文章编号　1004-924X(2025)01-0135-13

多模态语义交互的文本图像超分辨率重构

韩玉兰*，罗轶宏，崔玉杰，兰朝凤
(哈尔滨理工大学 测控技术与通信工程学院，黑龙江 哈尔滨 150080)

摘要：针对现有方法在文本图像特征表示缺乏尺度变换，分辨率不足导致识别器难以提取到正确的文本内容信息指导重构网络的问题，提出多模态语义交互的文本图像超分辨率重构方法。利用语义推理模块中的注意力掩码对文本内容信息进行校正，获得语义先验信息，约束并指导网络重构语义正确的文本超分辨率重构图像。为增强网络的表征能力，适应不同形状和长度的文本图像，设计了多模态语义交互块，其基本单元由视觉双流集成块、跨模态自适应融合块和正交双向门控循环单元组成。视觉双流集成块利用全局统计特性和局部拟合能力互补优势，获得包含上下文理解的多粒度视觉信息，跨模态自适应融合块动态执行语义信息与多粒度视觉特征之间的交互协作，缩小模态间的特征差异；最后，正交双向门控循环单元建立多模态特征在垂直和水平方向上的文本依赖。实验结果表明，在 TextZoom 测试集上，本文提出的方法在 PSNR 和 SSIM 定量指标上相比于其他主流方法均有所提升，并且在 ASTER，MORAN，CRNN 3 种识别器的平均识别精度相比 TPGSR 模型分别提高了 2.9%，3.6% 和 3.7%。由此表明，采用多模态语义交互方法的文本图像超分辨率重构，可以有效提高文本识别精度。

关　键　词：超分辨率重构；文本图像；多粒度；语义先验；多模态
中图分类号：TP391　　**文献标识码**：A
doi：10.37188/OPE.20253301.0135　　**CSTR**：32169.14.OPE.20253301.0135

Super-resolution reconstruction of text image with multimodal semantic interaction

HAN Yulan*, LUO Yihong, CUI Yujie, LAN Chaofeng

(*College of Measurement and Control Technology and Communication Engineering*, *Harbin University of Science and Technology*, *Harbin* 150080, *China*)
* *Corresponding author*, *E-mail*: *hanyulan@hrbust.edu.cn*

Abstract: The accurate extraction of text content from images is hindered by the absence of scale transformation in feature representation and insufficient resolution, which misguides the reconstruction network. To address this challenge, this paper proposes a novel multi-modal semantic interactive text image super-resolution reconstruction method. By incorporating an attention mask within the semantic inference module, the method corrects text content information and employs semantic prior knowledge to constrain and guide the reconstruction of semantically accurate super-resolution text images. To enhance the network's representational capacity and accommodate text images of varying shapes and lengths, a multimodal se-

收稿日期：2024-07-31；**修订日期**：2024-09-13.
基金项目：国家自然科学基金资助项目(No. 11804068)；黑龙江省自然科学基金资助项目(No. LH2020F033)；黑龙江省省属高等学校基本科研业务资助项目(No. 2020-KYYWF-0342)

mantic interaction block is introduced. This block consists of three key components: a visual dual-flow integration module, a cross-modal adaptive fusion module, and an orthogonal bidirectional gated recurrent unit. First, the visual dual-flow integration module captures multi-granularity visual information, including contextual understanding, by leveraging the complementary strengths of global statistical features and robust local approximations. Next, the cross-modal adaptive fusion module dynamically facilitates interaction and alignment between semantic information and multi-granularity visual features, effectively reducing cross-modal feature discrepancies. Finally, the orthogonal bidirectional gated recurrent unit establishes multimodal feature dependencies in both vertical and horizontal orientations. Experimental results on the TextZoom test set demonstrate that the proposed method outperforms state-of-the-art approaches in terms of quantitative metrics, achieving significant improvements in PSNR and SSIM. Compared to the TPGSR model, the proposed method increases the average recognition accuracy of ASTER, MORAN, and CRNN by 2.9%, 3.6%, and 3.7%, respectively. These findings highlight the effectiveness of multimodal semantic interaction in enhancing text image super-resolution and improving text recognition accuracy.

Key words: super-resolution reconstruction; text image; multi-granularity; feature semantic prior; multimodal

1 引　言

场景文本识别(Scene Text Recognition, STR)在自动驾驶、移动支付、教育和视障人士的服务等领域有着广泛的应用。尤其随着深度学习的发展,STR研究取得了巨大进展。目前,大部分STR算法[1-2]都是以字符形状清晰的高分辨率文本图像为基础和保障,然而受光线、变焦、远距离传输和采集设备等因素的影响,采集到的真实场景图像往往是字符模糊,丢失大量细节信息的低分辨率(Low Resolution,LR)图像,从而严重影响文本识别效果。超分辨率(Super Resolution,SR)重构技术可以有效解决上述问题。

