计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):103−112 [doi: 10.15888/j.cnki.csa.009479]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于模糊模式感知模块的场景文本图像超分辨率算法①

张　密, 余海洋

(复旦大学 计算机科学技术学院, 上海 200438)
通信作者: 张　密, E-mail: mizhang19@fudan.edu.cn

摘　要: 现有的场景文本识别器容易受到模糊文本图像的困扰, 导致在实际应用中性能较差. 因此近年来研究人员提出了多种场景文本图像超分辨率模型作为场景文本识别的预处理器, 以提高输入图像的质量. 然而, 用于场景文本图像超分辨率任务的真实世界训练样本很难收集; 此外, 现有的场景文本图像超分辨率模型只学习将低分辨率 (LR) 文本图像转换为高分辨率 (HR) 文本图像, 而忽略了从 HR 到 LR 图像的模糊模式. 本文提出了模糊模式感知模块, 该模块从现有的真实世界 HR-LR 文本图像对中学习模糊模式, 并将其转移到其他 HR 图像中, 以生成具有不同退化程度的 LR 图像. 本文所提出的模糊模式感知模块可以为场景文本图像超分辨率模型生成大量的 HR-LR 图像对, 以弥补训练数据的不足, 从而显著提高性能. 实验结果表明, 当配备提出的模糊模式感知模块时, 场景文本图像超分辨率方法的性能可以进一步提高, 例如, SOTA 方法 TG 在使用 CRNN 文本识别器进行评估时, 识别准确率提高了 5.8%.

关键词: 场景文本图像超分辨率; 场景文本识别; 图像模糊模式; 条件生成对抗网络; 深度学习

引用格式: 张密,余海洋.基于模糊模式感知模块的场景文本图像超分辨率算法.计算机系统应用,2024,33(4):103–112. http://www.c-s-a.org.cn/1003-3254/9479.html

Scene Text Image Super-resolution Algorithm Based on Blurring Patterns Aware Module

ZHANG Mi, YU Hai-Yang

(School of Computer Science, Fudan University, Shanghai 200438, China)

Abstract: Existing scene text recognizers are prone to be troubled by blurred text images, leading to poor performance in practical applications. Therefore, several scene text image super-resolution models have been proposed as the pre-processor for text recognizers to improve the quality of input images. However, real-world training samples for the scene text image super-resolution task are difficult to collect. In addition, existing STISR models only learn to transform low-resolution (LR) text images into high-resolution (HR) text images while ignoring blurring patterns from HR to LR images. This study proposes a blurring pattern aware module (BPAM), which learns blurring patterns from existing real-world HR-LR pairs and transfers them to other HR images for generating LR images with different degrees of degradation. Therefore, the proposed BPAM can produce massive HR-LR pairs for STISR models to compensate for the deficiency of training data, significantly improving performance. The experimental results show that when equipped with the proposed BPAM, the performance of SOTA STISR methods can be further improved. For instance, the SOTA method TG achieves a 5.8% improvement in recognition accuracy with CRNN for evaluation.

Key words: scene text image super-resolution (STISR); scene text recognition; image blurring pattern; conditional generative adversarial network (CGAN); deep learning

① 收稿时间: 2023-09-28; 修改时间: 2023-11-03, 2023-12-04; 采用时间: 2023-12-20; csa 在线出版时间: 2024-03-01
CNKI 网络首发时间: 2024-03-07

近年来, 场景文本识别在自动驾驶、车票识别和身份证识别等行业中得到了广泛应用. 然而, 在这些行业中捕获的文本图像往往容易受到成像过程中出现的低分辨率以及画面模糊等问题的影响, 这严重限制了场景文本识别器的性能. 因此, 场景文本超分辨率算法[1–6]可以用作场景文本识别器之前的预处理器, 获取模糊文本图像中的文本信息, 以提升低分辨率场景下文本图像的质量.

早期的方法[7,8]将场景文本图像超分辨率 (scene text image super-resolution, STISR) 视为一般的超分辨率任务, 忽略了场景文本图像中蕴含的文本特有属性. 最近, 一些方法试图引入与文本相关的先验知识, 以帮助模型更好地重建文本区域的细节, 从而获得更具视觉效果的超分辨率 (super-resolution, SR) 图像. PlugNet[9]基于多任务学习策略, 提出了一种可插拔的超分辨率分支, 该分支与识别分支联合训练, 提取具有文本语义的特征, 进一步提高了识别性能. TSRN[1]在特征提取模块中使用循环神经网络 (RNN) 来捕获文本序列相关性信息. TG[2]通过预训练一个基于 Transformer 的文本识别器, 使用其注意力图获取笔画级别的监督来恢复图像中文本区域的笔画细节. C3-STISR[3]在 TG 的基础上, 还引入了文本视觉和文本语义信息作为超分辨率过程的指导. TPGSR[4]以迭代的方式将字符概率序列作为文本先验添加到超分辨率模型中. 尽管这些方法可以在一定程度上提升低分辨率的模糊图像的质量, 但它们受到真实场景中训练数据集不足的限制. 例如, 目前场景文本图像超分辨率任务使用的数据集 TextZoom[1]仅包含 17 367 个用于训练的样本, 这对于当前基于深度学习的 STISR 模型来说仍然是远远不够的.

大多场景文本图像超分辨率方法将场景文本图像视为一般图像, 忽略了场景文本图像的文本先验信息. 如图 1(b) 所示, 尽管一些方法[5,6]提出引入额外的文本相关监督, 使得这些模型专注于图像中的文本区域, 但这些方法只试图学习从低分辨率 (LR) 图像到高分辨率 (HR) 图像的映射, 完全忽略了从 HR 到 LR 图像的模糊模式.

