基于文本语义指导的自然场景文本图像超分辨方法

习晨晨[1]，何　昕[1]，孟雅蕾[2]，张凯兵[2]
（1.西安工程大学电子信息学院，西安，710048；2.西安工程大学计算机科学学院，西安，710048）

摘要　在自然场景文本图像超分辨中，针对先验信息利用不准确、不充分以及文本边缘恢复不完整的问题，提出了一种基于文本语义指导的自然场景文本图像超分辨方法。该网络结构由超分辨重建模块和文本语义感知模块组成。为进一步提高超分辨网络的表达能力，提出使用循环十字交叉注意力，捕获全局上下文信息，使得模型在训练的过程中更加关注文本区域，同时，提出软边缘损失、梯度损失对重建过程进行约束，生成具有锐利边缘的超分辨结果。采用公开的自然场景文本图像超分辨数据集 TextZoom 对提出模型的性能进行验证，与 8 种主流深度网络模型进行了对比，结果表明：该模型在 3 个不同识别器下的平均识别率相比 TSRN 分别提升了 2.06%、1.80%和 2.89%，在 PSNR 和 SSIM 指标上也具有一定的优势。

关键词　场景文本图像超分辨；文本语义；注意力机制；软边缘损失；梯度损失
DOI　10.3969/j.issn.2097-1915.2023.06.013
中图分类号　TP391.41　　**文献标志码**　A　　**文章编号**　2097-1915(2023)06-0095-09

A Scene Text Image Super-Resolution Method Guided by Text Semantics in Wild

XI Chenchen[1], HE Xin[1], MENG Yalei[2], ZHANG Kaibing[2]
(1. School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China;
2. School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract　Aimed at the problems that in scene text image super-resolution, prior information is inaccurate and insufficient in utilization and text edge is incomplete in recovery, a scene text image super-resolution method guided by text semantics is proposed. This network structure is composed of a super-resolution reconstruction module and a text semantic-aware module. To further improve the expression ability of the super-resolution network, a recurrent crisscross attention mechanism is used to capture global contextual information, making the model pay more attention to the text region during training. And simultaneously, in order to generate sharp edges, a soft-edge loss and a gradient loss are proposed to constrain the reconstruction process. The performance of the proposed model is verified on the public scene text image super-resolution dataset TextZoom with eight mainstream deep network models. Compared with TSRN, the average recognition accuracy of the proposed model is promoted to 2.06%, 1.80%, and 2.89% by three different recognizers respectively, and the proposed model also has advantages in PSNR and SSIM indicators.

Key words　scene text image super-resolution; text semantic; attention mechanism; soft-edge loss; gradient loss

收稿日期：2023-03-31
基金项目：国家自然科学基金(61971339)
作者简介：习晨晨(1997－)，男，陕西乾县人，硕士生，研究方向为图像超分辨等。E-mail：xichenchen0423@163.com
通信作者：张凯兵(1975－)，男，湖北孝感人，教授，博士生导师，研究方向为图像增强、检测与识别等。E-mail：zhangkaibing@xpu.edu.cn

引用格式：习晨晨，何昕，孟雅蕾，等. 基于文本语义指导的自然场景文本图像超分辨方法[J]. 空军工程大学学报，2023，24(6)：95-103. XI Chenchen, HE Xin, MENG Yalei, et al. A Scene Text Image Super-Resolution Method Guided by Text Semantics in Wild [J]. Journal of Air Force Engineering University, 2023, 24(6): 95-103.

文本图像作为一种特殊图像存在于人们的生活当中，人类大脑时刻在对看到的场景进行分析，并根据场景中的文字指导行为。但是受环境、设备等因素的影响，采集的文本图像往往存在模糊、失真等低质量的情况。因此如何正确提取低质量文本图像中的信息来获得更高质量的图像已经成为一个日益紧迫的问题。文本图像超分辨重建技术应运而生[1-2]。文本图像超分辨重建技术已经在交通安全监控、笔迹识别、证件识别、自动驾驶以及书法文物保护与恢复等领域具有极大的应用价值。

相比于规整的扫描文档图像，自然场景中拍摄的图像所包含的文本有水平、倾斜甚至弯曲的文字，而且受制于硬件设备、摄像机抖动、相机与目标对象间的相对运动等拍摄条件的限制导致图像存在不同程度的模糊、昏暗或者分辨率低等情况，多种因素表明自然场景文本图像超分辨(scene text image super-resolution，STISR)非常困难。近年来，随着深度学习技术的快速发展，基于深度学习的自然场景文本图像超分辨技术克服了传统方法复杂度高、泛化性差且需要较多的先验信息等的局限性，取得令人瞩目的成就。Wang 等[3]引入条件生成对抗网络(conditional generative adversarial networks，cGAN)来重建 STISR，去除了 cGAN 中的批归一化(batch normalization，BN)层，引入了 Inception 结构，有效扩展了网络的宽度，使生成器能自适应地捕捉图像中不同大小的文本线索，更适合 STISR 重建任务。Xue 等[4]采用残差密集网络(residual in residual dense network，RRDN)提取比普通残差网络更深层的高频特征，并利用注意力机制增强空间和通道特征，同时引入了梯度损失监督网络训练，以获取更加清晰的文本边缘，该方法在 STISR 任务上取得了不错的结果。Zhang 等[5]设计了一种不需要预训练的 STISR 重建网络，该网络主要由卷积层、BN 层、LeakyReLU 激活层以及上采样层和下采样层组成，利用深度图像先验(deep image prior，DIP)的特点，设计了一种新的加权 MSE 损失函数来突出文本图像的高频细节。

