多域字符距离感知的场景文本图像超分辨率重建

黄俊炀,陈宏辉,王嘉宝,陈平平*,林志坚
(福州大学物理与信息工程学院,福建福州 350108)

摘 要: 场景文本图像超分辨率(Scene Text Image Super-Resolution, STISR)旨在提高文本在低分辨率图像中的分辨率和可读性. 但是在空间变形或低分辨率的文本图像中,由于缺乏文本区域细节,语义线索和视觉特征信息难以与字符位置匹配对齐,文本识别效果不佳. 针对该问题,本文提出多域字符距离感知的场景文本图像超高分辨率重建方法(Perceiving Multi-Domain Character distance super-resolution, PMDC),强化视觉语义特征,提高文本区域和纹理信息. 首先,采用非对称卷积以及语义先验信息模块,提取文本图像的视觉和语义特征信息;其次,融合字符距离感知模块中的视觉和语义特征,得到增强位置编码感知字符间的间距变化和语义相似性;最后,结合引导线索和视觉特征对像素进行重组得到超分辨率文本图像. 在公开数据集TextZoom上的实验结果,与最近TATT文本超分网络性能相比,在峰值信噪比指标上提高0.11 dB,有效提高文本清晰度和边缘纹理细节,同时提升1.5%的平均识别准确率,改进文本图像的可读性.

关键词: 计算机视觉;场景文本图像;超分辨率;注意力机制;特征信息关联

基 金 项 目: 国家自然科学基金(No. 62171135);福建省杰青项目(No. 2022J06010);福建省教育厅重点攻关项目(No.2023XQ004);福州科技局项目(No.2023-P-001)

中图分类号: TN911.73;TP391.43 **文献标识码:** A **文章编号:** 0372-2112(2024)07-2262-09

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240090

Scene Text Image Super-Resolution Reconstruction Based on Perceiving Multi-Domain Character Distance

HUANG Jun-yang, CHEN Hong-hui, WANG Jia-bao, CHEN Ping-ping*, LIN Zhi-jian
(*College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian* 350108, *China*)

Abstract: Scene text image super-resolution (STISR) aims to enhance the resolution and legibility of text in low-resolution images. In cases of spatial deformation or low-resolution text images, the lack of details in text regions and the difficulty in aligning semantic cues and visual features with character position make it difficult to recognize text effectively. In order to address these challenges, this paper proposes a perceiving multi-domain character distance for scene text image super-resolution method (PMDC), which improves the image text region and edge texture details. Firsly, the visual and semantic features are extracted by using the asymmetric convolution module along with the semantic prior module. Then the enhanced position coding is obtained by the character distance perception module to perceive the distance change and semantic similarity between characters. Finally, the guiding cues and visual features are combined to restructure the pixels and generate a super-resolution text image. In comparison to TATT, experimental results from the public dataset TextZoom showed an increase of 0.11 dB in the fidelity of the peak signal-to-noise ratio index. This improvement effectively enhances the clarity of the text area and the detailed edge texture. Additionally, the recognition accuracy was improved by 1.4%, which effectively enhances the readability of the text image.

Key words: computer vision; scene text images; super-resolution; attention mechanism; feature information association

Foundation Item(s): National Natural Science Foundation of China (No. 62171135); Fujian Talent Project (No. 2022J06010); Fujian Provincial Department of Education Key Research Project (No. 2023XQ004); Fuzhou Science and Technology Planning (No.2023-P-001)

收稿日期:2024-01-22;修回日期:2024-05-14;责任编辑:覃怀银

*通讯作者:陈平平

1 引言

图像中的文字是我们日常生活中重要的信息来源,可以根据不同的目的进行提取和解释. 场景文本识别的目标是读取自然图像中的文本,在提取视觉相关应用的文本信息方面具有关键作用,并广泛应用于自动驾驶[1]和基于场景文本的图像理解[2]等领域. 随着深度学习的发展,场景文本识别取得了很大的进展[3,4]. CRNN[5]使用CTC损失[6]进行训练,预测序列准确地与目标序列对齐. MORAN[7]和ASTER[8]有效解决弯曲文本的问题. SEED[9]和ABINet[10]将语义信息结合到识别模型中.

然而,由于成像过程中场景文本图像经常遇到各种质量退化,导致分辨率低且结构模糊,严重降低识别任务的性能[11],对低分辨率文本图像的识别性能仍然不理想. 因此,场景文本图像超分辨率(Scene Text Image Super-Resolution,STISR)[11]作为一种恢复低分辨率图像中缺失细节的预处理技术,可以有效提高场景文本图像分辨率,保障图像的视觉质量和字符的可读性.

近年来的场景文本图像超分辨率方法[12]主要分为高分辨率方法和线索引导方法的两类解决方案. CPIGAN[13]和SRCNN[14]等高分辨率方法将文本图像看作普通图像,直接使用图像超分辨率的方法来实现STISR. 但是在STISR的研究中,不仅要提高图像质量评价指标,还要提高下游识别任务的准确率. 高分辨率方法往往忽视文本图像的文本特性,无法达到令人满意的识别性能. TPGSR[15]、STT[16]和李滔等人的研究[17]等线索引导方法试图以文本特性为线索来引导超分辨率,并在图像质量和识别精度方面取得更好的表现.

但Wang等人的研究[18]表明这视觉和语义领域的线索是相互依赖的. 也就是说,如果其中一种线索较弱,另一种线索就无法稳定地找到对应的线索. 因此,在低分辨率的场景文本图像中,视觉特征和语义特征很容易不匹配. 此外,Yue等人的研究[19]还发现在长文本图像中,这种不匹配现象更为常见. 随着译码字符的增加,语义特征逐渐增强并主导识别,相邻时间步长之间的特征变得相似,容易引起注意漂移[20]和精度下降[21].