随着卷积神经网络和注意力机制的发展,图像超分辨率重构技术取得了比较瞩目的成就,也进一步推动了超分辨率重构的应用。Dong等[3]将卷积神经网络应用于超分辨率重构提出SRCNN (Super Resolution Convolutional Neural Network)。Niu等[4]提出HAN(Holistic Attention Network),通过引入层注意力模块来提升重构性能。相比于传统方法,基于深度学习的方法[5-7]的重构效果得到了明显提升。然而,这些重构方法往往是面向自然场景的通用模型,主要关注图像整体的清晰度和细节表现,缺乏对特定场景文本内容的处理能力。针对这一问题,场景文本图像超分辨率重构(Scene Text Image Super Resolution, STISR)应运而生,旨在提高LR文本图像分辨率、改善文本图像视觉效果,重构语义正确的文本结构和形状,进而提高下游场景文本识别准确率,STISR目前已成为计算机视觉领域的研究热点[8-9]。

2019年,Wang等[10]提出TextSR (Content-Aware Text Surper-Resolution Network),利用文本感知损失指导网络训练,使模型关注图像的文本信息。Wang等[11]提出CGAN(Text-Attentional Conditional Generative Adversarial Network),在生成对抗网络上结合密集残差连接与通道注意力机制,学习更有效的文本特征表示。Mou等[12]提出PlugNet (Pluggable Super Resolution Unit),引入轻量级的可插拔的超分辨率单元来处理模糊的场景文本图像,降低了网络模型的复杂度和参数量。上述方法虽然性能优异,但是多数以双三次下采样生成的高低分辨率图像对作为训练数据集,人工模糊的文本图像与真实的LR文本图像之间存在域差异,难以推广到复杂的真实场景中。

针对这一问题,2020年Wang等[13]构建了第一个名为TextZoom的真实场景高低分辨率文本图像对数据集,并在此数据集上提出文本超分辨率网络(Text Super Resolution Network,

TSRN)。近年来，研究发现利用先验信息有助于恢复物体形状和纹理，越来越多的研究将各种文本属性作为先验来引导文本图像重构网络。Ma 等[14]提出了文本先验引导超分辨率网络(Text Prior Guided Super Resolution，TPGSR)，将文本的类别信息作为先验，嵌入到重构网络，为模型训练提供引导。Ma 等[15]提出文本注意网络(Text ATTention Network，TATT)，利用 Transformer 将变形的文本图像与文本先验对齐，进一步提高模型性能。Yang 等[16]提出退化先验引导的超分辨率(Degradation Prior Guided Super Resolution，DPGSR)网络，利用设计的退化先验提取器来获取 LR 图像中的文本先验信息，以引导 SR 模块生成可识别的 SR 图像。Ma 等[17]提出了场景文本超分辨率重构网络(Scene Text SR Network，TextSRNet)，利用 Otsu 方法对文本图像进行阈值分割，经卷积网络得到文本图像轮廓先验信息，以获得精细的字符细节。

文本先验信息进一步提高了文本图像的超分辨率重构效果，但多数模型没有充分考虑图像中文本内容带来的语义信息，并且只通过诸如逐元素相加或级联等简单的线性操作与图像视觉特征融合，缺乏自适应模态间的对齐机制，限制了文本先验信息的指导作用。另外，多数研究过于关注文本先验信息而忽略了文本视觉特征提取，均以 SRB (Sequential Residual Blocks)[13]作为特征提取模块，仅利用简单的两个 CNN 层进行特征提取，由于卷积计算的本质局限性，难以捕捉文本图像多粒度上的长程依赖和细微的空间变化，在文本特征表示方面尤其不足，缺乏视觉特征的多粒度表示。针对以上问题，本文联合文本语义和视觉高级语义信息提出一种基于多模态语义交互的文本图像超分辨重构方法(Super resolution of text image with multimodal semantic interaction，MSISR)，通过语义推理模块(Semantic Resoning Module，SRM)进行校正，获得语义正确的文本内容信息，并将此作为先验引导重构网络。提出视觉双流集成块(Visual Dual Flow Integration Module，VDFI)，通过关注特征图不同层次信息来学习远近不同距离的依赖关系，在字符间和字符内的不同粒度上，获得包含多粒度的视觉高级语义特征。提出跨模态自适应融合块(Cross-modal Adaptive Fusion Module，CAFM)，深度挖掘视觉特征和语义先验之间的关联，缩小模态间的特征鸿沟。

2　原　理

2.1　网络整体结构

受 TPGSR 启发，本文提出了 MSISR，其整体框架如图 1 所示，主要由语义先验生成、浅层特征提取、深层特征提取和图像重构 4 个部分组成。与普通的自然图像相比，文本图像中包含文本内容带来的重要信息。针对场景文本图像的这一特点，MSISR 网络利用文本识别器和语义推理模块提取文本的语义信息，并将其作为先验引导重构网络构建深层特征，提高重构图像视觉效果的同时，进一步提高后续文本图像识别的准确率。

2.2　语义先验生成

语义先验生成主要包括文本识别器和语义推理模块两部分。

2.2.1　文本识别器

文本识别器采用预训练的卷积循环神经网络(Convolutional Recurrent Neural Network，CRNN)[18]。与基于注意力机制的文本识别器相比，CRNN 模型简单，在预测字符时考虑背景区域，有助于模型理解字符间的界限。CRNN 的具体结构如图 1 所示，使用 CNN 卷积结构对输入图像提取特征，得到特征图。然后，利用双向 RNN 对特征序列进行预测，并输出预测标签。网络利用转录层 CTC 损失，将获取的一系列标签分布转换成最终的标签序列。