2021 年，Fang 等[6]提出文本超分辨生成对抗网络(text super-resolution generative adversarial networks，TSRGAN)，引入生成对抗网络来防止网络产生过平滑图像，同时加入三元组注意力机制提高网络的表征能力，并引入小波损失来重构更清晰的边缘。Honda 等[7]提出了一种基于多任务学习的 STISR 网络(multi-task super-resolution，MTSR)，该网络使用了 2 个并行任务：图像重建和图像超分辨(super-resolution，SR)，将重建模块和 SR 模块的特征进行融合然后送入下一层进行迭代，使 SR 网络能够学习到重建任务中所提取的特征，最后得到一个训练完备的 STISR 模型，获得不错的重建效果。但上述方法缺少先验信息的利用，导致恢复图像缺少细节信息，不能达到令人满意的效果。

本文受文本先导超分辨(text-prior guided super-resolution，TPGSR)网络[8]启发，以文本超分辨网络(text super-resolution network，TSRN[13])为基础，从先验信息利用和损失函数 2 个角度考虑自然场景文本图像超分辨任务，提出了一个新的文本语义指导的超分辨网络(text-semantic guided super-resolution network，TSGSRN)。针对 TPGSR 方法中使用低分辨文本先验指导网络训练导致先验信息利用不准确的问题，本文提出使用预训练语义感知网络建立 SR 图像和真实高分辨(high resolution，HR)图像之间的文本语义监督，以有效提高网络模型对文本字符的语义理解能力。除此之外，针对现有的十字交叉注意力机制只关注局部特征的问题，本文使用循环十字交叉注意力[9]，提升远距离像素之间的相关性，更好地融合周围像素的上下文信息，从而捕获全局信息。最后，考虑到现有方法使用边缘检测算子提取边缘导致的边缘特征丢失问题，采用软边缘损失和梯度损失对重建结果进行优化。在相同的实验条件下，提出的 TSGSRN 能获得比现有方法更好的质量评价指标[10-11]。

1 TSGSRN 整体框架

本文提出的 TSGSRN 的整体框架如图 1 所示，由超分辨重建模块和文本语义感知模块组成。

超分辨重建模块以 LR 图像及其二进制掩码图作为输入。其中，LR 图像为 RGB 图像，二进制掩码图为二值图(文字区域置为 1，背景区域置为 0)。首先，网络的输入经过中心对齐网络进行对齐，然后通过单个卷积层提取特征；其次，通过 7 个相同的超分辨残差块；最后，使用 Pixel-Shuffle 对处理后的特征映射进行上采样，以生成 SR 结果，并通过 L_2 损失、梯度损失和软边缘损失计算重建图像和真实图像之间的差异。文本语义感知模块通过预训练识别

网络建立 SR 图像和 HR 图像之间的字符类别概率分布差异，获得更多面向文本的信息。相比于 TSRN，TSGSRN 有以下改进：①使用预先训练的语义感知网络感知文本自身的语义信息，使得模型具有更好的语义理解能力；②TSGSRN 在每个超分辨残差块中加入了注意力机制进一步提升超分辨效果；③使用软边缘损失对生成图像的边缘进行约束，得到边缘更准确、清晰的超分辨结果。

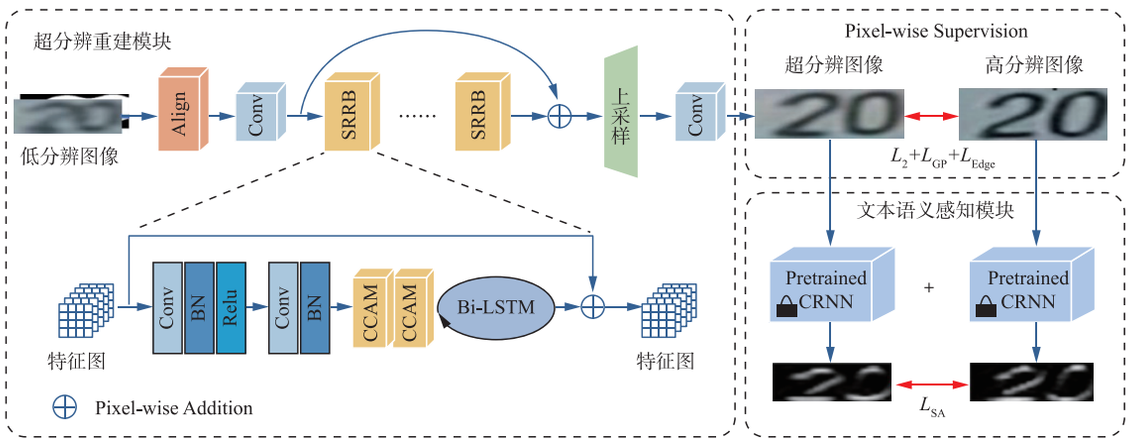


图 1　基于文本语义指导的自然场景文本图像超分辨方法整体框架

2　TSGSRN 设计

2.1　超分辨重建模块

SR 重建模块主要由对齐模块、基于残差网络的重建主体、后上采样模块组成。首先，LR 文本图像及其二进制掩码图像作为输入，送入到对齐网络中，使得输入的 LR 图像与真实的 HR 图像具有中心对齐的效果，以减小数据本身存在的像素误差。对齐网络采用薄板样条变换（thin plate spline，TPS），对齐过程可以表示为：

$$F_{\text{in}}=f_{\text{TPS}}(I_{\text{LR}}) \tag{1}$$

式中：f_{TPS} 表示薄板样条变换；F_{in} 表示对齐网络的输出特征。然后，输出的特征经过一个卷积核大小为 9×9 的卷积和 PRelu 激活函数，表示为：