鉴于上述问题,Star-net网络[22]使用添加字符位置编码的方法,在一定程度上缓解了失配. 然而,它们的位置编码是与内容无关的,使用固定的位置编码来关联不同文本图像的视觉特征. Luo等人的研究[7]认为它更像是一个占位符约束,而不是内容感知受体. 它不能表示出现在固定位置的各种字符模式,可能会错误地在一个时间步长中处理多个字符. 这种类似占位符的用法低估了位置编码的效用. 因此,通过建模视觉空间、语义空间和位置空间之间的特征交互来产生一种增强位置编码可以实现更健壮的特征字符对齐. Ma等人的研究[23]认为该增强位置编码除了可以感知低分辨场景文本图像中字符的空间位置外,还可以用来描述字符的语义相似性. 结合这些线索生成一种内容感知嵌入,可以提取字符之间的间隔变化和语义相似性. 它被形象地理解为在视觉域和语义域上描述字符距离,即多域字符距离. 这种联合表征有利于感知低分辨率场景文本中字符的位置,从而引导文本图像的重建来提高图像分辨率和可读性.

针对以上问题,本文提出一种多域字符距离感知的场景文本超分辨率重建方法(Perceiving Multi-Domain Character distance super-resolution,PMDC),通过增强位置编码感知低分辨率文本图像中字符的位置,关注图像中的文本内容以及边缘纹理,提高图像文本区域清晰度和边缘纹理细节,同时提高模糊文本识别的准确性.

2 网络模型

在该部分,我们将对所提出的PMDC场景文本超分网络进行详细描述,包括整体网络结构的介绍以及特征信息提取网络,编码器和解码器的网络模块介绍.

2.1 整体网络结构

该方法旨在通过融合增强位置编码的感知内容位置线索引导,以输入低分辨率(Low-Resolution,LR)文本图像 $\boldsymbol{Y}_{\mathrm{LR}} \in \mathbb{R}^{h\times w\times 3}$,其中,$h$ 和 w 为图像 $\boldsymbol{Y}$ 的高度和宽度,3为图像的通道数,通过融合增强位置编码 $\boldsymbol{F}_p$ 的感知内容位置线索 $\boldsymbol{F}_{\mathrm{tp}}$ 引导,生成超分辨率(Super-Resolution,SR)图像 $\boldsymbol{Y}_{\mathrm{SR}} \in \mathbb{R}^{h\times w\times 3}$. 整体网络结构如图1所示.

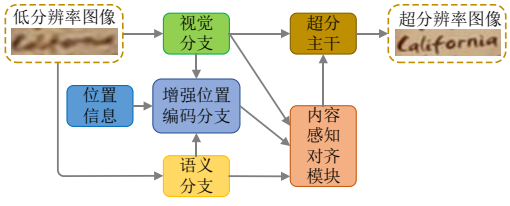

图1　整体网络结构

该方法是一个基于注意力机制的编码器-解码器框架,并可进行端到端训练. 具体来说,编码器对视觉分支、位置分支和语义分支进行编码融合. 位置分支用于查询视觉和语义特征,并将它们融合生成增强位置编码,增强位置编码记录了字符在视觉域和语义域之间的距离. 解码器方面,通过内容感知对齐模块将三种特征信息融合,生成感知内容位置线索. 该线索用作超分辨率主干网络部分的引导线索,引导像素进行重组重建,最终输出视觉和语义都得到提升的超分辨率文本图像.

2.2 特征提取网络

特征信息提取部分分为视觉分支、语义分支和位置分支. 特征提取网络框架图如图2所示.

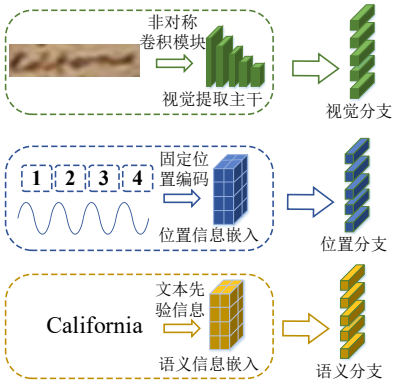

图2 特征提取网络框架

在视觉分支中,为了更好地获取低分辨率空间变形的文本图像视觉特征,采用非对称卷积[24]构建视觉特征提取模块. 输入LR图像 $\boldsymbol{Y}_{\mathrm{LR}}$ 通过非对称卷积层(Asymmetric Convolution Block,ACBlock)构建的主干特征提取网络,同时感知低分辨率文本图像在正方形、水平和垂直方向上的视觉特征 $\boldsymbol{F}_v$,计算过程如式(1)所示:

$$\boldsymbol{F}_v=\mathrm{ACBlock}(\boldsymbol{Y}_{\mathrm{LR}}) \tag{1}$$

其中,$\boldsymbol{F}_v\in\mathbb{R}^{h\times w\times c}$,$c$ 为特征的通道数. 这样可以增强视觉特征的表达能力,并提高模型对旋转失真图像的鲁棒性.