对于输入的低分辨率文本图像 I_{LR}，利用 CRNN 获得的文本识别概率序列：

$$I_{RK}=\mathrm{CRNN}(I_{LR}), \tag{1}$$

其中：CRNN(•)为文本识别器 CRNN 算子。若 CRNN 学习到的字符数为 L，I_{LR} 中字符的个数为 N，那么 I_{RK} 为 $L\times N$ 大小的矩阵。I_{RK} 的每一个列向量表示为该位置可能出现的所有字符的概率。CRNN 学习到的字符主要包含不区分大小写的字母数字 0～9，A～Z 字符和空白标签。

2.2.2　语义推理模块

本文采用预训练双向完型填空网络(Bidirectional Cloze Network，BCN)[19]作为语义推理模

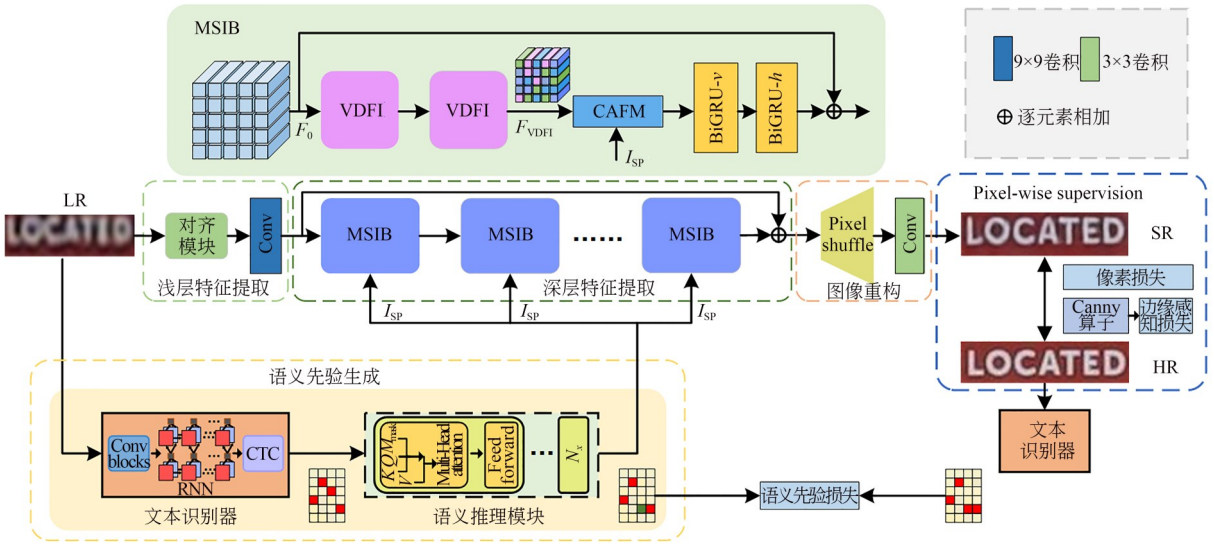


图 1　MSISR 整体架构

Fig. 1　Overall architecture of MSISR

块，对文本序列中的字符相关性进行建模，预测上下文信息，进而校正文本识别概率序列 I_{RK}。具体结构如图 1 所示。BCN 由一系列多头注意力和前馈网络构成，并且在多头注意中通过加入注意力掩码 M_{mask} 避免过度关注当前字符。加入掩码的多头自注意力 F_M 的计算公式如下：

$$F_M = soft\max\left[\frac{Q(I_{RK}W_1)^{\mathrm{T}}}{\sqrt{d}} + M_{mask}\right] I_{RK}W_2, \quad (2)$$

其中：d 为多头自注意力维度，W_1 和 W_2 为变换矩阵，$M_{mask}(i,j) = \begin{cases} 0, i \neq j \\ -\infty, i = j \end{cases} i,j \in [1,L]$，$soft\max(\cdot)$ 为激活函数。当预测第 i 个字符时，如果 $i=j$，那么 $M_{mask}(i,j) = -\infty$，权重系数为 0，即忽略当前字符本身信息，通过结合其上下文字符信息进行预测，避免过度关注当前字符，提高预测能力。

对于文本识别概率序列 I_{RK}，经过 SRM 语义推理后得到的语义信息为：

$$I_{SP} = SRM(I_{RK}), \quad (3)$$

其中 $SRM(\cdot)$ 为语义推理模块。

2.3　浅层特征提取

如图 1 所示，MSISR 浅层特征提取部分由对齐模块和 9×9 的卷积层构成。本文使用的数据集为通过相机改变焦距获得的真实场景 LR-HR 图像对，LR 和 HR 图像像素之间难免存在错位现象。因此，本文使用基于空间变换网络（Spatial Transformer Network，STN）的薄板样条变换（Thin Plate Spline，TPS）作为对齐模块。TPS 变换通过求解一个薄板样条插值函数来实现非刚性形变，将 LR 和 HR 图像中对应的字符区域转换成统一尺寸和形状的区域，防止网络学习错误的对应信息，缓解 LR 和 HR 图像的水平、垂直、斜向等像素错位问题。具体过程可表示为：

$$F_0 = STN(I_{LR}) \cdot W_s^{9\times9} + b_s, \quad (4)$$

其中：F_0 为 LR 图像浅层特征，I_{LR} 表示输入的 LR 图像，$W_s^{9\times9}$ 为 9×9 的卷积核，b_s 表示偏置，$STN(\cdot)$ 表示对齐模块操作。