$$F_{\text{in}}^{m}=f_{\text{c}}(F_{\text{in}}) \tag{2}$$

式中：f_{c} 表示卷积；F_{in}^{m} 表示卷积层的输出特征，再将其输入到重建主体网络当中，重建主体网络采用残差结构，将底层的结构信息送到网络的更深层，不仅充分利用了底层特征，而且避免了网络发生过拟合。过程表示为：

$$F_{\text{out}}^{n}=f_{\text{SRRB}}^{(1\sim7)}(F_{\text{in}}^{m}) \tag{3}$$

式中：$f_{\text{SRRB}}^{(1\sim7)}$ 表示由 7 个相同的超分辨残差块（super-resolution residual block，SRRB）组成的残差网络；F_{out}^{n} 表示主体网络输出的特征映射。最后，该特征映射经过后上采样模块进行 2 倍放大，获得最终的 SR 重建图像，过程表示为：

$$I_{\text{SR}}=f_{\text{up}}(F_{\text{out}}^{n}) \tag{4}$$

式中：f_{up} 表示 2 倍上采样操作；I_{SR} 表示整个超分辨重建模块的输出结果。

2.2　文本语义感知模块

为了使得网络能够充分理解文本的内容信息，具有更好的感知能力，本文提出文本语义感知模块见图 2，为文本语义感知（semantic-aware，SA）模块的内部结构。

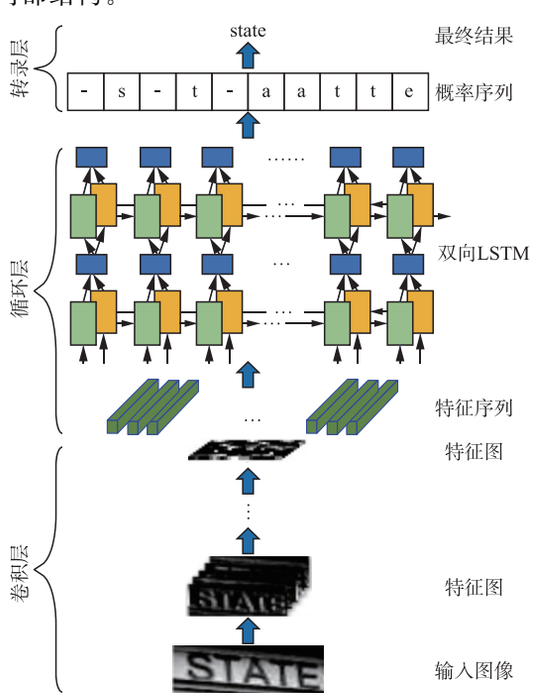


图 2　文本语义感知模块结构

SA 模块使用 CRNN 网络[12]，该网络结构包含 3 个部分：卷积层、循环层和转录层。卷积层使用卷积神经网络（convolutional neural network，CNN），从输入图像中提取图像特征；循环层使用循环神经网络（recurrent neural network，RNN），对图像特征

的语义信息进行建模，用来预测从卷积层获取的特征序列的标签分布；转录层使用 CTC 损失使得预测序列更准确地与目标序列对齐，把从循环层获取的标签分布去重整合得到最终的分类文本先验。

SR 重建模块得到的 SR 图像 I_{SR} 和真实的 HR 图像分别送入 CRNN 网络中，以 SR 图像为例：首先经过 6 个卷积层，得到卷积层的输出特征：

$$F_{CNN}=f_{CNN}^{(1\sim6)}(F_{out}) \tag{5}$$

然后，特征 F_{CNN} 送入循环层，循环层使用双向长短时记忆网络，根据输入的特征进行预测，得到所有字符的 SoftMax 概率分布，该分布是长度为字符类别数，高度为字母表 a～z 和数字表 0～9 的向量。将该分布送入第 3 部分转录层，使用 CTC 损失使得预测序列更准确地与目标序列对齐，把从循环层获取的标签分布去重整合得到最终的分类文本先验，如图 3 所示。白点越明显，表示属于该类别的概率越高；越模糊，表示属于该类别的概率越低。

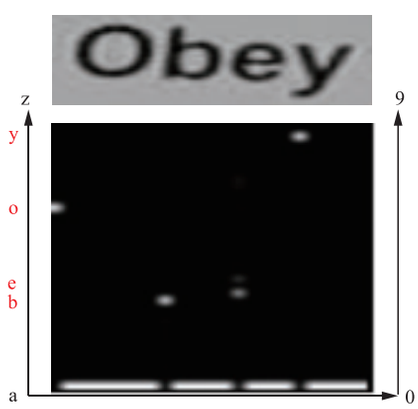


图 3 字符分类概率图

2.3 循环十字交叉注意力

随着注意力机制被提出，超分辨任务也取得了进一步的发展。通道注意力首先被提出，其旨在建立不同通道之间的相关性，通过对每个通道的特征赋予不同的权重，从而强化重要特征，抑制非重要特征，更关注于全局特征；空间注意力旨在增强关键区域的特征表达，通过对空间中每个位置生成权重掩膜进行加权，增强感兴趣区域表达，弱化无关的背景区域；三元组注意力通过利用三分支结构实现跨维交互，建立维度间的依赖关系；坐标注意力则是将位置信息嵌入到通道中，分别沿 2 个方向聚合特征，可以在一个空间方向上捕获远程依赖关系，同时在另一个空间方向上保存精确的位置信息，其只能捕获某一个坐标的信息，不能捕获周围相邻像素的信息，而循环十字交叉注意力通过级联 2 个相同的十字交叉注意力，更好地融合全局上下文信息。