在语义分支部分,使用输入的低分辨率图像 $\boldsymbol{Y}_{\mathrm{LR}}$ 生成文本图像的语义特征 $\boldsymbol{F}_s$,并将该特征传递给增强位置编码分支和感知内容位置模块作为指导,以获得更加准确的场景文本超分辨率结果. 语义特征 $\boldsymbol{F}_s$ 是通过识别器Rec预测的概率分布计算得到的,将输入图像 $\boldsymbol{Y}_{\mathrm{LR}}$ 传递给识别器Rec,预测识别概念序列作为文本的先验信息 $\boldsymbol{T}_p$. 计算过程如式(2)所示:

$$\boldsymbol{T}_p=\mathrm{Rec}(\boldsymbol{Y}_{\mathrm{LR}}) \tag{2}$$

其中,$\boldsymbol{T}_p\in\mathbb{R}^{l\times|A|}$ 是由大小为 $|A|$ 的先验概率向量组成的长度为 l 的序列. A 表示由“0”到“9”、“A”到“Z”和一个空类(共37个字符)组成的字符集. 然后,将序列 $\boldsymbol{T}_p$ 投影到 c 通道来与图像特征通道相匹配,得到语义特征 $\boldsymbol{F}_s$.

在位置分支部分,训练过程中对特征进行一次性编码,而在推理过程中逐步更新特征. 为了在训练中对字符的位置进行编码,本文使用了固定位置编码方法对位置进行编码. 首先生成一个向量序列,其中每个向量在其位置索引维度上都具有固定的常量值,否则为0. 具体本文采用了与Vaswani等人[25]研究中相同的正弦位置嵌入方法,并通过两个多层感知器MLP层进行嵌入处理得到位置特征 $\boldsymbol{F}_{\mathrm{pos}}$. 而在推理过程中,位置嵌入首先被初始化为一个占位符,然后随着解码字符的积累而逐渐增长.

2.3 编码器

在编码器部分如图3左侧部分所示,对特征提取部分得到的三个特征分支进行跨分支交互以及融合得到增强位置编码 $\boldsymbol{F}_p$.

注意力机制的任务是获取局部关注的信息. 给定某个元素的查询向量,计算当前查询向量与所有元素键向量的相似性,得到对应值向量的注意力权重系数,最终对所有元素的值向量进行加权求和得到注意力值. 其中将特征提取部分得到的位置嵌入 $\boldsymbol{F}_{\mathrm{pos}}$ 作为查询向量,将其并行输入视觉分支和语义分支. 在视觉分支的注意力机制中,将位置信息作为查询向量,视觉信息作为键向量和值向量,计算位置信息 $\boldsymbol{F}_{\mathrm{pos}}$ 与视觉特征 $\boldsymbol{F}_v$ 之间的交叉注意力(Multi-head Cross-Attention,MCA)并经过前馈网络(Feed-Forward Network layers,FFN),即使用先前解码的字符位置来搜索文本图像中识别的字符区域. 同时,注意力机制应用于语义分支时,同样将位置信息作为查询向量,语义特征作为键向量和值向量,计算位置信息 $\boldsymbol{F}_{\mathrm{pos}}$ 与语义特征 $\boldsymbol{F}_s$ 之间的交叉注意,感知先前解码的字符与待识别字符之间的语义亲和力.

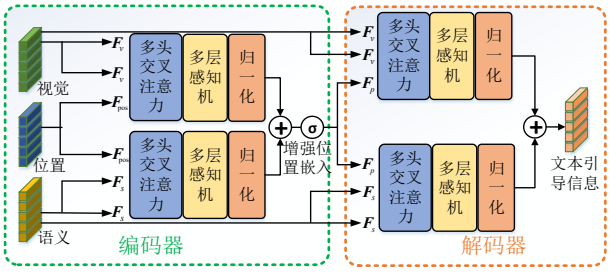

图3 编码器解码器结构图

因此,两个分支在相互作用后都使位置嵌入得到加强. 此外,上三角掩码[23] $\boldsymbol{M}_{\mathrm{pos}}$ 应用于查询向量,以防止“看到自己”,或者在时间步中泄漏信息. 两种交叉注意操作如式(3)和式(4)所示:

$$\boldsymbol{F}_{\mathrm{pv}}=\mathrm{Atten}((\boldsymbol{F}_{\mathrm{pos}},\boldsymbol{F}_v,\boldsymbol{F}_v),\boldsymbol{M}_{\mathrm{pos}})+\mathrm{FFN} \tag{3}$$

$$\boldsymbol{F}_{\mathrm{ps}}=\mathrm{Atten}((\boldsymbol{F}_{\mathrm{pos}},\boldsymbol{F}_s,\boldsymbol{F}_s),\boldsymbol{M}_{\mathrm{pos}})+\mathrm{FFN} \tag{4}$$

其中,FFN表示前馈网络,Atten(·)代表的是多头注意力机制,其中第一项参数中的三项依次表示多头注意力机制中的查询向量、键向量和值向量.

将交互表述为基于位置的视觉和语义领域的增强,它不仅在视觉特征方面强化了位置编码,而且还在语义特征方面强化了位置编码,可以理解为刻画字符位置之间的空间差异和语义相似性. 然后将两个位置增强交互分支得到的特征进行有效融合. 将这两个特

征串联起来形成一个混合特征,鼓励特征融合形成跨视觉和语义域的位置特征动态融合,最终得到增强位置编码 $\boldsymbol{F}_p$. 此操作的计算过程如式(5)所示:

$$\boldsymbol{F}_p=\text{sigmoid}((\boldsymbol{F}_{\text{pv}},\boldsymbol{F}_{\text{ps}})\boldsymbol{W}_{\text{conv}}) \tag{5}$$

其中,$\boldsymbol{W}_{\text{conv}}\in\mathbb{R}^{2\times c\times c}$表示相应的卷积.