2.4　深层特征提取

深层特征提取部分是由若干多模态语义交互块（Multimodal Semantic Interaction Block，MSIB）构建的一个残差组，在促进特征有效传递的同时，防止网络训练不稳定。MSIB 作为特征提取的重要组成部分，主要包含视觉双流集成块、跨模态自适应融合块以及正交的双向门控循环单元（Bidirectional Gated Recurrent Unit，BiGRU）。

2.4.1　视觉双流集成块

卷积神经网络具有局部连接性和平移不变性，可以较好地捕获输入图像的局部相关性，但缺乏长程依赖。尽管全局注意力机制擅长捕捉全局特征，但往往伴随较大的计算量和更高的资源消耗。Swin Transformer[20] 层的窗

口自注意力和移动窗口机制，可以有效捕获窗口内的像素关联信息，并实现窗口间的信息交互，进而增强网络全局建模的能力。为此，本文提出了一种高效的视觉双流集成块 VDFI，以关注特征图的不同层次信息，对局部的像素相关性和全局的语义依赖进行建模，将局部信息和全局信息相结合，包含文本字符的整体布局和局部结构细节在内的多粒度特征，为图像重构提供丰富的视觉信息，有助于处理变形和弯曲的文本图像。

在文本中，粗粒度指字符间信息，包含文本行的空间变形，使用 Transformer 自注意力机制和移动窗口机制建模，学习粗粒度的字符间远程依赖关系。细粒度指字符内信息，使用卷积网络，以及 transformer 局部窗口的自注意力来协同作用学习字符间细粒度的近程依赖关系。

VDFI 模块的具体结构如图 2 所示，本文在 Swin Transformer 层中将加入了由卷积、批量归一化（BN）、激活函数（GELU）和高效通道注意力机制（Efficient Channel Attention，ECA）共同组成的卷积块（Convolution Block，CB），以增强网络的表征能力。CB 模块使用两个 3×3 的卷积层进行局部特征提取，为了降低计算成本，通过第一个卷积层压缩通道数，并在第二个卷积层后恢复，最后通过 ECA 自适应调整通道特征，对特征进行细化。VDFI 模块中，第一个 LN（Layer Norm）层后，CB 块与多头自注意（Multi-head Self-attention，MSA）模块并行，以利用全局统计特性和较强的局部拟合能力互补优势。为了确保 CB 块和 MSA 块在优化过程中模块之间的协调和稳定性，在 CB 的输出中通过平衡参数 γ 进行调控。随后，LN 层和多层感知机（Multilayer Perceptron，MLP）级联，外层使用残差连接。

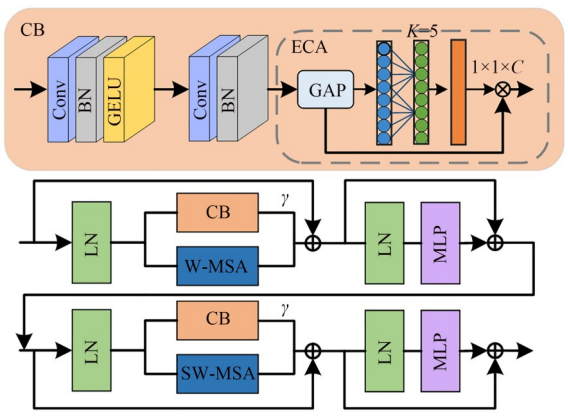


图 2　视觉双流集成块

Fig. 2　Visual dual flow integration module

对于输入特征 F_0，VDFI 模块的输出特征可以表示为：

$$F_1 = MSA\left(LN\left(F_0\right)\right) + \gamma CB\left(LN\left(F_0\right)\right) + F_0, \tag{5}$$

$$F_{\mathrm{VDFI}} = MLP\left(LN\left(F_1\right)\right) + F_1, \tag{6}$$

其中：F_{VDFI} 表示输出特征，$MSA(\cdot)$ 为多头自注意力操作，$LN(\cdot)$ 为层归一化操作，$MLP(\cdot)$ 为多层感知机操作，$CB(\cdot)$ 为卷积块操作，平衡参数 $\gamma = 0.01$，F_1 为中间特征。

2.4.2　跨模态自适应融合块

文本图像中包含重要的内容信息，对图像重构任务具有重要的指导意义。然而，多数场景文本图像超分辨率重构处理往往忽略了这一信息。针对这一问题，本文利用 CRNN 和语义推理预测得到包含丰富内容信息的语义特征 I_{SP}。语义特征 I_{SP} 是对文本图像内容的高层次理解，与多粒度视觉特征 F_{VDFI} 具有不同的模态。为了解决不同模态之间的底层特征隔阂，自适应学习视觉特征和语义特征之间的信息关联，充分发挥语义信息对文本图像重构网络的引导作用，本文提出了 CAFM，将语义信息融合到深层特征构建中。该模块的具体结构如图 3 所示。