为此, 本文提出了模糊模式感知模块 (blurring pattern aware module, BPAM), 从现有训练数据中的 HR-LR 文本图像对中学习真实世界的模糊模式. 采用本文提出的模糊模式感知模块将学习到的模糊模式转移到其他高分辨率图像中, 以生成更多符合真实场景下低分辨率模糊图像分布范式的 HR-LR 文本图像对供场景文本图像超分辨率模型进行训练, 这将极大地缓解训练数据短缺的问题, 并使场景文本图像超分辨率模型对真实世界的低分辨率文本图像更具鲁棒性与泛化性. 具体而言, 本文使用条件生成对抗性网络 (conditional generative adversarial network, CGAN)[10]来构建模糊模式感知模块, 其中输入条件表示即将生成的低分辨率模糊图像的不同退化程度. 为了在训练阶段提供退化程度, 在 TextZoom 数据集中手动注释了低分辨率图像的退化程度. 本文提出的模糊模式感知模块仅在训练阶段与场景文本超分模型联合训练使用, 而不参与推理阶段. 因此在训练完成后, 可以在不增加文本超分模型推理开销的条件下, 显著地提升其性能. 本文使用两种图像质量指标 (即 *SSIM* 和 *PSNR*) 和识别准确度来验证模糊模式感知模块的有效性. 实验结果表明, 当配备本文所提出的模块时, 现有场景文本图像超分辨率方法的性能可以获得明显提高. 具体地, 当配备所提出的 BPAM 模块时, TG[2]在 CRNN 文本识别器上的识别精度提高了 5.8%.

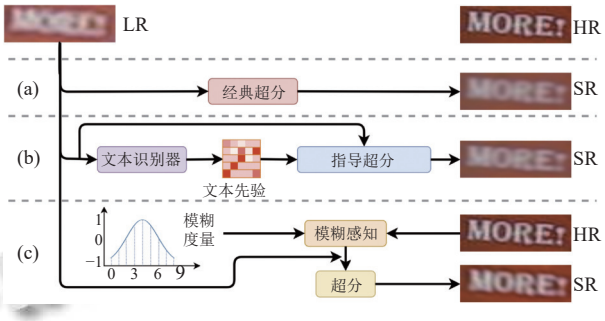

图 1 与先前方法的对比

本文的主要贡献可以总结为如下 3 点.

(1) 本文首次尝试从真实世界的高分辨率-低分辨率 (HR-LR) 场景文本图像对中学习模糊模式, 以便为场景文本图像超分辨率模型生成更多的训练样本, 进一步释放其生成能力.

(2) 本文提出了模糊模式感知模块 (BPAM) 来学习真实世界场景文本图像中所蕴含的模糊模式, 并生成符合真实场景中低分辨率图像分布范式的模糊文本图像.

(3) 本文在公开数据集 TextZoom 上验证了所提出的方法的有效性. 实验结果表明, 现有场景文本图像超分辨率模型配备本文提出的模糊模式感知模块时, 其

性能可以进一步提高.

1 相关工作

1.1 场景文本图像超分辨率

单图像超分辨率 (single image super-resolution, SISR) 的目的是从低分辨率图像重建高分辨率图像, 并保持高分辨率图像具有丰富的细节信息. 在早期的工作中, 研究人员结合了人工特征来提高重建图像的质量. 近年来, 基于深度学习的模型在 SISR 中发挥着重要作用. SRCNN 模型[11]是深度学习在超分辨率重建领域的开山之作. EDSR[12]受到 ResNet 的启发, 采用了残差结构, 并且在标准的残差结构中去除了批量归一化层, 节省了内存空间的同时, 也避免了批量归一化所导致的对比度、色彩拉伸的问题. RDN[13]则是在 EDSR 的基础上, 进一步地扩大了网络的规模, 引入了密集残差块 (residual dense block, RDB) 使得模型更加充分地获取局部特征信息和全局特征信息. 经典的 RCAN[14]模型则是引入了通道注意力机制, 促进了特征在通道维度进行进一步融合, 提高重要的通道的权重, 舍弃不重要的通道的特征.

与 SISR 相比, 场景文本图像超分辨率更加侧重于利用图像中蕴含的文本语义信息. 因此, 基于场景文本识别器的识别准确度 (recognition accuracy) 是被用来评估场景文本图像超分辨率模型性能优劣的最重要指标. TPGSR[4]利用预先训练的场景文本识别算法获得字符概率序列作为文本先验, 并提出了一个文本先验引导的超分辨率框架来引导模型基于文本先验重构文本区域的细节. STT[5]引入了位置感知模块和内容感知模块, 使得场景文本图像超分辨率模型专注于图像的文本区域. PCAN[15]设计了一个并行的上下文注意力块, 以自适应地选择文本序列中的关键信息来指导模型生成文本图像. TG[2]通过预训练一个基于 Transformer 的文本识别器, 使用其注意力图获取笔画级别的监督来恢复图像中文本区域的笔画细节. TATT[6]通过基于 Transformer 的文本注意力网络提取文本语义, 以应对空间变形的文本图像. 此外, TATT 提出了文本结构一致性损失, 通过在规则文本和变形文本的重建上施加结构一致性来细化视觉外观. C3-STISR[3]使用场景文本识别器的反馈、视觉和语言信息作为线索来引导模型生成更易识别的超分辨率文本图像, 其中语言信息由预先训练好的字符级语言模型进行直接生成.

1.2 场景文本识别

场景文本识别 (scene text recognition, STR) 的目标是将场景文本图像转换为文本序列. 在过去的数年里, 涌现了许多基于深度学习的方法来解决场景文本识别问题. CRNN[16]结合卷积神经网络和循环神经网络作为编码器来提取文本序列特征. 在特征提取之后, 使用 CTC 损失函数 (connectionist temporal classification loss) 用于对齐预测文本序列和目标文本序列. 近年来, 基于注意力的文本识别方法取得了巨大的成功. MORAN[17]和 ASTER[18]使用空间变换网络 (spatial Transformer network, STN) 来校正输入图像, 对于倾斜、弯曲的图像表现良好, 然后使用注意力机制来提高对空间变形文本的鲁棒性. 此外, 在 SEED[19]中使用预先训练的 FastText 来学习文本语义信息. 尽管现有的场景文本识别算法已经取得了不错的性能, 但低分辨率的模糊文本图像仍然对场景文本识别提出了巨大的挑战, 直接将模糊图像输入场景文本识别模型, 往往难以获得令人满意的结果. 因此, 有必要设计一种超分辨率预处理器来提高文本图像的质量.