十字交叉注意力结构如图 4 所示，对于输入特征 X，首先使用 3 个不同的 1×1 卷积核获取注意力模型中的 Q，K，V；通过 Q 和 K 来获取当前像素下横向和纵向像素点之间的相关性。最后将相关性矩阵与 V 整合，再加上原始的特征 X，得到最终的注意力特征 X'，但是该注意力只计算了“十字”结构中像素点的相关性，对于周围的像素点未遍历，只关注到局部特征。因此，通过级联双层的十字交叉注意力可对周围像素点进行遍历，从而融合全局上下文信息。循环十字交叉注意力在语义分割任务中已经取得不错的效果。由于文本超分辨的目的是增强文字区域，弱化背景区域，因此该注意力可应用于文本超分辨任务。

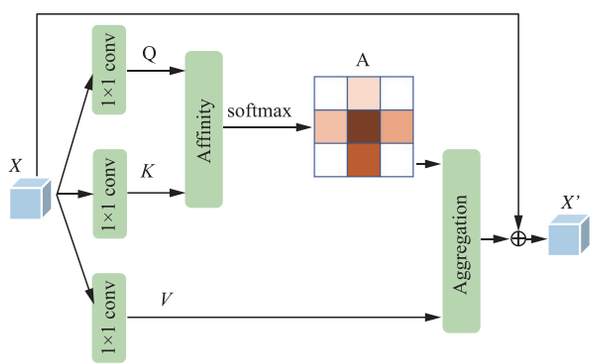


图 4 十字交叉注意力

2.4 损失函数设计

在 SR 任务中，联合不同损失函数对网络模型进行优化，有利于加快网络训练时的收敛速度，从而提升模型的重建性能。因此，本文将像素损失、梯度损失、软边缘损失和文本语义感知损失联合起来共同训练所提出的深度网络。本文方法使用的损失函数如下：

1)像素损失。像素损失表示 SR 结果和目标图像之间的曼哈顿距离，相比于 L_1 损失，L_2 损失有利于恢复清晰的边缘，提高模型收敛速度。因此，本文采用 L_2 损失度量重建图像与目标图像之间的误差。像素损失表示为：

$$L_{pixel}=\frac{1}{N}\sum_{n=1}^{N}\|I_{SR}-I_{HR}\|_2^2 \tag{6}$$

式中：I_{SR} 为 SR 图像；I_{HR} 为真实的 HR 图像。

2)梯度损失。图 5(a)、(b)和(c)分别表示 LR、SR 和 HR 图像，图 5(d)、(e)和(f)分别表示其梯度图。可以看出，LR 图像的梯度场为矮胖型，而 HR 图像的梯度场为高瘦型，为了减小 SR 图像和真实 HR 图像之间的梯度分布差异，引入梯度损失，从而进一步减小 SR 图像和真实 HR 图像之间的差异，表达式为：

$$L_{grad}=\|\nabla I_{SR}-\nabla I_{HR}\|_1 \tag{7}$$

式中：∇表示梯度操作。

3)软边缘损失。为了保证恢复图像的边缘完整

性，本文直接通过软边缘损失对 SR 图像和 HR 图像进行监督，表达式为：

$$L_{\text{edge}} = \| I_{\text{SR}}^{\text{edge}} - I_{\text{HR}}^{\text{edge}} \|_1 \tag{8}$$

式中：$I_{\text{SR}}^{\text{edge}} = \text{div}(u_x, u_y)$ 和 $I_{\text{HR}}^{\text{edge}} = \text{div}(u_x, u_y)$ 分别表示 SR 图像和 HR 图像的软边缘特征，div 表示散度操作。$u_i = \dfrac{\nabla_i I_{\text{HR}}}{\sqrt{1 + |\nabla I_{\text{HR}}|^2}}$，$i \in \{x, y\}$，$x$ 和 y 分别表示水平和垂直方向，∇表示梯度操作。

4）文本语义感知损失。由于 CRNN 中的 CNN 的浅层特征和深层特征分别关注局部结构信息和全局语义信息，因此，文本语义感知损失可以同时保证低级笔画结构和高级文本上下文之间的一致性。相比于一般的自然图像超分辨方法侧重图像的局部细节，对文本语义和字符的形状理解不佳，因此，从预训练的文本语义感知模型中可以获得更多面向文本的信息，它可以更好地衡量 SR 图像和 HR 图像中前景字符之间的相似性，表达式为：

$$L_{\text{tsa}} = \lambda_1 | t_{\text{SR}} - t_{\text{HR}} | + \lambda_2 D_{\text{KL}}(t_{\text{SR}} \| t_{\text{HR}}) \tag{9}$$

式中：t_{SR} 和 t_{HR} 分别表示 SR 图像和 HR 图像的语义类别概率；$|\cdot|$表示 L_1 范数；D_{KL} 表示 KL 散度操作；λ_1 和 λ_2 为很小的常数，均设置为 1.0。本文联合以上 4 个损失对网络模型参数进行优化，整个网络的损失函数表示为：

$$L = \alpha L_{\text{pixel}} + \beta L_{\text{grad}} + \gamma L_{\text{edge}} + \lambda L_{\text{tsa}} \tag{10}$$