2.4　解码器

多域感知解码网络如图3右侧部分所示,接受编码器模块得到的增强位置编码 $\boldsymbol{F}_p$ 和特征提取部分得到的视觉特征 $\boldsymbol{F}_v$ 以及语义特征 $\boldsymbol{F}_s$. 将 $\boldsymbol{F}_p$ 作为查询向量并行输入视觉分支和语义分支,同时对视觉和语义进行全局交叉关注,计算位增强位置编码与视觉特征和与语义特征之间的交叉注意 $\boldsymbol{F}_{\text{vis}}$ 和 $\boldsymbol{F}_{\text{sem}}$,然后进行融合映射得到文本引导映射向量 $\boldsymbol{F}_{\text{tp}}$. 计算过程如式(6)~(8)所示:

$$\boldsymbol{F}_{\text{vis}}=\text{Atten}((\boldsymbol{F}_p,\boldsymbol{F}_v,\boldsymbol{F}_v),\boldsymbol{M}_{\text{pos}})+\text{FFN} \tag{6}$$

$$\boldsymbol{F}_{\text{sem}}=\text{Atten}((\boldsymbol{F}_p,\boldsymbol{F}_s,\boldsymbol{F}_s),\boldsymbol{M}_{\text{pos}})+\text{FFN} \tag{7}$$

$$\boldsymbol{F}_{\text{tp}}=\boldsymbol{F}_{\text{vis}}+\boldsymbol{F}_{\text{sem}} \tag{8}$$

通过互注意力运算,将语义域以及视觉域的每个元素与空间域的位置关联起来,使得文本引导映射 $\boldsymbol{F}_{\text{tp}}$ 增强了空间域的语义以及视觉的意义区域,用于调整图像特征进行语义特定的文本重构.

最后,像素重建重组流程如图4所示. 将多域感知解码得到的文本引导映射 $\boldsymbol{F}_{\text{tp}}$ 和图像特征 $\boldsymbol{F}_v$ 传递到顺序循环块[11]重建高分辨率图像特征,再经过像素重组层[26]输出得到超分辨文本图像.

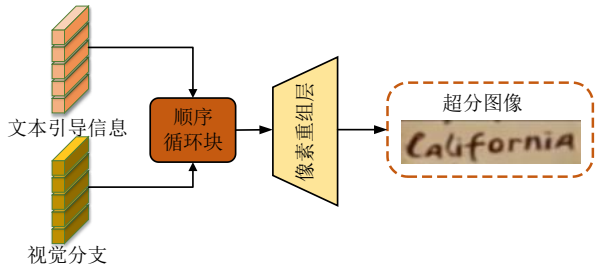

图4　像素重建重组流程图

像素重组层采用的是经典的上采样方法ESPCN[26],采用亚像素卷积的方法来对图像进行超分辨率重建. 像素重组层将低分辨率像素划分为$[r\times r]$份,默认是由特征图对应像素位置的 r^2 个特征像素组成一个低分辨率像素,在组成的过程中通过不断优化每组组合的权重来达到最好的上采样效果.

2.5　损失函数

在训练中,总体损失函数 L 包括超分辨率损失 L_{pix}、文本先验损失 L_{txt} 和文本结构一致性损失 L_{TSC}. 整体损耗函数描述如式(9)所示.

$$L=L_{\text{pix}}+\alpha L_{\text{txt}}+\beta L_{\text{TSC}} \tag{9}$$

其中,α 和 β 为平衡参数. 两种不同的损失 L_{pix} 和 L_{txt} 分别用于提供像素和文本特定的监督.

超分辨率损失 L_{pix} 采用 $L2$ 范数进行计算超分辨率图像SR输出 $\boldsymbol{I}_{\text{SR}}$ 和真实高分辨率图像 $\boldsymbol{I}_{\text{HR}}$ 之间的差值. 计算过程如式(10)所示:

$$L_{\text{pix}}=\left\|\boldsymbol{I}_{\text{HR}}-\boldsymbol{I}_{\text{SR}}\right\|_2 \tag{10}$$

采用文本聚焦损耗[17] L_{txt} 来权衡保真度和识别性能,计算过程如式(11)所示:

$$L_{\text{txt}}=\lambda_1\left\|\boldsymbol{A}_{\text{HR}}-\boldsymbol{A}_{\text{SR}}\right\|_1+\lambda_2\,\text{KL}(p_{\text{SR}},p_{\text{HR}}) \tag{11}$$

其中,$\boldsymbol{A}$ 和 p 分别表示的是固定识别器所预测的注意图和概率分布;KL 表示 Kullback-Leibler 散度;λ_1 和 λ_2 代表的是两个超参数.

L_{TSC} 采用的是TATT[23]所提出的文本结构一致性丢失,计算过程如式(12)所示:

$$L_{\text{TSC}}(\boldsymbol{X},\boldsymbol{Y})=1-\text{TSSIM}(D(F(\boldsymbol{Y})),F(D(\boldsymbol{Y})),D(\boldsymbol{X})) \tag{12}$$

其中,TSSIM 表示 TATT[23]中提出的三元 SSIM,$\boldsymbol{X}$ 为高清图片,$\boldsymbol{Y}$ 为低分辨率图片,D 表示随机变形,F 表示超分辨过程. $D(F(\boldsymbol{Y}))$ 表示超分辨率文本图像变形后的版本,$F(D(\boldsymbol{Y}))$ 表示低分辨率文本图像变形后的超分结果,$D(\boldsymbol{X})$ 表示真实高清标签图像变形后的结果.