2 模型方法设计

本文提出了基于模糊模式感知的场景文本图像超分辨率算法, 该方法的框架如图 2 所示. 模型主要由两个模块组成, 模糊模式感知模块 (BPAM) 和场景文本图像超分辨率模块. 基于条件生成对抗网络 (CGAN)[10]的模糊模式感知模块包含生成器G和鉴别器D; 场景文本图像超分辨率模块可以被任何其他场景文本图像超分辨率模型取代. 整体流程可以描述为: 将高分辨率场景文本图像与模糊度量输入模糊模式感知模块, 生成大量与模糊度量相匹配的低分辨率文本图像, 随后将这些生成的低分辨率图像送入场景文本超分辨率模块, 进行统一训练. 模糊模式感知模块作为一个可插拔的模块, 在模型训练完成后可直接去除. 图 2 中使用红色线表示数据流向的模块仅在模型训练阶段参与, 只有黑色线及其关联的模块参与推理测试阶段. 也就是说, 后续的推理测试阶段可直接使用场景文本图像超分辨率模块进行推理, 不会增加超分模型的推理开销.

2.1 模糊度量的注释

在现实生活中, 可以发现相同分辨率的图像通常拥有着不同的模糊程度. 正如第 3.1 节中提到的 Text-Zoom[1]数据集中测试集分为 Easy、Medium、Hard 这

3 个子集. 明显地, Hard 子集中的数据样本的模糊程度远超 Easy 子集. 在一定程度上, 模糊程度决定了该低分辨率场景文本图像能否被识别. 此时, 能否正确地表示图像的模糊程度成为研究的重点.

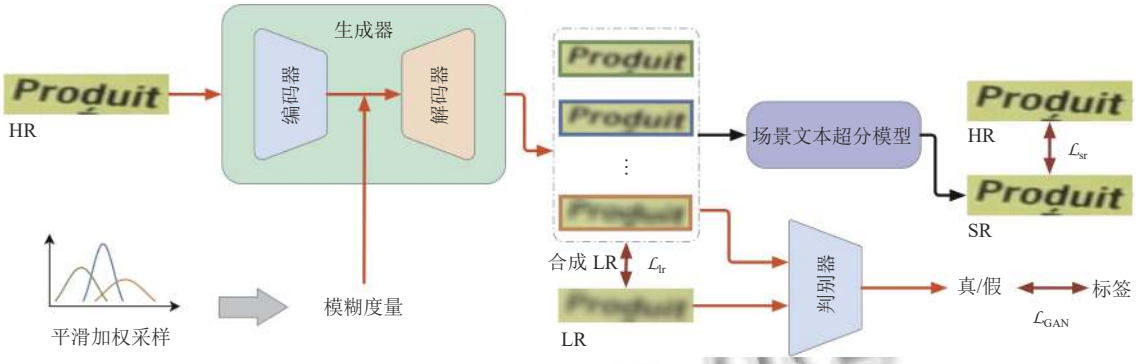

图 2 模型整体网络图

为了衡量图像的模糊程度, 本方法针对 TextZoom 数据集[1]进行重新设计. 将衡量图像的模糊程度的变量称为模糊度量 (degradation degree), 同时, 模糊度量也是设计模糊模式感知模块所必不可少的一部分. 手动注释 TextZoom 训练集中具有相应退化程度的低分辨率文本图像, 退化程度共有 10 个级别, 范围为{0, 1, ⋯, 9}.

具体地, 如图 3 所示, 对于 TextZoom 数据集中的高分辨率文本图像, 使用高斯模糊 (模糊核大小设置为 3) 对其进行模糊处理, 共计 10 次. 模糊 1 次的图像标记为 1, 模糊 2 次的图像标记为 2, ⋯, 模糊 10 次的图像标记为 10. 然后通过计算真实低分辨率文本图像 (TextZoom 训练集中的低分辨率文本图像) 与合成低分辨率图像之间的结构相似性 (*SSIM*) 度量:

$$i = \arg\max_{j}(SSIM(L, L'_j)), \quad j = 0, 1, \cdots, 9 \tag{1}$$

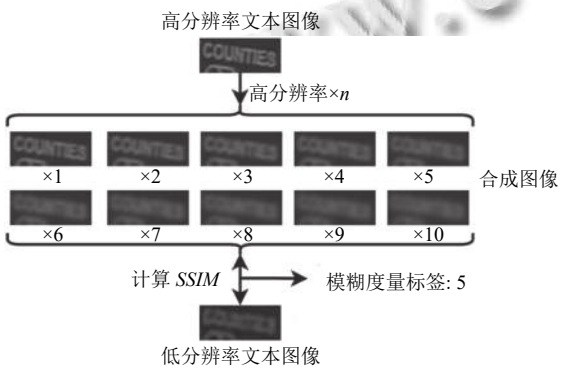

图 3 计算模糊度量的策略

选定与真实低分辨率文本图像具有最高 *SSIM* 的合成低分辨率图像的标记, 减 1 后, 作为真实低分辨率文本图像的模糊度量. 此外, 由于 TextZoom 数据集中的图像对存在像素点错位的问题, 需先挑选出 *SSIM* 计算值异常的图像对, 并一一对其进行人工位置矫正对齐, 以便获得正确的模糊度量值.

2.2 模糊模式感知模块

模糊模式感知模块 (BPAM) 基于条件生成对抗网 (conditional generative adversarial network, CGAN)[10] 设计而成, 包含一个生成器 (generator) 和一个判别器 (discriminator).