式中：α，β，γ，λ 为用于平衡 4 个损失的权衡因子。本文将权重分别设置为：20、0.1、0.1 和 0.1。

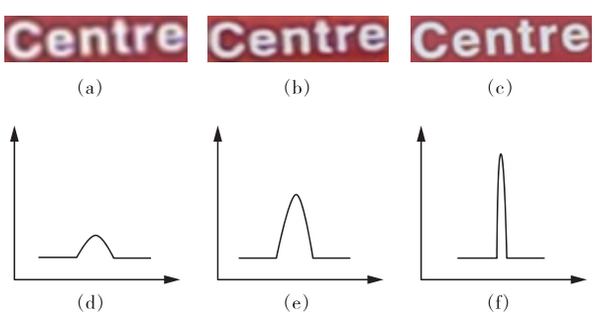
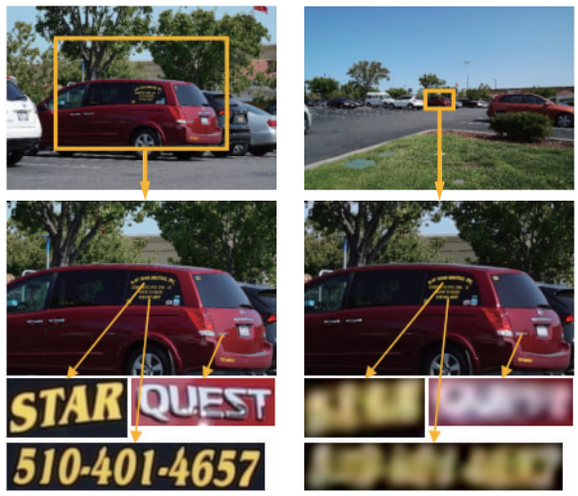


图 5　低分辨、超分辨和高分辨图像及其对应的梯度图

3　实验结果与分析

3.1　实现细节

本文方法使用 WANG 等[13]提出的 TextZoom 数据集进行训练和测试，该数据集是从 CAI 等[14]提出的 RealSR 和 ZHANG 等[15]提出的 SRRAW 中裁剪得到。该数据集是第一个用于自然场景文本图像超分辨任务的数据集，由相机在不同焦距的真实场景中捕获（如图 6 所示），其包含 LR-HR 图像对，但由于人为抖动等原因，存在像素不对齐问题。



(a)150 mm　　(b)24 mm

图 6　不同焦距捕获的低分辨和高分辨图像

TextZoom 数据集中 18 986 张图像用于训练，4 373 张用于测试。测试集根据恢复难易程度分为 3 个等级：easy，medium 和 hard（如图 7 所示）。Easy 包含 1 619 张图像，medium 包含 1 411 张图像，hard 包含 1 343 张图像。与合成的文本数据集的不同之处在于，该数据集的 LR 图像不是经过对 HR 图像下采样获得。并且 TextZoom 数据集在真实场景中经历了复杂的退化，这使得 SR 模型难以恢复高质量的文本图像。低分辨图像大小为 16×64，HR 图像大小为 32×128。本算法模型基于 Pytorch 平台实现，GPU 使用 Nvidia 2080Ti，学习率设置为 0.001。

(a)easy

(b)medinm

(c)hard

图 7　TextZoom 数据集 3 个测试子集

3.2　对比结果与分析

为了验证本文提出方法的有效性，在公共的自然场景文本超分辨数据集 TextZoom 上进行了验证实验。本文方法对比了 8 种主流的超分辨方法：BICUBIC[16]、SRCNN[17]、SRResNet[18]、RDN[19]、VDSR[20]、LapSRN[21]、TSRN[13]、TSRGAN[6]。在 TextZoom 数据集上进行 2 倍放大的识别率评定结果如表 1 所示。ASTER，MORAN 和 CRNN 为常用的 3 种文本识别器。ASTER 由矫正网络和识别网络组成，矫正网络使用 TPS，识别网络是一种加入注意力机制的序列-序列模型，对矫正后的图像进行字符预测；MORAN 由矫正子网络 MORN 和识

别子网络 ASRN 组成，针对弯曲等不规则文本图像具有较好的识别效果；CRNN 的详细介绍见 2.2 节。表 1 中，average 为 3 个测试子集识别率的加权平均值，由于 3 个测试子集数量分别为 1 619，1 411 和 1 343，因此将权重分别设置为 0.37，0.32 和 0.31。PSNR[22] 和 SSIM[23] 指标的定量评定结果如表 2 所示。在表中最优值均加粗表示。（注：由于 TSRGAN 方法源码未公开，所有数据均摘录于原论文）

表 1 识别率对比实验结果

方法	ASTER				MORAN				CRNN			
	easy	medium	hard	average	easy	medium	hard	average	easy	medium	hard	average
BICUBIC	64.11	40.96	31.35	46.55	55.90	35.22	28.07	40.66	36.38	21.12	21.07	26.75
SRCNN	67.26	43.80	33.06	49.15	59.36	36.64	29.64	42.88	40.89	21.19	21.30	28.51
SRResNet	66.58	49.82	34.33	51.22	59.91	42.88	30.53	45.35	42.37	29.13	23.68	32.34
RDN	66.65	47.91	32.91	50.19	60.72	40.11	30.16	44.65	42.74	27.29	23.53	31.84
VDSR	69.24	48.62	34.62	51.91	63.13	42.10	31.35	46.55	43.17	27.85	24.42	32.46
LapSRN	72.82	49.82	35.44	53.87	65.47	44.15	32.46	48.42	47.50	30.33	25.32	35.13
TSRN	73.32	56.27	39.24	57.30	67.03	52.30	37.02	53.01	54.73	41.03	31.50	43.15
TSRGAN	75.70	57.30	40.90	59.02	**72.00**	54.60	39.30	**56.30**	56.20	42.50	32.80	44.56
本文方法	74.49	**59.11**	**41.55**	**59.36**	68.19	**55.07**	38.57	54.81	**56.95**	**45.36**	**33.73**	**46.04**