CAFM 模块主要由特征转换、细化、聚合、双尺度通道注意四部分构成。为了与视觉特征 F_{VDFI} 的尺寸匹配，将语义特征 I_{SP} 进行特征转换，经过 3 个步长为（2，2）的反卷积层和一个步长为（2，1）的反卷积层，获得语义特征图 F_{SP}。F_{VDFI} 和 F_{SP} 沿通道维度进行级联，并通过串行的空间通道重构卷积（SCConv）[21] 对级联后的特征进行细化处理，利用空间重构单元（SRU）和信道重构单元（CRU）分别在空间和通道维度上降低冗余信息，进而减少背景信息和错误的语义信息带来的干扰。接着，采用 3 个并行的 1×1 卷积进行通道压缩，实现特征聚合，分别将特征投影到 F_{1n}，F_{2n}，F_{3n} 特征空间。通过全局平均池化（GAP）调节，在 F_{1n} 上并行地执行局部注意力和全局注意力权重计算，将注意力权重与 F_{2n} 相乘，自适应地选择特征，为不同信息差异化分配权重。最后，通过残差连接得到增强后的特征 F_{m}。该过程可

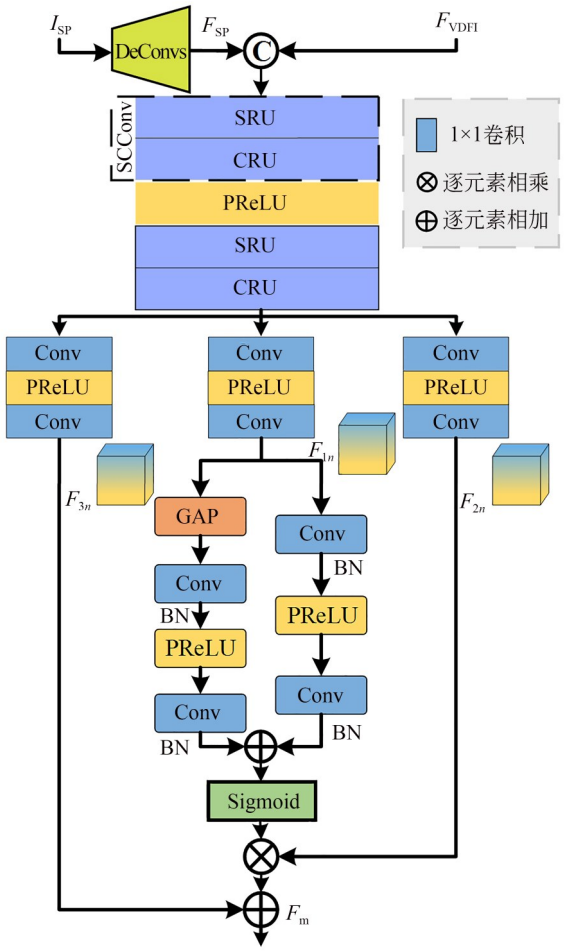


图 3　跨模态自适应融合模块

Fig. 3　Cross-modal adaptive fusion module

具体表示为：

$$F_{local}=BN(H_{conv1}(\delta_{prelu}(BN(H_{conv1}(F_{1n}))))), \quad (7)$$

$$F_{global}=BN(H_{conv1}(\delta_{prelu}(BN(H_{conv1}(H_{gap}(F_{1n})))))), \quad (8)$$

$$F_m=\delta_{sig}(F_{local}\oplus F_{global})\otimes F_{2n}\oplus F_{3n}, \quad (9)$$

其中：$H_{conv1}(\cdot)$为 1×1 卷积操作，$\delta_{prelu}(\cdot)$为 PReLU 激活函数，$H_{gap}(\cdot)$为全局平均池化操作，$\delta_{sig}(\cdot)$为 sigmoid 激活函数，$BN(\cdot)$为归一化操作，$\oplus$为逐元素相加，$\otimes$为逐元素相乘，F_{local}为提取的局部特征，F_{global}为提取的全局特征。

2.4.3　BiGRU 模块

场景文本图像中文本信息主要集中在水平和垂直两个方向，水平方向的上下文信息提供字符之间语义关联，垂直方向的上下文信息提供诸如笔画等字符内部特征。如图 1 所示，基于文本的序列数据特性，本文分别在垂直和水平方向上使用双向门控循环单元 BiGRU-v 和 BiGRU-h 捕捉多模态特征 F_m 的垂直和水平方向信息，并建立这两个方向上的文本依赖。具体过程可表示为：

$$F=H_{BiGRU_h}(H_{BiGRU_v}(F_m)), \quad (10)$$

其中：$H_{BiGRU_v}(\cdot)$和$H_{BiGRU_h}(\cdot)$分别为在垂直和水平方向使用 BiGRU 提取的特征，F为最终提取的特征。

2.5　图像重构

MSISR 网络采用一个亚像素卷积(Pixel Shuffle)层和一个 3×3 的卷积层对图像进行重构。亚像素卷积和转置卷积是两种常用的上采样方法。与转置卷积相比，亚像素卷积不引入待学习参数，仅通过重新排列不同通道的特征图达到上采样的目的，速度较快，且上采样前不需要进行零填充。对于输入特征 F，重构过程可表示为：

$$I_{SR}=W_{rec}^{3\times3}\cdot H_{up}(F)+b_{rec}, \quad (11)$$

其中：$H_{up}(\cdot)$表示使用亚像素卷积进行上采样操作，$W_{rec}^{3\times3}$为 3×3 卷积核，b_{rec}为偏置，I_{SR}为 SR 图像。