在 CGAN 中的生成器, 给定一个输入噪声z和额外信息 y, 之后将两者通过全连接层连接到一起, 作为隐藏层输入. 同样地, 在判别器中输入图像 x 和 额外信息 y 也将连接到一起作为隐藏层输入. 通过式 (2) 的对抗损失函数对生成器与判别器依次进行优化.

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x\mid y)] \\ &+ \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z\mid y)))]\end{aligned} \tag{2}$$

模糊模式感知模块网络架构如图 4 所示. 将高分辨率场景文本图像$H \in \mathbb{R}^{H\times W\times C}$作为输入, 生成器首先通过 5 个卷积模块将其编码, 提取出深度特征表示$F \in \mathbb{R}^{\frac{W}{32}\times C'}$. 具体地, 卷积模块包含两个3×3卷积、批量归一化、ReLU 和最大池化层, 实现逐级降维. 随后将其压缩, 得到压缩后的特征$\tilde{F} \in \mathbb{R}^{\frac{W}{32}\times \tilde{C}}$, 公式表示为:

$$\tilde{F} = \tanh(FW_d) \tag{3}$$

其中, tanh(·)表示激活函数, $W_d \in \mathbb{R}^{C'\times \tilde{C}}$代表可学习的参数. 随后将对应的模糊度量向量$d \in \mathbb{R}^{C^d}$与压缩后的特征$\tilde{F}$进行融合:

$$F=\begin{bmatrix}\text{Concat}(\tilde{F}_1,d)\\ \text{Concat}(\tilde{F}_2,d)\\ \text{Concat}(\tilde{F}_i,d)\\ \vdots\\ \text{Concat}(\tilde{F}_K,d)\end{bmatrix} \tag{4}$$

其中, Concat(·)表示在通道维度上进行拼接融合操作, $F'\in\mathbb{R}^{\frac{W}{32}\times(\tilde{C}+C^d)}$表示融合后的特征表示, K表示$\tilde{F}$的宽度, 例如$W/32$. $\tilde{C}$和C^d分别表示压缩特征$\tilde{F}$和模糊度量d的通道数量. 然后, F'进一步转换为与F相同维度的特征表示$\widehat{F}$:

$$\widehat{F}=\text{ReLU}(F'W_u) \tag{5}$$

其中, ReLU(·)表示 ReLU 激活函数, $W_u\in\mathbb{R}^{(\tilde{C}+C^d)\times C'}$表示可学习的参数矩阵. 在完成特征的融合后, 使用解码器将$\widehat{F}\in\mathbb{R}^{\frac{W}{32}\times C'}$解码生成低分辨率场景文本图像$\widehat{L}$. 具体地, 解码器同样包含 5 个卷积模块, 每个模块拥有两个 卷积、批量归一化、ReLU 以及上采样层, 实现特征的逐级扩张. 在此过程中, 使用类似 U-Net 的跳跃连接辅助浅层特征与深层特征充分融合.

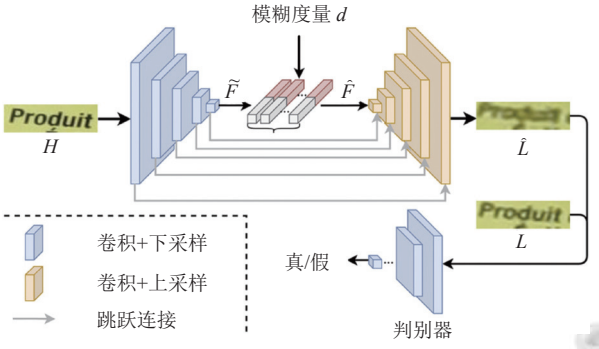

图 4 模糊模式感知模块详细网络图

基于条件对抗生成网络的模糊模式感知模块还需要判别器对生成图像与真实世界低分辨率图像进行识别, 其结构较为简单, 如表 1 所示, 经过 3 次卷积操作降低维度后使用全局平均池化进而将维度变为 1, 判断输入图像的真伪, 1 表示判别为真, 0 则表示判别器判别为假. 最终希望生成器能够生成让判别器难辨真假的低分辨率场景文本图像.

2.3 训练策略

本方法采用两阶段训练策略, 包括热身阶段 (warm-up stage) 与联合训练阶段 (joint-training stage). 具体而言, 首先对模糊模式感知模块进行初步训练, 使其感知到不同的模糊模式, 并将 one-hot 模糊度量作为提示条件输入模糊模式感知模块. 然后, 通过协同训练模糊模式感知模块和场景文本图像超分辨率模块. 在此阶段, 使用平滑加权采样 (smooth weighted sampling) 采样模糊度量向量作为模糊模式感知模块的条件输入, 为场景文本图像超分辨率模块提供具有不同模糊模式的低分辨率文本图像.

表 1 判别器 D 的参数设置

网络模块	超参数
卷积	k=4, s=2, p=1
LeakyReLU	ns=0.2
卷积	k=4, s=2, p=1
归一化	—
LeakyReLU	ns=0.2
卷积	k=4, s=2, p=1
归一化	—
LeakyReLU	ns=0.2
卷积	k=4, s=1, p=1
全局平均池化	—

2.4 损失函数

整体的损失函数设计包含 3 个部分: 像素均方误差损失、对抗损失、超分辨率模块损失.

(1) 像素均方误差损失$\mathcal{L}_{\text{lr}}$. $\mathcal{L}_{\text{lr}}$监督低分辨率场景文本图像的生成, 公式表示为:

$$\mathcal{L}_{\text{lr}}=\left\|L-\widehat{L}\right\|_2^2 \tag{6}$$

其中, L表示真实低分辨率图像, $\widehat{L}$表示模糊模式感知模块生成的低分辨率图像.

(2) 对抗损失$\mathcal{L}_{\text{GAN}}$. 对抗性损失$\mathcal{L}_{\text{GAN}}$用来衡量生成器G与判别器D之间的博弈, 公式表示为:

$$\mathcal{L}_{\text{GAN}}=\frac{1}{N}\sum_{n=1}^{N}\log D(L_n)+\frac{1}{N}\sum_{n=1}^{N}\log(1-D(G(H_n))) \tag{7}$$

其中, N表示样本数量, G表示模糊模式感知模块, D表示判别器, L、H分别表示低分辨率图像与高分辨率图像.