在所有的比较方法中，前 6 种方法为一般图像超分辨方法，没有加入任何的图像先验信息，受模型性能制约，效果较差；TSRN 使用梯度损失加强边缘的构建，效果略有提升；TSRGAN 在 TSRN 基础上增加对抗损失和小波损失，进一步提升了超分辨效果；本文方法在 TSRN 基础上加入文本语义先验和软边缘损失，识别率进一步提升。从表 1 可以看出，本文方法在 3 个识别器上的平均识别率相比于 TSRN 分别提升了 2.06%、1.80% 和 2.89%。在 ASTER 和 CRNN 识别器上的平均识别率相比于 TSRGAN 分别提高了 0.34% 和 1.48%。在 MORAN 上的平均识别率却稍低于 TSRGAN。

表 2 PSNR 和 SSIM 指标对比实验结果

方法	PSNR/dB			SSIM/dB		
	easy	medium	hard	easy	medium	hard
BICUBIC	22.31	19.04	19.44	0.788 3	0.625 9	0.659 1
SRCNN	22.92	18.99	19.68	0.817 7	0.635 1	0.683 1
SRResNet	21.89	19.03	19.74	0.833 3	0.643 8	0.709 7
RDN	22.86	18.99	19.68	0.841 1	0.647 8	0.714 3
VDSR	23.69	19.12	19.64	0.833 1	0.646 9	0.699 4
LapSRN	23.89	19.11	19.98	0.845 4	0.652 1	0.710 4
TSRN	23.45	18.82	19.69	0.866 8	0.657 5	0.731 0
TSRGAN	24.22	19.17	19.99	0.879 1	0.677 0	0.742 0
本文方法	23.92	19.16	19.91	0.874 9	0.671 8	**0.743 7**

由表 2 可以看出，本文方法相比于 TSRN 在 3 个测试子集的结构相似性（structual similarity，SSIM）指标分别提升了 0.008 1、0.014 3 和 0.012 7；峰值信噪比（peak signal to noise ratio，PSNR）指标分别提升了 0.47、0.34 和 0.22。相比于 TSRGAN 方法，本文方法的 SSIM 指标在测试子集 easy 和 medium 上略低，原因在于 TSRGAN 引入了对抗网络，使得生成的文本图像具有更丰富的细节。

由于 PSNR 指标具有争议性，模糊的图像可能具有较高的 PSNR 值，而清晰的图像可能倾向于表现出较低的 PSNR 值，不一定符合人眼的视觉感知质量，因此，不以 PSNR 指标作为主要评价指标。综上，本文方法相比于其他对比方法表现出了一定的优势。

为了更直观地对比不同 SR 方法的重建性能，图 8 展示了所有对比方法在 TextZoom 数据集上的 SR 重建效果对比。本文选取一些最具有代表性且边缘细节及文字完整性较好的图像进行视觉质量对比。可以看到，方法 SRCNN、SRResNet、RDN、VDSR、LapSRN 和 TSRN 方法的重建结果较为平滑，边缘完整性较差，而本文方法获得的结果均表现出较为完整的字符边缘，这主要得益于模型加入了文本语义信息和软边缘损失。尽管 TSRN 也能够重建出较好效果的图像，但是在细节上仍然存在问题，字符的分离度较差，存在相邻字符之间的粘连问题。其原因在于该网络在训练的过程中只针对边缘结构进行了优化，而缺少文本本身的语义信息参与指导，导致训练得到的模型在重建过程中很难对相邻字符之间的特征进行精准表示。

综上，本文方法在相邻字符的处理上具有一定的优势，且效果逼真，识别错误率最低。此外，本文方法与 TSRGAN 相比在参数量上也有明显的优势。本文提出的基于文本语义指导的 STISR 方法具有较好的重建性能，更适合 STISR 重建任务。

图 8　不同超分辨方法视觉对比结果

3.3　消融实验

1）循环十字交叉注意力。为了验证提出方法使用的循环十字交叉注意力的有效性，对比了几种具有代表性的注意力：通道注意力（CA）[24]、通道-空间注意力（CBAM）[25]、三元组注意力[26]（TAM）和坐标注意力（CoA）[27]，在 3 个测试子集的对比结果如表 3 所示。

表 3　不同注意力的对比实验结果

测试子集	easy			medium			hard		
	Accuracy/%	PSNR/dB	SSIM/dB	Accuracy/%	PSNR/dB	SSIM/dB	Accuracy/%	PSNR/dB	SSIM/dB
CA	**75.11**	23.57	0.856 5	57.41	18.91	0.648 9	40.13	19.79	0.725 4
CBAM	74.24	20.68	0.807 2	60.17	18.21	0.609 5	41.25	18.85	0.673 2
TAM	74.00	23.29	0.862 6	57.76	19.07	0.661 9	**41.77**	19.77	0.733 7
CoA	74.74	23.38	0.858 6	57.41	**19.31**	0.661 9	40.06	**20.15**	0.732 3
RCCA	74.49	**23.92**	**0.874 9**	**59.11**	19.16	**0.671 8**	41.55	19.91	**0.743 7**

由表 3 可见，相比于其它注意力模型，使用的循环十字交叉注意力在 easy 和 medium 测试子集上的识别率、PSNR 和 SSIM 指标具有一定的优势，能显著提升重建图像质量。

2）文本语义感知模块。为了验证 SA 模块的有效性，对该模块进行了消融实验，从定量和定性 2 个层面证明 SA 模块的有效性，定量对比结果如表 4 所示，加入 SA 模块后，在测试集的 3 个子集上的平均识别率、平均 PSNR 和 SSIM 值都高于没有 SA 模块的模型。重建图像的视觉质量对比如图 9 所示。从图 9 可以看出，在 SA 模块的作用下，模型具有较高的字符语义理解能力，字符的完整程度明显较高，与 HR 图像的相似性更高。

