZoom [5] 数据集规模远小于以上两种，且数据分布存在一定差异，适合作为微调数据集。经过调研，本文识别器选取 CRNN [8] 作为主干网络，并用该网络在 TextZoom 数据集上做微调训练。如图 3-2 所示，实线部分为

3.5 超分辨模块设计

超分辨率模块主要将输入的特征解码为 SR 超分图像。

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.

[22] SHI W, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[23] MA J, LIANG Z, ZHANG L. A text attention network for spatial deformation robust scene text image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5911-5920.

[24] MA J, GUO S, ZHANG L. Text Prior Guided Scene Text Image Super-Resolution[J/OL]. IEEE Transactions on Image Processing, 2023, 32: 1341-1353. DOI:10.1109/TIP.2023.3237002.

[25] ZHU S, ZHAO Z, FANG P, et al. Improving scene text image super-resolution via dual prior modulation network[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 3843-3851.

[26] ZHAO C, FENG S, ZHAO B N, et al. Scene Text Image Super-Resolution via Parallelly Contextual Attention Network[C/OL]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: Association for Computing Machinery, 2021: 2908-2917. https://doi.org/10.1145/3474085.3475469. DOI:10.1145/3474085.3475469.

[27] CHEN J, YU H, MA J, et al. Text gestalt: Stroke-aware scene text image super-resolution[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 36. 2022: 285-293.

[28] ZHAO M, WANG M, BAI F, et al. C3-stisr: Scene text image super-resolution with triple clues[J]. arXiv preprint arXiv:2204.14044, 2022.

[29] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[30] SAHARIA C, HO J, CHAN W, et al. Image super-resolution via iterative refinement[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(4): 4713-4726.

[31] LIU B, YANG Z, WANG P, et al. TextDiff: Mask-Guided Residual Diffusion Models for Scene Text Image Super-Resolution[J]. arXiv preprint arXiv:2308.06743, 2023.

[32] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Reading text in the wild with convolutional neural networks[J]. International journal of computer vision, 2016, 116: 1-20.

[33] GUPTA A, VEDALDI A, ZISSERMAN A. Synthetic data for text localisation in natural images[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2315-2324.

[34] JI X, CAO Y, TAI Y, et al. Real-World Super-Resolution via Kernel Estimation and Noise Injection[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): Vol. 0. 2020: 1914-1923. DOI:10.1109/CVPRW50498.2020.00241.

[35] ZHANG X, CHEN Q, NG R, et al. Zoom to learn, learn to zoom[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3762-3770.

[36] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.

[37] BOOKSTEIN F L. Principal warps: Thin-plate splines and the decomposition of deformations[J]. IEEE Transactions on pattern analysis and machine intelligence, 1989, 11(6): 567-585.

附录 A 模型推理结果

致　　谢