(3) 超分辨率模块损失$\mathcal{L}_{\text{sr}}$. 超分辨率模块损失$\mathcal{L}_{\text{sr}}$由选择的场景文本图像超分辨率模型决定, 一般包含像素均方误差损失以及感知损失等.

于是, 完整的损失函数$\mathcal{L}$如下所示:

$$\mathcal{L}=\begin{cases}\mathcal{L}_{\text{lr}}+\lambda\min\limits_{G}\max\limits_{D}\mathcal{L}_{\text{GAN}}, & \text{启动阶段}\\ \mathcal{L}_{\text{sr}}+\lambda\min\limits_{G}\max\limits_{D}\mathcal{L}_{\text{GAN}}, & \text{联合训练阶段}\end{cases} \tag{8}$$

其中, λ是用来平衡两个损失的超参数, 第 3.3 节中进行了关于超参数λ选择的实验及分析. 启动热身阶段,

损失函数仅包含$\mathcal{L}_{\rm lr}$损失和$\mathcal{L}_{\rm GAN}$损失; 在联合训练阶段, 模糊模式感知模块与场景文本图像超分辨率模块作为一个整体, 此时$\mathcal{L}_{\rm lr}$损失将不再计入总损失中.

3 实验结果分析

本节中, 首先介绍本文训练以及测试使用的数据集, 随后介绍评价指标以及具体的实验设置细节, 最后再详细分析实验结果.

3.1 数据集

本文主要对广泛使用的场景文本图像超分辨率数据集 TextZoom[1]进行训练和评估. TextZoom 中的部分示例如图 5 所示. 此外, 使用合成场景文本识别数据集 SynthText 进行预训练, 并使用 3 个场景文本识别基准数据集 (IC15、IC13 和 CUTE80) 来评估所提出方法的有效性. 在下文中, 较详细地介绍上述 5 个数据集.

Easy 测试子集

Medium 测试子集

Hard 测试子集

图 5 TextZoom 测试子集样例

(1) TextZoom. TextZoom 数据集[1]是从 RealSR 和 SR-RAW 两个数据集中收集而来. TextZoom 数据集包含训练集与测试集两部分, 其中训练集 17 367 对图像及文本标签, 测试集拥有 4 373 对图像及文本标签. 相机拍摄时的焦距一般决定了拍摄图像的清晰度. 如图所示, Wang 等人[1]基于焦距的长短将测试集划分成 3 个不同难度的测试子集, 其中 Easy 测试子集包含 1619 对数据, Medium 测试子集包含 1 411 对数据, Hard 测试子集包含 1 343 对数据. 可以看到相比 Easy 测试子集. HR 图像的大小调整为32×128, LR 图像的大小则调整为16×64. Hard 测试子集中的低分辨率图像更加模糊, 让人难以辨认出图像中的文本内容, 这给场景文本超分辨率模型带来了巨大的挑战性.

(2) SynthText. 在场景文本识别领域, SynthText 数据集是最常用的合成文本图像数据集之一, 该数据集包含大约 7 200 000 张利用自然场景图像与标准字体合成的文本图像. 通常 SynthText 数据集用于文本识别模型的预训练过程中.

(3) IC13. IC13 数据集 (全称 ICDAR13) 是由 ICDAR 在 2013 年比赛中公开的数据集, 其中训练集包括 848 张图像, 而测试集包括 1 095 张图像. 部分图像中拥有特殊字符, 因本文中方法不涉及特殊字符, 故将其删除, 最终测试集包括 1 015 张图像.

(4) IC15. IC15 数据集 (全称 ICDAR15) 是由 ICDAR 在 2015 年比赛中公开的数据集, 其中训练集包括 4 468 张图像, 而测试集包括 2 077 张图像. 该数据集较难, 包含大量不规则文本图像, 如斜向文本、弯曲文本以及带有仿射变换的文本图像等.

(5) CT80. CT80 数据集 (全称 CUTE80) 包含 288 张测试使用的文本图像, 文本图像大多为包含不规则文本的高分辨率图像, 难度较高.

3.2 评价指标

在超分辨率图像的质量方面, 本文使用峰值信噪比 (*PSNR*) 和结构相似性指标 (*SSIM*) 来评估. 与单图像超分辨率任务不同, 场景文本图像超分辨率更关注文本区域是否变得更清晰、更容易识别. 因此, 识别准确率被用作衡量模型性能的最重要指标. 本文遵循以前的方法[1−6]进行公平比较.

3.3 训练设置

本文方法使用 PyTorch 实现. 本文中的所有实验都是在 12 GB 内存的 NVIDIA 2080Ti GPU 上进行的. Batch 大小设置为 16. 联合训练之前, 在 TextZoom[1]中训练 50 个 epoch, 为模糊模式感知模块预热. 在联合训练阶段中, 将学习率设置为场景文本图像超分辨率模型原始学习率的 1/10. 本文使用 4 种 SOTA 方法 (TATT[6]、TPGSR[4]、C3-STISR[3]和 TG[2]) 作为基线模型. 我们使用了 6 种场景文本识别方法, 包括 CRNN[16]、MORAN[17]、ASTER[18]、AutoSTR[20]、SEED[19]和 ABINet[21], 来评估识别的准确性, 并使用他们的官方 PyTorch 代码和发布的预训练模型.

3.4 超参数选择实验

本节讨论关于超参数选择的实验, 方便选取能够发挥模型最佳性能的参数.

首先是关于损失函数的超参数λ的实验, 该参数用于平衡两个损失函数之间的权重关系, 当λ越大时, 模型越关注对抗性损失部分, 反之, 当λ越小时, 模型则越关注于生成的超分辨率图像与高分辨率 GT 图像之间的相似关系. 如表 2 所示, 随着λ逐渐变大时, 模型生成

的文本图像的峰值信噪比与结构相似性指标呈现先上升后下降的态势, 当λ取 0.1 时, 峰值信噪比与结构相似性同时取得最佳的结果. 这也就说明, 此时模型能够生成与真实世界的文本图像最为相似的图像.