2.6　损失函数

MSISR 结合多个损失函数进行训练，主要有像素损失、边缘感知损失和语义先验损失。像素损失主要采用对 SR 图像和 HR 图像之间对应像素，使用 L_2 损失，具体如下：

$$L_2=\|I_{SR}-I_{HR}\|^2, \quad (12)$$

其中 I_{SR} 为 SR 图像，I_{HR} 为真实的 HR 图像。

文本的边缘信息包含文本形状、轮廓和结构等关键特征，有助于理解和处理文本内容。为了避免 SR 图像中文本字符边缘过度平滑，本文提出边缘感知损失 L_{EP}，具体如下：

$$L_{EP}=\|f(I_{HR})-f(I_{SR})\|_1, \quad (13)$$

其中 $f(\cdot)$为边缘提取算子。这里采用 Canny 算子提取边缘。与常用 Sobel 边缘提取算子相比，Canny 算子因为运用“非极大值抑制”和“形态学连接操作”，所以提取边缘比较完整，并且边缘连续性很好，具有良好的抗噪性能。

语义先验损失 L_{TP} 用以增强文本语义信息的引导作用，进一步提高图像重构效果，具体如下：

$$L_{TP}=\lambda_1\|L_{SP}-H_{HP}\|+\lambda_2 D_{KL}(L_{SP}\|H_{HP}), \quad (14)$$

其中：λ_1 和 λ_2 为平衡参数；I_{SP} 和 I_{HP} 分别为 LR 图像和 HR 图像中提取的语义信息；$D_{KL}(I_{SP} \| I_{HP})$ 为 I_{SP} 和 I_{HP} 的 KL 散度，具体为：

$$D_{KL}(I_{SP} \| I_{HP}) = \sum_{i=1}^{L}\sum_{j=1}^{|A|} I_{HP}^{ij} \ln \frac{I_{HP}^{ij}+\sigma}{I_{SP}^{ij}+\sigma}, \tag{15}$$

其中：I_{SP}^{ij} 和 I_{HP}^{ij} 分别为 I_{SP} 和 I_{HP} 中第 i 个位置第 j 个维度元素；σ 为一个很小的正数，避免运算中出现数值错误，本文设置 $\sigma = 10^{-6}$。

综合以上，MSISR 网络的总损失可表示为：

$$L = \alpha L_2 + \beta L_{EP} + \lambda L_{SP}, \tag{16}$$

其中 α，β，λ 为平衡参数，本文分别设置为 1，1×10^{-4}，1。

3　实验与结果分析

3.1　数据集与评价指标

本文使用的数据集为专门针对场景文本图像超分辨率重构问题的真实场景文本图像数据集 TextZoom[13]，该数据集包含数码相机拍摄的 21 740 个高低分辨率图像对。其中，17 367 对样本用于训练，其他用于测试。根据数码相机拍摄的焦距不同，通常将测试图像分为简单样本（easy）1 619 对、中等样本（medium）1 411 对和复杂样本（hard）1 343 对 3 个测试子集。针对文本识别器固定输入 32×128 的设计，进行 2 倍超分辨率重构，LR 尺寸为 16×64，HR 尺寸为 32×128。

场景文本图像超分辨率重构的核心目标是提升文本识别模型对 LR 文本图像的识别精度。因此，本文使用 3 种主流的文本识网络 ASTER[22]，CRNN[18]，MORAN[23] 对重构后的文本图像进行识别，并将识别精度作为重构网络的主要评价指标，通用的图像超分辨率重构评价指标结构相似度（Structural Similarity，SSIM）和峰值信噪比（Peak Signalto Noise Ratio，PSNR）作为辅助参考指标。

3.2　环境及参数设置

所有实验都是在单个 NVIDIA GTX 3090 GPU 上使用 PyTorch1.9 和 python3.9 进行实现。使用 Adam 优化器进行参数优化，动量设置为 0.9，Batch 大小设置为 48，将学习率设置为 10^{-3}，进行了 500 epoch 的训练。

3.3　消融实验

3.3.1　MSIB 模块数量分析

本文模型中，若干 MSIB 模块组成的残差结构用于提取深层特征，MSIB 的堆叠数量直接影响模型性能。在 easy，medium 和 hard 3 个测试子集上，保持其他参数不变，通过设置不同的 MSIB 模块堆叠数量进行实验，进而研究 MSIB 模块堆叠数量对模型性能的影响。实验采用 CRNN 网络对重构图像进行文本识别，结果如表 1 所示。其中，avg 为根据各子集样本数量计算的加权平均值，最佳识别率用黑体加粗表示。实验结果表明，增加 MSIB 的数量并不能一直提升性能，当 MSIB 堆叠数量为 5 时，网络的性能达到饱和，此时模型性能最优。