3　实验结果与分析

在本节中,首先介绍了实验中使用的数据集、度量方式和实现细节. 然后,将本文所提出的方法与其他方法进行了比较. 最后,进行消融研究来验证该方法的有效性.

3.1　数据集

本文基于TextZoom[11]场景文本超分辨率数据集来评估算法的性能. TextZoom 数据集由 21 740 个低分辨率和高分辨率文本图像对组成,通过改变真实场景中相机的焦距收集,其中 17 367 个样本用于训练. 其余的样本根据相机焦距分为三个子集进行测试,即容易样本(1 619 个)、中等样本(1 411 个)困难样本(1 343 个)用来测试. 同时,TextZoom 数据集提供了文本识别的标签.

3.2　评价指标

本文使用识别精度来评估该方法的文本超分辨率识别性能. 根据先前研究的研究设置[17],去掉所有标点符号,并将大写字母转换为小写字母来计算识别精度. 此外,还使用峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)和结构相似度指数(Structural SIMilarity index, SSIM)来评估图像的保真度.

3.3　实验细节

本文算法模型基于 PyTorch1.8 实现. 所有实验都是在 2 块 NVIDIA GeForce RTX3090 GPU 上进行的,内

存为48 GB. 该模型使用Adam[27]优化器进行训练,学习率为0.001,批量大小设置为64. 所设计模型的输入图像宽度为64,高度为16,输出为2×HR结果. 将整体损失函数中的α和β设置为1和0.1,将文本聚焦损耗[17] L_{txt}中的λ_1和λ_2设置为10和0.000 5. L_{TSC}沿用TATT[23]中的变形操作D,通过在[−10, 10]度范围内应用随机旋转,在[0.5, 2.0]范围内应用剪切和纵横比来实现. 语义生成部分选用在SynthText和MJSynth上进行预训练的CRNN[5]. MCA层的头数设置为4. MCA、FFN计算中的图像特征通道数c均设为64. 模型参数量总计为15.95 M. 在训练时,使用CRNN的预训练权值初始化语义生成部分,其他部分随机初始化. 训练和评估是基于以下做法:在训练过程中保存平均最佳的模型,以CRNN作为识别器,并使用该模型评估其他识别器和三种设置[11](easy,medium,hard)得到最终的测试指标.

3.4 主流模型性能对比分析

本文在TextZoom[11]数据集评估了该方法,并在三个识别模型上将其与现有的超分辨率模型进行比较,包括CRNN[5]、MORAN[7]和ASTER[8]. 其中与MORAN和ASTER在三个不同难度分组的识别准确率的对比结果,结果如表1所示,其中加粗字体的为最佳值,下划线的为次优值.

表1 与现有主要方法在TextZoom测试集上进行文本识别准确率实验结果

算法	ASTER[8]				MORAN[7]			
	easy	medium	hard	avg	easy	medium	hard	avg
Bicubic	64.7%	42.4%	31.2%	47.2%	60.6%	37.9%	30.8%	44.1%
C3-STISR[28]	<u>79.1%</u>	63.3%	**46.8%**	<u>64.1%</u>	74.2%	<u>61.0%</u>	<u>43.2%</u>	<u>60.5%</u>
SRCNN[29]	69.4%	43.4%	32.2%	49.5%	63.2%	39.0%	30.2%	45.3%
HAN[30]	71.1%	52.8%	39.0%	55.3%	67.4%	48.5%	35.4%	51.5%
TSRN[17]	75.1%	56.3%	40.1%	58.3%	70.1%	53.3%	37.9%	54.8%
PCAN[31]	77.5%	60.7%	43.1%	61.5%	73.7%	57.6%	41.0%	58.5%
TPGSR[15]	77.0%	60.9%	42.4%	60.9%	72.2%	57.8%	41.3%	57.8%
TG[32]	77.9%	60.2%	42.4%	61.3%	**75.8%**	57.8%	41.4%	59.4%
TATT[23]	78.9%	<u>63.4%</u>	45.4%	63.6%	72.5%	60.2%	43.1%	59.5%
PMDC(Ours)	**79.8%**	**64.1%**	<u>46.5%</u>	**64.5%**	<u>75.2%</u>	**61.6%**	**44.6%**	**61.4%**
HR	94.2%	87.7%	76.2%	86.6%	91.2%	85.3%	74.2%	84.1%

以ASTER和MORAN为基准模型的条件下,除了在MORAN基本模型下的easy测试组的测试结果为次优值以外,PMDC在两个基准模型下的其他测试组以及平均识别准确率都达到最佳值. 具体的相较C3-STISR[28]在ASTER和MORAN识别器上分别提高了0.4%以及0.9%的字符平均识别准确率. 由实验结果可以验证,本文所提方法有效提高识别的准确性,证明了该方法的有效性和优越性.

以CRNN[5]为识别器在TextZoom[11]数据集的三个不同难度分组的识别准确率,以及平均识别准确率与现有超分辨率算法进行对比,实验结果如表2所示. 同时在表2中给出了实验得到的保真度指标(PSNR和SSIM)与现有主要方法的比较结果,其中加粗字体的为最佳值,下划线的为次优值.

以CRNN为识别器的情况下,除在easy测试组的识别结果略低于C3-STISR[28]算法的测试结果以外,PMDC的实验结果在medium和hard测试组以及整体的识别准确率都为最优值. 平均识别准确率比C3-STISR的次优值高了0.3%,相较TATT算法的结果高出1.4%. 同时本文所提出的PMDC方法在保真度指标上优于现有的主要超分辨率方法.