表 2 关于超参数λ的选择实验

评价指标	0.01	0.05	0.1	1.0
PSNR (dB)	23.11	22.36	**23.88**	22.60
SSIM ($\times10^{-2}$)	88.10	87.50	**89.14**	84.90

在第 2.2 节中, 高分辨率文本图像输入模糊模式感知模块, 经过 5 次的降维处理后, 在进行模糊度量的特征融合之前进行了特征的压缩操作, 而其压缩后的特征维度是可选的. 过低的维度容易导致特征包含的内容丢失; 而过大的维度则增加了模型的参数, 同时使得模糊度量特征所占的比重过低, 影响其发挥该有的能力, 难以获得更优的效果. 如表 3 所示, 实验选择$\tilde{C}$为{50, 100, 200, 400}, 当$\tilde{C}$设置为 100 时, 峰值信噪比与结构相似性同时取得最佳的结果. 也说明, 此时模型能够生成与真实世界的文本图像最为相似的图像.

表 3 关于超参数$\tilde{C}$的选择实验

评价指标	50	100	200	400
PSNR (dB)	23.42	**23.88**	22.03	21.92
SSIM ($\times10^{-2}$)	87.24	**89.14**	86.08	85.97

3.5 文本识别数据集实验

本文还在 3 个文本识别基准数据集上进行了实验 (即 IC13、IC15 和 CT80), 以验证本文提出的模糊模式感知模块的有效性. 按照前面的方法[5], 我们通过从 3 个基准数据集中选择低分辨率图像重新构建测试数据集.

本文从 3 个基准数据集中挑选 361 张低分辨率图像, 其中 79 张来自 IC13, 216 张来自 IC15, 66 张来自 CT80, 并将所选图像的尺寸调整为16×64. 使用 3 个识别器 (CRNN[16]、MORAN[17]和 ASTER[18]) 来评估场景文本图像超分辨率模型的性能. 根据表 4 中所示的实验结果, 当配备所提出的模糊模式感知模块时, 场景文本图像超分辨率模型的性能得到了进一步提高. 得益于模糊模式感知模块, TG[2]在 CRNN[16]的识别精度上提高了 5.7%, 这表明我们的方法对真实世界的模糊更具鲁棒性.

3.6 TextZoom 数据集实验

在本节中, 使用目前的 4 个 SOTA 方法 (TATT[6]、TPGSR[4] 、C3-STISR[3]以及 TG[2]) 作为基准模型. 使用 3 个场景文本识别模型 (CRNN[16]、MORAN[17] 、ASTER[18]) 从识别率衡量场景文本超分辨率模型的重建质量. 此外, AutoSTR[20]、SEED[19]以及 ABINet[21]作为补充, 进一步比较模型的性能.

表 4 在场景文本识别基准数据集上的比较实验 (%)

算法	BPAM	CRNN[16]	MORAN[17]	ASTER[18]
TATT[6]	—	27.8	**38.3**	36.6
	√	**29.9**	37.1	**38.7**
TPGSR[4]	—	26.1	36.1	**36.8**
	√	**29.8**	**37.8**	36.4
C3-STISR[3]	—	26.1	34.7	35.8
	√	**28.9**	**36.2**	**37.3**
TG[2]	—	28.0	36.5	36.3
	√	**33.7**	**41.6**	**42.6**

现有的场景文本超分辨率任务以 TextZoom 数据集为最重要的测试数据集, 因此本节的对比实验也聚焦于该数据集. 如表 5 所示, 详细对比了低分辨率文本图像 (LR)、高分辨率文本图像 (HR) 以及场景文本图像超分辨率模型 (TATT、TPGSR、C3-STISR 和 TG) 在 6 个场景文本识别模型 (CRNN 、MORAN 、ASTER、AutoSTR、SEED 以及 ABINet) 的表现, 并且对于是否使用模糊模式感知模块 (BPAM) 进行了两遍实验, 共计 72 个实验, 使用相关论文 GitHub 库的官方代码, 重新实现并训练. 综合 Easy、Medium、Hard 这 3 个子数据集以及总体平均识别准确率 (average recognition accuracy) 上的表现, 在使用 BPAM 模块后, 模型的性能普遍获得了明显提升. 例如, 相比基线 (baseline) 模型, 使用 CRNN、MORAN 和 ASTER 识别模型进行评估, 在配备本文所提出的模糊模式感知模块后, SOTA 方法 TG 实现了 5.8%、4.3% 和 3.6% 的提升, 远比 SOTA 模型 (TATT、TG 等) 之间的差距来得巨大. 另外, 在更多的模型 (如 AutoSTR、SEED、ABINet) 上, BPAM 模块表现依然突出.

可视化结果如图 6 所示. 与基线模型相比, 它们可以在 BPAM 的帮助下恢复更多文本区域的细节. 由于 BPAM 可以为 STISR 模型生成具有真实世界模糊的 LR 图像, 因此可以显著提高它们的鲁棒性.

3.7 关于数据增强的比较实验

超分辨率任务的现有增强策略依赖于人工模糊核, 导致生成的图像和真实世界的图像之间存在明显的数据域差异. 与现有的数据增强不同, 本文提出的 BPAM 使用现有的 HR-LR 文本图像对来学习真实场

景中的模糊模式, 并将学习到的模糊模式转移到其他 HR 图像上, 以生成与真实场景相似的 LR 图像. 为了验证所提出的 BPAM 相对于以前的数据增强方法的优越性, 本文进行了相应的实验. 具体地, 利用中值模糊、高斯模糊、归一化模糊和运动模糊来增强 TextZoom 的训练数据集. 如表 6 所示, 当使用数据增强策略时, 基线模型的性能略有提高, 这是由于人工模糊图像和真实世界的 LR 图像之间存在明显的域差异. 相反, 本文提出的 BPAM 模块可以显著提高现有方法的性能.