表 1　MSIB 数量对识别精度的影响

Tab.1　Influence of recognition accuracy on number of MSIB　　(%)

数量	CRNN 准确率			
	easy	medium	hard	avg
3	62.6	50.8	37.9	51.2
4	64.0	52.7	39.1	52.7
5	**64.8**	**54.0**	**39.8**	**53.6**
6	64.1	53.2	39.4	53.0
7	63.4	51.7	38.3	51.9

3.3.2　模型模块分析

本文利用 VDFI 来提取文本图像的多粒度特征，通过 CAFM 学习不同模态之间的信息关联，并结合边缘损失 L_{EA} 监督网络对于文本边缘的重建。为了研究模型中不同模块对最终重构结果的影响，本节在 easy，medium 和 hard 3 个测试子集上，对 VDFI，CAFM，L_{EA}、语义先验信息的有效性进行分析，实验结果如表 2 所示。其中，avg 为根据各子集样本数量计算的加权平均值，最佳识别率用黑体加粗表示，×表示没有使用相应模块，√表示使用了相应模块，Swin 为使用 Swin Transformer 提取图像深层特征。实验结果表明，模型同时加入各模块时取得了最佳表现。与 Swin Transformer 相比，VDFI 模块模型识别精度可以提升 0.7%。

表 2　不同模块对应的识别准确率

Tab. 2　Recognition accuracy of different modules

Swin	语义先验	VDFI	CAFM	L_{EA}	avg/%
×	×	√	×	×	44.2
×	√	√	×	×	52.3
×	√	×	√	×	52.0
×	√	×	×	√	51.4
×	√	√	√	×	53.2
√	√	×	√	√	52.9
×	√	√	√	√	**53.6**

3.3.3　融合策略分析

为了融合多粒度视觉特征 F_{VDFI} 和语义信息 F_{SP}，充分发挥语义先验对文本图像重构的促进作用，本文提出了CAFM融合模块。为了进一步验证该融合策略的有效性，这里在easy，medium和hard 3个测试子集上，在同一模型上使用不同的融合策略进行对比实验。需要说明的是，基于内存的考虑采用小的基线模型，使用SRB[13]作为特征提取模块。实验过程中，采用CRNN网络对重构图像进行文本识别，结果如表3所示。其中，avg为根据各子集样本数量计算的加权平均值，最佳识别率用黑体加粗表示，C为级联，A为逐元素相加，CA为通道注意力。实验结果表明，通道注意力可以在级联的基础上，将识别率提高0.4%，而相较于级联和逐元素相加这两种简单的线性操作，CAFM方法的平均识别精度分别提高了1.5%，1.7%。

表 3　不同融合策略对识别精度的影响

Tab. 3　Impact of different fusion strategy over recognition accuracy　(%)

融合策略	CRNN准确率			
	easy	medium	hard	avg
C	61.7	50.6	37.0	50.5
A	61.2	50.8	36.7	50.3
C+CA	61.9	51.2	37.3	50.9
CAFM	**63.1**	**52.6**	**38.1**	**52.0**

3.3.4　边缘损失函数分析

文本边缘中包含许多关键特征，本文模型利用Canny算子提取边缘，并运用边缘损失引导模型训练，避免重构图像过于平滑。为了进一步验证该方法的有效性，在easy，medium和hard 3个测试子集上，并在同一模型上使用不同的边缘损失进行实验。与常规的计算图像梯度损失 L_{GP}[13]和基于Sobel算子的边缘损失 L_{EG}[24]进行对比实验，结果如表4所示。通过CRNN网络对重构图像进行文本识别，avg为根据各子集样本数量计算的加权平均值，最佳识别率用黑体加粗表示。实验结果表明，Canny算子运用“非极大值抑制”和“形态学连接操作”，重构图像识别结果最佳。

表 4　不同损失函数对识别精度的影响

Tab. 4　Impact of different loss function over recognition accuracy　(%)

损失	CRNN准确率			
	easy	medium	hard	avg
L_{GP}	61.7	50.6	37.0	50.5
L_{EG}	62.5	51.1	37.3	51.1
L_{EA}	**62.8**	**51.6**	**37.5**	**51.4**

3.3.5　边缘损失函数参数分析

如式(16)所示，MSISR网络总损失包含3个部分，参考TPGSR方法，这里设置 $\alpha=1$，$\lambda=1$。为了分析边缘感知损失的权重 β 对重构效果的影响，保持 α 和 λ 值不变，通过设置不同权重的 β 值进行实验，通过CRNN网络对重构图像进行文本识别，avg为根据各子集样本数量计算的加权平均值，如表5所示。实验结果表明，当

表 5　不同 β 值对识别精度的影响

Tab. 5　Impact of different β values over recognition accuracy

β	平均识别精度/%
1×10^{-5}	53.2
1×10^{-4}	53.6
1×10^{-3}	53.1
1×10^{-2}	52.9
1×10^{-1}	52.5