表2 与现有主要方法进行保真度和识别性能的比较

算法	评价指标		识别准确率/%			
	PSNR/ dB	SSIM	easy	medium	hard	avg
Bicubic	20.35	0.696 1	36.4	21.1	21.1	26.8
SRCNN[29]	20.78	0.722 7	38.7	21.6	20.9	27.7
TSRN[17]	21.42	0.769 1	52.5	38.2	31.4	41.4
PCAN[31]	21.49	0.775 3	59.6	45.4	34.8	47.4
TG[32]	21.4	0.745 6	61.2	47.6	35.5	48.9
TATT[23]	<u>21.52</u>	<u>0.793 0</u>	62.6	53.4	39.8	52.6
C3-STISR[28]	21.51	0.772 1	**65.2**	<u>53.6</u>	<u>39.8</u>	<u>53.7</u>
PMDC(Ours)	**21.63**	**0.795 4**	<u>64.1</u>	**55.2**	**40.6**	**54.0**

在PSNR指标上,相比TATT[23]和C3-STISR分别提高了0.11 dB和0.12 dB. 由实验结果可以验证,本文所提出的PMDC算法有效提高了图像中文本区域的清晰度和边缘的纹理细节,同时改进了文本图像可读性. 在hard类别中,由于图像过于模糊,难以获得准确的语义信息,引导超分辨率模型生成不完美的超分图像,误导后续识别. 因此,所提方案在hard类别的性能提升较不明显. 同时,在easy类别中,我们能够更好地利用语义信息辅助超分辨率过程,使得超分识别的效果提升较为显著.

同时,本节评估了PMDC在其他场景文本图像识别数据集上的泛化性能,包括ICDAR15[33]、CUTE80[34]和SVTP[35]. 这些数据集是为文本识别目的而构建的,包含自然场景中空间变形的文本图像. 依照TATT[23]中的实验设置,我们挑选数据集中低分辨率(即低于16×64)的图像来组成测试集,共有533个样本(ICDAR15中的391个,CUTE80中的3个和SVTP中的139个). 将PMDC网络在TextZoom数据集上以CRNN为识别器进行训练,在挑选的场景文本识别数据集中的低质量图像上进行测试,并与TSRN[17]、TBSRN[16]和TPGSR[15]以及TATT[23]进行比较. 实验结果如表3所示,其中加粗字体为最优结果,下划线字体表示次优结果. 本文所提出的PMDC算法获得了最高的平均识别准确率. 表明该网络虽然是在TextZoom数据集上训练的,但可以有效地应用到其他数据集中. 利用PMDC算法构造出超分辨率文本图像,有利于后续的场景文本识别等任务.

表3　在场景文本识别数据集中低分辨率测试集的实验结果

算法	识别准确率/%
Bicubic	18.1
TSRN[17]	26.6
TBSRN[16]	38.3
TPGSR[15]	42.5
TATT[23]	47.2
PMDC(Ours)	**49.6**

为了证明所提出算法的普适应,本文对图像进行超分辨率增强后,采用新近提出的SVTR[36]和ABINet[10]识别器来测试超分结果对识别准确率的提升效果,实验结果如表4所示.

表4　在场景文本识别数据集中低分辨率测试集的实验结果

算法	SVTR[36]	ABINet[10]
—	50.8%	60.0%
TG[32]	61.6%	66.0%
TPGSR[15]	64.5%	67.4%
C3-STISR[28]	65.1%	67.7%
PMDC(Ours)	**65.3%**	**67.8%**

直接使用典型的SVTR和ABINet场景文本识别算法,对低分辨率场景文本图像进行字符识别的效果不佳. STISR方法提高了低分辨率图像的识别效果. 与其他STISR方法相比,本文所提出的PMDC算法在不同的识别器中都获得最好的识别性能.

此外,在图5中可视化了一些例子,其中最上面一行为低分辨率场景文本图像样本,中间一行为算法所输出的超分辨率场景文本图像,最后一行为数据集中所提供的真实对照样本. 与其他方法相比,本文所提出的方法可以更好的恢复模糊像素,增加字符和背景的对比度有利于下游识别任务.

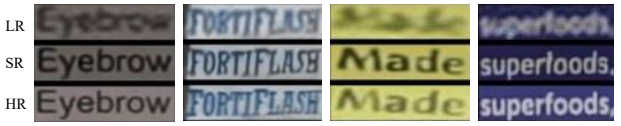

图5　PMDC在TextZoom[11]数据集上的可视化效果

3.5　消融实验

为了证明所提出的算法的有效性与可行性,本节设计三组消融实验证明算法模块对场景文本图像超分辨率结果的影响,以及可视化体现增强位置编码在视觉和语义领域所起的引导作用. 实验选用TextZoom[11]数据集,以CRNN[5]作为识别器.

3.5.1　增强位置编码的有效性

在本小节,评估感知字符间的间距变化和语义相似性的增强位置编码对场景文本图像超分辨率过程引导作用对实验结果的影响. 设计两组结构与增强位置编码进行对比,第一组使用整型值标记位置作为位置编码,第二组使用正弦位置嵌入作为位置编码来对超分过程进行引导. 消融实验结果如表5所示,其中,加粗字体是最佳值.