表 5　在 TextZoom 数据集上的对比实验 (%)

算法	BPAM	CRNN[16]				MORAN[17]				ASTER[18]			
		Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
LR	—	36.4	21.1	21.1	26.8	60.6	37.9	30.8	44.1	67.4	42.4	31.2	48.2
HR	—	76.4	75.1	64.6	72.4	91.2	85.3	74.2	84.1	94.2	87.7	76.2	86.6
TATT[6]	—	62.6	53.4	39.8	52.6	72.5	60.2	43.1	59.5	78.9	63.4	45.4	63.6
	√	**64.2**	**55.8**	**40.8**	**54.3**	**75.5**	**62.5**	**44.5**	**61.8**	**80.1**	**64.6**	**47.4**	**65.1**
TPGSR[4]	—	61.4	49.6	36.2	49.8	74.2	58.9	42.2	59.4	78.4	60.9	43.6	62.1
	√	**63.3**	**52.8**	**39.4**	**52.6**	**75.1**	**61.0**	**44.2**	**61.1**	**79.0**	**62.9**	**45.4**	**63.5**
C3-STISR[3]	—	60.5	50.4	37.0	50.0	69.4	54.8	38.9	55.3	76.5	60.6	42.2	60.8
	√	**62.3**	**51.1**	**38.0**	**51.2**	**73.1**	**59.1**	**41.8**	**59.0**	**78.3**	**61.7**	**44.6**	**62.6**
TG[2]	—	61.2	47.6	35.5	48.9	75.8	57.8	41.4	59.4	77.9	60.2	42.2	61.3
	√	**67.2**	**55.4**	**38.9**	**54.7**	**80.1**	**62.2**	**45.6**	**63.7**	**80.9**	**63.9**	**46.6**	**64.9**

算法	BPAM	AutoSTR[20]				SEED[19]				ABINet[21]			
		Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
LR	—	64.9	43.2	30.1	47.2	67.6	45.4	32.7	49.7	77.0	57.3	42.3	60.0
HR	—	93.6	88.2	77.4	86.9	95.1	88.2	77.6	87.5	98.4	90.4	82.4	90.9
TATT[6]	—	**73.5**	48.5	36.7	54.1	74.2	49.6	36.9	54.8	**77.9**	54.0	41.0	58.9
	√	73.3	**49.3**	**37.5**	**54.6**	**75.6**	**50.1**	**37.9**	**55.8**	77.5	**54.1**	**42.7**	**59.3**
TPGSR[4]	—	75.4	57.0	40.1	58.6	76.1	57.9	40.0	59.1	78.3	60.7	46.2	62.8
	√	**77.2**	**60.0**	**42.2**	**60.9**	**76.7**	**59.1**	**42.5**	**60.5**	**80.7**	**61.9**	**47.3**	**64.4**
C3-STISR[3]	—	71.3	48.5	36.7	53.3	70.5	50.3	36.9	53.7	73.6	54.2	41.7	57.5
	√	**76.7**	**55.1**	**40.4**	**58.6**	**76.0**	**54.8**	**41.3**	**58.5**	**78.9**	**59.1**	**45.6**	**62.3**
TG[2]	—	77.8	61.4	42.7	61.7	78.3	61.5	42.5	61.9	81.2	65.5	46.7	65.5
	√	**81.0**	**66.6**	**46.8**	**65.9**	**81.2**	**64.8**	**47.7**	**65.6**	**83.7**	**66.3**	**52.3**	**68.4**

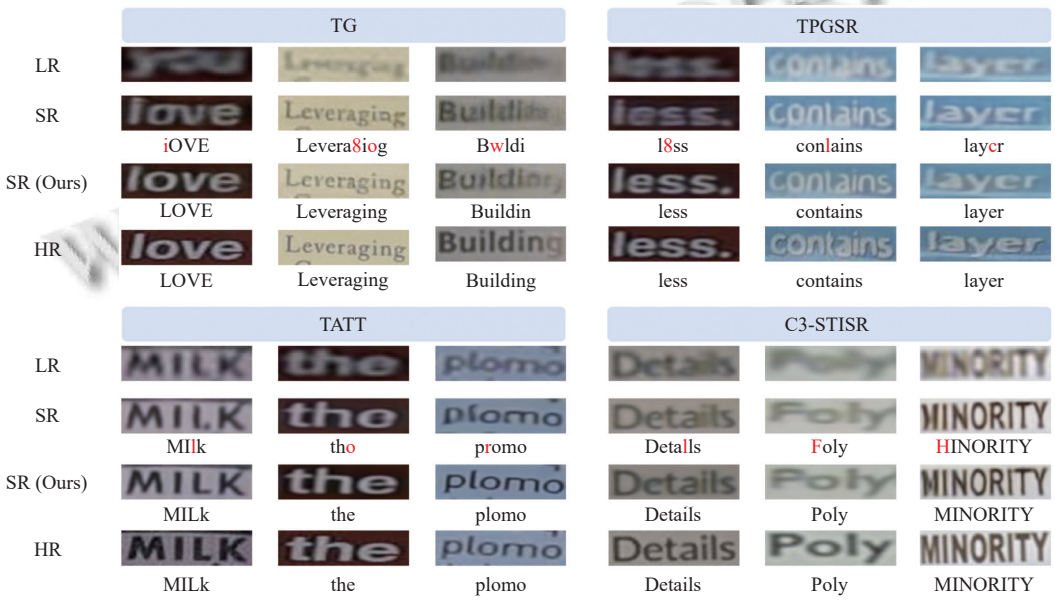

图 6　可视化结果展示

表 6 使用 CRNN[16]评估数据增强的比较实验 (%)

算法	Easy	Medium	Hard	Average
TG[2]	61.2	47.6	35.5	48.9
TG+Aug	62.0	49.5	35.5	49.8
TG+BPAM	67.2	55.4	38.9	54.7
TG+BPAM+Syn	**67.6**	**55.9**	**40.5**	**55.5**

本文继续探究在更多训练数据上模型的表现. TextZoom 只有 17 367 个训练样本, 这对于现有的基于深度学习的场景文本图像超分辨率模型来说是不够的. 我们试图通过引入 SynthText 来扩大训练数据的规模, 进一步提高场景文本超分辨率模型的性能. 如表 6 最后一行 (TG+BPAM+Syn) 结果所示, 在联合训练阶段使用更多的数据可以进一步提高我们方法的性能.