表 4　语义感知模块有效性定量对比实验结果

SA 模块	Accuracy/%	PSNR/dB	SSIM/dB
无	57.09	20.61	0.727 0
有	**59.36**	**21.15**	**0.769 2**

(a)无 SA 模块

(b)有 SA 模块

(c)HR 图像

图 9　语义感知模块有效性定性对比实验结果

3)损失函数。为了验证本文方法所用损失函数 的有效性,对其进行了消融实验,如表 5 所示。

表 5 不同损失函数的消融实验对比结果

损失函数		Accuracy /%	PSNR /dB	SSIM/dB
L_{pixel}	easy	73.83	22.06	0.802 3
	medium	57.12	17.81	0.601 1
	hard	39.85	18.79	0.700 2
$L_{pixel}+L_{grad}$	easy	74.49	23.61	0.870 2
	medium	57.83	18.76	0.655 3
	hard	40.47	19.43	0.720 5
$L_{pixel}+L_{grad}+L_{edge}$	easy	74.68	23.56	0.824 3
	medium	58.49	19.27	0.639 8
	hard	40.76	19.65	0.701 0
$L_{pixel}+L_{grad}+L_{edge}+L_{tsa}$	easy	74.49	23.92	0.874 9
	medium	59.11	19.16	0.671 8
	hard	41.55	19.91	0.743 7

由表 5 可以看出,相比于单一的损失函数,联合所有的损失函数能够显著提升模型的重建性能,得到更好的重建效果。表 5 中,第 1 行只使用像素损失,模型的重建效果不理想;第 2 行表示在像素损失的基础上加入梯度损失,可以看出,在 3 个测试子集的识别、PSNR 和 SSIM 指标均有所提高;第 3 行表示在像素损失、梯度损失的基础上加入软边缘损失,可以看出,在 medium 测试子集的识别率提高了 0.66%,在 medium 测试子集上的 PSNR 指标提高了 0.51 dB;第 4 行表示在像素损失、梯度损失和软边缘损失的基础上加入文本语义感知损失,可以看出,在 3 个测试子集的识别率、PSNR 和 SSIM 均有所提高,相比于只使用像素损失的模型,对比指标有大幅度提升。上述实验结果验证了本文提出的 3 个损失函数对模型性能提升均有贡献。

4)SRRB 的数量。此外,还验证了 SRRB 的数量对网络模型重建性能的影响,结果如图 10 和图 11 所示,对于 STISR 任务,并不是越深的网络效果越好,主要在于图像先验信息的引入,由图 10 可以看出,SRRB 数量为 7 时,模型在 3 个测试子集上均具有最好的识别率。

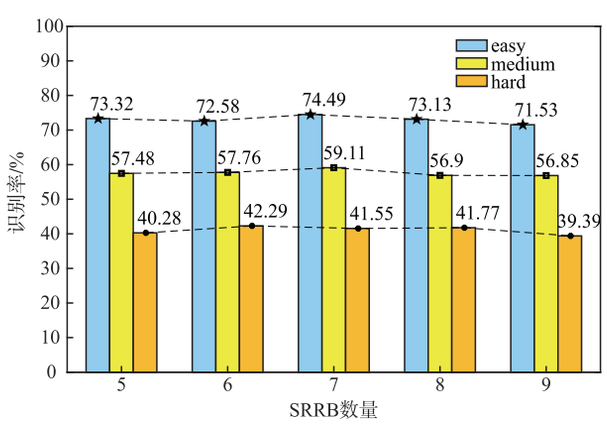
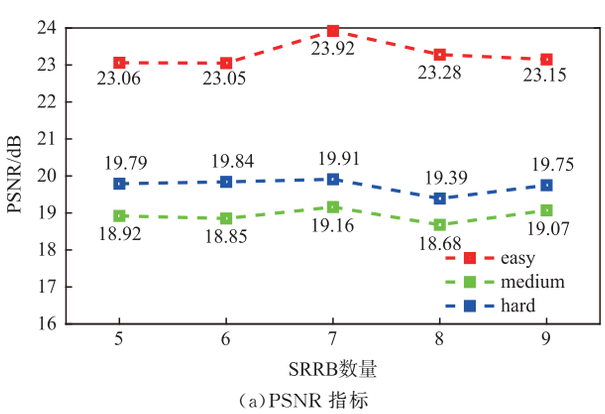
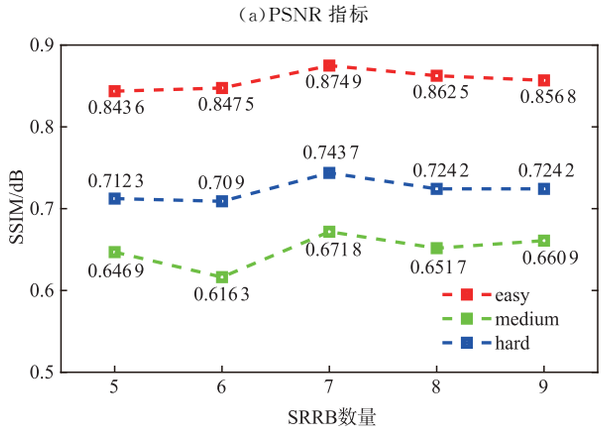


图 10 SRRB 数量的消融实验在识别率上的对比结果

SRRB 的数量对 PSNR 和 SSIM 指标的影响结果如图 11 所示,可以看出,当 SRRB 数量为 7 时,模型具有最佳的 PSNR 和 SSIM 指标。