表5　采用不同的位置编码在TextZoom测试集上的实验结果

位置编码	PSNR/dB	SSIM	识别准确率/%
整型标记位置	21.07	0.753 9	40.3
正弦位置嵌入	21.23	0.768 0	46.5
增强位置编码	**21.63**	**0.795 4**	**54.0**

表5的实验结果显示,与前两组实验相比,使用增强位置编码对场景文本图像超分辨率进行引导的方式,在TextZoom数据集上的PSNR提高了0.4~0.56 dB,SSIM提高了0.027 4~0.041 5,识别的平均准确率提高了7.5%~13.7%. 实验证明,本文提出的增强位置编码可以很好地感知字符之间的空间和语义距离,提高低分辨率场景文本图像的视觉效果和可读性.

3.5.2　非对称卷积视觉特征提取网络模块有效性

在本小节,评估在场景文本图像超分辨率中采用非对称卷积搭建视觉特征提取网络对实验结果的影响. 在TextZoom[11]数据集上设置了一组对比实验,即在使用普通卷积搭建特征提取网络,其他设置均相同. 消融实验结果如表6所示.

由表6可知,在TextZoom数据集验证下,采用非对称卷积搭建视觉特征提取网络比使用普通卷积进行搭

表6　采用不同的卷积搭建视觉特征提取网络在TextZoom测试集上的实验结果

卷积	PSNR/dB	SSIM	识别准确率/%
普通卷积	21.47	0.786 9	52.3
非对称卷积	**21.63**	**0.795 4**	**54.0**

建的结构PSNR提高了0.16 dB,SSIM提高了0.008 5,识别的准确率提高了1.7%. 实验结果证明该特征提取模块增强了模型对空间变形文本图像的鲁棒性,有效提升了低分辨率场景文本图像的视觉效果和可读性.

3.5.3 非对称卷积模块、增强位置编码模块以及内容感知对齐模块消融实验

在本小节,评估了所提出的非对称卷积模块、增强位置编码模块以及内容感知对齐模块对实验结果的影响. 在TextZoom[11]数据集上以CRNN为识别器,将每个模块分别删去并与完整网络进行识别准确率和保真度的比较. 消融实验结果如表7所示,"×"和"√"分别表示不包含和包含该模块.

在实验组1中,本文删除了非对称卷积模块,降低视觉特征提取模块对不规则文本图像的视觉特征提取能力,与所提出的PMDC相比,识别准确率下降了1.7%. 然后在实验组2中,本文设计删除增强位置编码模块的实验组,即使用与内容无关的位置编码进行超分过程的引导,识别准确率较PMDC下降了7.5%. 最后在实验组3中,本文将内容感知对齐模块删除作为实验组比较,即仅使用图像视觉特征进行超分过程,识别准确率下降了8.6%. 实验结果证明非对称卷积模块、增强位置编码模块以及内容感知对齐模块的有效性和必要性.

表7 非对称卷积模块、增强位置编码模块以及内容感知对齐模块的消融实验结果

实验组	非对称卷积模块	增强位置编码模块	内容感知对齐模块	PSNR/dB	识别准确率/%
1	×	√	√	52.3	21.47
2	√	×	√	46.5	21.23
3	√	√	×	45.4	21.19
PMDC(Ours)	√	√	√	**54.0**	**21.63**

3.5.4 多域字符距离感知引导线索可视化

在本小节,将所提出的PMDC中结合图像视觉特征用来引导图片像素重组的感知多域字符距离线索进行可视化. 具体做法为将解码器最终输出得到的文本先验引导层通过热力图的形式渲染,得到最终可视化结果,具体可视化结果如图6所示.

图6 引导线索在TextZoom数据集上的可视化效果

其中第一行左侧为输入的低分辨率场景文本图像,右侧为图像经过CRNN[5]模型识别得到的识别结果,黑色字体为识别正确字符,红色字体为识别错误字符,第二行为将多域字符距离感知引导线索可视化的效果图,第三行左侧为通过本文所提出的算法得到的超分辨率文本图像,右侧为图像经过CRNN模型识别得到的识别结果. 由实验结果的热力图可以发现,多域字符距离感知引导线索可以正确引导从像素重建模块着重对图像中的字符内容进行超分辨率,并且由于融合了语义特征信息使得了引导线索具有上下文语义关联信息,同时能够明确地将所注重的字符内容与冗余的背景信息进行区分. 通过将处理前后的图像输入CRNN模型得到的识别结果进行对比. 可以发现本文所提出的算法可以有效的提升低分辨率文本图像的可读性,有利于下游的识别任务进行.

4 结论

本文提出了一种PMDC场景文本图像超分网络,解决低分辨率文本图像存在的语义线索和识别特征薄弱难以匹配的问题. 在视觉特征提取方面,使用非对称卷积设计特征提取模块,增强算法对空间变形文本图像的鲁棒性. 通过位置编码结合语义先验信息和视觉特征得到增强编码,感知字符之间的间距变化和语义的相似性. 融合三重特征得到感知多域字符距离线索引导像素重建模块,提升了低分辨率场景文本图像的分辨率和可读性. 在TextZoom数据集的实验结果表明,本文提出的PMDC能够有效提高低分辨率文本图像的视觉效果和识别的准确性.

参考文献

[1] ZHANG C S, DING W P, PENG G W, et al. Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(7): 4727-4743.

[2] SINGH A, NATARAJAN V, SHAH M, et al. Towards VQA models that can read[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 8317-8326.

[3] 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): 1-20.

[4] CHENG Z Z, BAI F, XU Y L, et al. Focusing attention: Towards accurate text recognition in natural images[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5076-5084.

[5] SHI B G, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (11): 2298-2304.

[6] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning - ICML' 06. New York: ACM, 2006: 369-376.