4 讨论

4.1 模糊度量采样策略

在联合训练阶段, 需要对随机退化程度进行采样, 以提示模糊模式感知模块生成相应的 LR 图像. 为此进行了实验, 以选择合适的采样方法. 具体来说, 我们探索了 3 种采样方法: 1) 一次热采样, 2) 平滑加权采样, 3) 随机高斯采样. 此外, 我们尝试对每个退化程度进行随机高斯采样, 作为 BPAM 的条件. 表 7 中的实验结果表明, 平滑加权采样实现了最佳性能. 一个可能的原因是, 通过平滑加权采样方法采样的退化程度矢量与预热阶段的退化程度向量相似, 同时保持采样的多样性.

表 7 使用 CRNN[16]评估采样方法的比较实验 (%)

采样方法	Easy	Medium	Hard	Average
One-hot采样	**67.3**	54.7	38.6	54.4
随机高斯采样	66.0	53.9	**40.0**	54.1
平滑加权采样	67.2	**55.4**	38.9	**54.7**

4.2 时间效率

以场景文本超分模型 TG[2]为例, 结果如表 8 所示, 在训练阶段, 相比原 TG 模型, 装备 BPAM 模块的 TG 模型训练一个 epoch 所花费的时间增加了 10% 左右; 推理阶段, 则取得相近的速度. 增加的少许训练开销对于明显的模型性能提升而言是值得的.

表 8 关于训练和推理的时间效率实验

模型	训练 (s/epoch)	推理 (f/s)
TG[2]	251	223
TG+BPAM	277	219

4.3 失败样例

TextZoom[1]中的一些失败样例如图 7 所示. 根据可视化结果, 我们观察到所提出的方法在处理长文本的低分辨率图像时仍然存在困难. 此外, 低分辨率模糊图像中的艺术字体也是一个巨大的挑战, 因为在 TextZoom 的训练集中艺术字体的样本很少.

图 7 TextZoom 数据集中的一些失败样例

5 总结

本文提出了基于条件生成对抗网络的模糊模式感知模块 (BPAM). 为了训练基于 CGAN[10]的模型, 我们在 TextZoom[1]的训练数据集中手动注释每个低分辨率图像的退化程度. 在联合训练阶段, 对不同退化程度进行采样, 为场景文本超分辨率模型生成相应的低分辨率图像, 增强了其对真实世界低分辨率图像的鲁棒性. 实验表明, 本文提出的模糊模式感知模块可以显著提高现有 SOTA 场景文本图像超分辨率模型的性能, 同时在推理中不引入额外的时间成本.

参考文献

1 Wang WJ, Xie EZ, Liu XB, *et al*. Scene text image super-resolution in the wild. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 650–666.

2 Chen JY, Yu HY, Ma JQ, *et al*. Text gestalt: Stroke-aware scene text image super-resolution. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 285–293.

3 Zhao MY, Wang M, Bai F, *et al*. C3-STISR: Scene text image super-resolution with triple clues. Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna: IJCAI, 2022. 1707–1713.

4 Ma JQ, Guo S, Zhang L. Text prior guided scene text image super-resolution. IEEE Transactions on Image Processing, 2023, 32: 1341–1353. [doi: 10.1109/TIP.2023.3237002]

5 Chen JY, Li B, Xue XY. Scene text telescope: Text-focused scene image super-resolution. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12021–12030.

6 Ma JQ, Liang ZT, Zhang L. A text attention network for

spatial deformation robust scene text image super-resolution. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5901–5910.

7 Pandey RK, Vignesh K, Ramakrishnan AG, *et al*. Binary document image super resolution for improved readability and OCR performance. arXiv:1812.02475, 2018.

8 Xu XY, Sun DQ, Pan JS, *et al*. Learning to super-resolve blurry face and text images. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 251–260.

9 Mou YQ, Tan L, Yang H, *et al*. PlugNet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 158–174.

10 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784, 2014.

11 Dong C, Loy CC, He KM, *et al*. Learning a deep convolutional network for image super-resolution. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 184–199.

12 Lim B, Son S, Kim H, *et al*. Enhanced deep residual networks for single image super-resolution. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1132–1140.

13 Zhang YL, Tian YP, Kong Y, *et al*. Residual dense network for image super-resolution. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2472–2481.

14 Zhang YL, Li KP, Li K, *et al*. Image super-resolution using very deep residual channel attention networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 294–310.

15 Zhao CR, Feng SY, Zhao BN, *et al*. Scene text image super-resolution via parallelly contextual attention network. Proceedings of the 29th ACM International Conference on Multimedia. ACM, 2021. 2908–2917.

16 Shi BG, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(11): 2298–2304. [doi: 10.1109/TPAMI.2016.2646371]

17 Luo CJ, Jin LW, Sun ZH. MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recognition, 2019, 90: 109–118. [doi: 10.1016/j.patcog.2019.01.020]

18 Shi BG, Yang MK, Wang XG, *et al*. ASTER: An attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2035–2048. [doi: 10.1109/TPAMI.2018.2848939]

19 Qiao Z, Zhou Y, Yang DB, *et al*. SEED: Semantics enhanced encoder-decoder framework for scene text recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13525–13534s.

20 Zhang H, Yao QM, Yang MK, *et al*. AutoSTR: Efficient backbone search for scene text recognition. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 751–767.

21 Fang SC, Xie HT, Wang YX, *et al*. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7094–7103.

(校对责编: 牛欣悦)