(a)PSNR 指标



(b)SSIM 指标

图 11 SRRB 数量的消融实验在 PSNR 和 SSIM 指标上的对比结果

4 结语

本文提出了一种基于文本语义指导的 STISR 模型,该模型能够充分利用文本图像的文本语义信息指导超分辨模型训练,通过循环十字交叉注意力提升模型对文本上下文的理解能力,提升有效信息

的表达能力，将更多的注意力放在文字本身。在常用的基准数据集 TextZoom 上的实验结果表明，本文提出的方法在主观和客观质量评价方面都能够获得更好的重建结果，尤其在处理文本字符的粘连问题方面相比于其他方法具有显著优势。

尽管提出的基于文本语义指导的 STISR 重建方法能够获得更好的重建性能，但是仍然存在不足之处。首先，数据集中存在大量模糊图像，模型对其语义理解能力不佳，效果较差；其次，STISR 任务可以视为高频信息恢复后的颜色填充问题，如何只对图像的高频信息进行处理显得尤为重要，是未来需要进一步研究的问题。

参考文献

[1] 杨伟铭，张钰. 基于并行残差卷积网络的图像超分辨重建[J]. 空军工程大学学报(自然科学版)，2019，20(4)：84-89.

[2] 黄淑英，胡瀚洋，杨勇，等. 基于 EM 自注意力残差的图像超分辨率重建网络[J/OL]. 北京航空航天大学学报，2022：1-12. http://doi. org/10. 13700/j. bh. 1001-5965. 2022. 0401.

[3] WANG Y，SU F，QIAN Y. Text-Attentional Conditional Generative Adversarial Network for Super-Resolution of Text Images[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai：IEEE，2019：1024-1029.

[4] XUE M，HUANG Z，LIU R，et al. A Novel Attention Enhanced Residual-in-Residual Dense Network for Text Image Super-Resolution[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen：IEEE，2021：1-6.

[5] ZHANG Q，YE Z，ZHIWEN L，et al. A Text Image Super-Resolution Generation Network without Pretraining [C]//2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC). Piscataway：IEEE，2020：515-519.

[6] FANG C，ZHU Y，LIAO L，et al. TSRGAN：Real-World Text Image Super-Resolution Based on Adversarial Learning and Rriplet Attention[J]. Neurocomputing，2021，455：88-96.

[7] HONDA K，FUJITA H，KUREMATSU M. Improvement of Text Image Super-Resolution Benefiting Multi-Task Learning[C]//Advances and Trends in Artificial Intelligence. 2022：275-286.

[8] MA J，GUO S，ZHANG L. Text Prior Guided Scene Text Image Super-Resolution[J]. IEEE Transactions on Image Processing，2023，32：1341-1353.

[9] HUANG Z，WANG X，HUANG L，et al. CCnet：Criss-Cross Attention for Semantic Segmentation [C]//Proceedings of the IEEE/CVF International Conference on Computervision. Piscataway：IEEE，2019：603-612.

[10] 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 [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(11)：2298-2304.

[11] 张婷悦，张凯兵. 基于稀疏表示的无参考型超分辨图像质量评价方法[J]. 西安工程大学学报，2020，34(5)：20-26.

[12] 朱丹妮，许小华，贺静婧，等. 应用多层感知机回归的无参考型超分辨图像质量评价[J]. 西安工程大学学报，2022，36(5)：70-78.

[13] WANG W，XIE E，LIU X，et al. Scene Text Image Super-Resolution in the Wild[C] //Proceedings of the European Conference on Computer Vision (ECCV). Glasgow：Springer，2020：650-666.

[14] CAI J，ZENG H，YONG H，et al. Toward Real-World Single Image Super-Resolution：A new benchmark and a new model[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Seoul：IEEE，2019：3086-3095.

[15] 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 (CVPR). Long Beach，CA：IEEE，2019：3762-3770.

[16] KEYS R. Cubic Convolution Interpolation for Digital Image Processing[J]. IEEE Transactions on Acoustics，Speech，and Signal Processing，1981，29(6)：1153-1160.

[17] DONG C，LOY C C，HE K，et al. Image Super-Resolution Using Deep Convolutional Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，38(2)：295-307.

[18] LEDIG C，THEIS L，HUSZÁR F，et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI：IEEE，2017：4681-4690.

[19] ZHANG Y，TIAN Y，KONG Y，et al. Residual Dense network for Image Super-Resolution [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City，UT：IEEE，2018：2472-2481.

[20] KIM J，LEE J K，LEE K M. Accurate Image Super-Resolution Using very Deep Convolutional Networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas，NV：IEEE，2016：1646-1654.

(下转第 127 页)

(上接第 103 页)

[21] LAI W S, HUANG J B, AHUJA N, et al. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 624-632.

[22] HUYNH THU Q, GHANBARI M. Scope of Validity of PSNR in Image/Video Quality Assessment[J]. Electronics Letters, 2008, 44(13): 800-801.

[23] WANG Z, BOVVIK A C, SHEIKH H R, et al. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing, 2004, 1(4): 600-612.

[24] ZHANG Y, LI K P, LI K, et al. Image Super-Resolution Using Very Deep Residual Channel Attention Networks[C] //Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 286-301.

[25] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional Block Attention Module[C] //Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 3-19.

[26] MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to Attend: Convolutional Triplet Attention Module[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S. l.]:IEEE,2021: 3139-3148.

[27] XIE C, ZHU H, FEI Y. Deep Coordinate Attention Network for Single Image Super-Resolution[J]. IET Image Processing, 2022, 16(1): 273-284.

(编辑:徐楠楠)