[7] LUO C J, JIN L W, SUN Z H. MORAN: A multi-object rectified attention network for scene text recognition[J]. Pattern Recognition, 2019, 90(C): 109-118.

[8] SHI B G, YANG M K, WANG X G, et al. ASTER: An attentional scene text recognizer with flexible rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2035-2048.

[9] QIAO Z, ZHOU Y, YANG D B, et al. SEED: Semantics enhanced encoder-decoder framework for scene text recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 13528-13537.

[10] FANG S C, XIE H T, WANG Y X, et al. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 7098-7107.

[11] WANG W J, XIE E Z, LIU X B, et al. Scene text image super-resolution in the wild[C]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 650-666.

[12] MANCAS-THILLOU C, MIRMEHDI M. An introduction to super-resolution text[M]//Digital Document Processing. London: Springer London, 2007: 305-327.

[13] 刘杰, 葛一凡, 田明. 文物图像的超分辨率重建算法研究[J]. 电子学报, 2023, 51(1): 139-145.
LIU J, GE Y F, TIAN M. Research on super-resolution reconstruction algorithm of cultural relic images[J]. Acta Electronica Sinica, 2023, 51(1): 139-145. (in Chinese)

[14] XU X Y, SUN D Q, PAN J S, et al. Learning to super-resolve blurry face and text images[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 251-260.

[15] MA J Q, GUO S, ZHANG L. Text prior guided scene text image super-resolution[J]. IEEE Transactions on Image Processing, 2023, 32: 1341-1352.

[16] CHEN J Y, LI B, XUE X Y. Scene text telescope: Text-focused scene image super-resolution[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 12026-12035.

[17] 李滔, 董秀成, 林宏伟. 基于深监督跨尺度注意力网络的深度图像超分辨率重建[J]. 电子学报, 2023, 51(1): 128-138.
LI T, DONG X C, LIN H W. Depth map super-resolution reconstruction based on deeply supervised cross-scale attention network[J]. Acta Electronica Sinica, 2023, 51(1): 128-138. (in Chinese)

[18] WANG T W, ZHU Y Z, JIN L W, et al. Decoupled attention network for text recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12216-12224.

[19] YUE X Y, KUANG Z H, LIN C H, et al. RobustScanner: Dynamically enhancing positional clues for robust text recognition[C]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 135-151.

[20] WAN Z Y, HE M H, CHEN H R, et al. TextScanner: Reading characters in order for robust scene text recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12120-12127.

[21] LIAO M H, ZHANG J, WAN Z Y, et al. Scene text recognition from two-dimensional perspective[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8714-8721.

[22] LIU W, CHEN C F, WONG K Y, et al. STAR-net: A spatial attention residue network for scene text recognition[C]//Proceedings of the British Machine Vision Conference 2016. Glasgow: British Machine Vision Association, 2016: 22482128.

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

[24] DING X H, GUO Y C, DING G G, et al. ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1911-1920.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.
[26] SHI W Z, CABALLERO J, HUSZAR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 1874-1883.
[27] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22) [2024-05-14]. https://arxiv.org/abs/1412.6980.
[28] ZHAO M Y, WANG M S, BAI F, et al. C3-STISR: Scene text image super-resolution with triple clues[EB/OL]. (2022-04-29)[2024-05-14]. https://arxiv.org/abs/2204.14044.
[29] DONG C, LOY C C, HE K M, et al. Image super-resolution using deep convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(2): 295-307.
[30] NIU B, WEN W L, REN W Q, et al. Single image super-resolution via a holistic attention network[C]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 191-207.
[31] QUAN Y H, YANG J T, CHEN Y X, et al. Collaborative deep learning for super-resolving blurry text images[J]. IEEE Transactions on Computational Imaging, 2020, 6: 778-790.
[32] CHEN J Y, YU H Y, MA J Q, et al. Text gestalt: Stroke-aware scene text image super-resolution[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 285-293.
[33] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]// 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE, 2015: 1156-1160.
[34] 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.
[35] PHAN T Q, SHIVAKUMARA P, TIAN S X, et al. Recognizing text with perspective distortion in natural scenes[C]//2013 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2013: 569-576.
[36] DU Y K, CHEN Z N, JIA C Y, et al. SVTR: Scene text recognition with a single visual model[EB/OL]. (2022-04-30)[2024-05-14]. https://arxiv.org/abs/2205.00159.

作者简介

黄俊炀 男,1998年12月出生于福建省泉州市. 现为福州大学物理与信息工程学院研究生. 主要研究方向为计算机视觉、场景文本图像超分辨率、场景文本识别.
E-mail: vampire_hjy@foxmail.com

陈宏辉 男,1998年9月出生于福建省南平市. 现为福州大学物理与信息工程学院研究生. 主要研究方向为计算机视觉、场景文本检测、场景文本端到端检测识别.
E-mail: 726673517@qq.com

王嘉宝 男,1999年11月出生于福建省泉州市. 现为福州大学物理与信息工程学院研究生. 主要研究方向为计算机视觉、场景文本检测、场景文本识别.
E-mail: 13110543916@163.com

陈平平 男,1986年出生于福建省泉州市. 现为福州大学电子信息工程系教授,博士生导师. 主要研究方向为信息处理、人工智能与计算机视觉. 中国电子学会会员编号:E190021215M.
E-mail: ppchen.xm@gmail.com

林志坚 男,1984年出生于福建省漳州市. 现为福州大学电子信息工程系副教授,硕士生导师. 主要研究方向为车联网、计算机视觉与模式识别.
E-mail: zlin@fzu.edu.cn