β权重设置为1×10^{-4}时，平均识别精度最佳。

3.4　对比实验与结果分析

3.4.1　客观指标分析

为了验证本文提出方法的有效性，本文在公开数据集TextZoom上进行2倍超分辨率重构实验，并与11种主流超分辨率重构方法进行比较，包括双三次插值（Bicubic）、SRCNN[3]，HAN[4]，TSRN[13]，PCAN[24]，TBSRN[25]，TG[26]，MTSR[27]，TATT[15]，DPGSR[16]和TPGSR[14]。对于TPGSR，本文对比了一阶（TPGSR）和三阶（TPGSR-3）两种模型。表6列出了不同方法重构图像的文本识别精度的平均值，每组中识别精度最高的用粗体表示，对次优算法进行下划线表示，avg为根据各子集样本数量计算的加权平均值。结果显示，传统的Bicubic方法具有最低的识别率。经SRCNN，HAN方法重构后的图像识别精度优于Bicubic方法，但作为图像超分辨率重构的通用模型，因缺乏对特定场景文本图像的处理能力，重构结果并未达到最佳。与SRCNN，HAN方法相比，TSRN，PCAN方法利用LSTM获取文本上下文信息，文本识别准确率得到了明显提升。TBSRN，MTSR方法基于自注意力机制捕捉文本图像远距离依赖，取得了相对较好的结果，但由于缺乏局部细节信息，识别精度并没有得到明显提升。得益于文本先验信息的引入，TG，TATT，DPGSR，TPGSR方法利用各种文本属性作用于SR网络，重构图像获得了相对较好的识别精度。本文方法重构的图像具有最佳表现，与直接采用Bicubic上采样生成的SR图像相比，本文模型对ASTER，MORAN和CRNN的加权平均识别准确率分别提高了16.9%，17.6%和26.8%；与主流的TPGSR相比，本文的模型将平均识别准确率分别提高了2.9%，3.6%和3.7%。

不同方法重构图像的PSNR和SSIM值如表7所示，本文方法在PSNR和SSIM这两个技术指标上相比于其他方法仍有一定的优势。

表6　不同方法在TextZoom数据集上的文本识别准确率

Tab. 6　Recognition accuracy of different methods on TextZoom dataset　　(%)

算　法	ASTER				MORAN				CRNN			
	easy	medium	hard	avg	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	36.4	21.1	21.1	26.8
SRCNN[3]	69.4	43.4	32.2	49.5	63.2	39.0	30.2	45.3	38.7	21.6	20.9	27.7
HAN[4]	71.1	52.8	39.0	55.3	67.4	48.5	35.4	51.5	51.6	35.8	29.0	39.6
TSRN[13]	75.1	56.3	40.1	58.3	70.1	53.3	37.9	54.8	52.5	38.2	31.4	41.4
PCAN[24]	77.5	60.7	43.1	61.5	73.7	57.6	41.0	58.5	59.6	45.4	34.8	47.4
TBSRN[25]	75.7	59.9	41.6	60.0	74.1	57.0	40.8	58.4	59.6	47.1	35.3	48.1
TG[26]	77.9	60.2	42.4	61.3	75.8	57.8	41.4	59.4	61.2	47.6	35.5	48.9
MTSR[27]	75.6	59.8	43.4	58.9	73.9	57.2	41.8	56.0	56.2	47.0	35.3	45.4
TATT[16]	78.9	63.4	45.4	63.6	72.5	60.2	43.1	59.5	62.6	53.4	39.8	52.6
DPGSR[17]	75.5	57.8	41.9	59.4	69.7	53.4	39.7	55.2	57.6	43.0	33.4	45.5
TPGSR[14]	77.0	60.9	42.4	61.2	72.2	57.8	41.3	58.1	61.0	49.9	36.7	49.9
TPGSR-3[14]	78.9	62.7	44.5	62.8	74.9	60.5	44.1	60.5	63.1	52.0	38.6	51.8
Ours	**80.0**	**63.6**	**45.6**	**64.1**	**76.5**	**60.9**	**44.8**	**61.7**	**64.8**	**54.0**	**39.8**	**53.6**

表 7　不同方法在 TextZoom 数据集上的 PSNR 和 SSIM

Tab. 7　PSRN and SSIM of different methods on TextZoom dataset

算　法	PSNR				SSIM			
	easy	medium	hard	avg	easy	medium	hard	avg
Bicubic	22.35	18.98	19.39	20.35	0.788 4	0.625 4	0.659 2	0.696 1
SRCNN[3]	23.48	19.06	19.34	20.78	0.837 9	0.632 3	0.679 1	0.722 7
HAN[4]	23.30	19.02	20.16	20.95	0.869 1	0.653 7	0.738 7	0.759 7
TSRN[13]	**25.07**	18.86	19.71	21.42	0.889 7	0.667 6	0.730 2	0.769 0
PCAN[24]	24.57	19.14	20.26	21.49	0.883 0	0.678 1	0.747 5	0.775 2
TBSRN[25]	23.46	19.17	19.68	20.91	0.872 9	0.645 5	0.745 2	0.760 3
TG[26]	23.82	19.17	19.68	21.05	0.866 0	0.653 3	0.749 0	0.761 4
MTSR[27]	23.55	19.88	19.64	21.16	0.873 4	0.684 3	0.747 6	0.773 9
TATT[16]	24.72	19.02	20.31	21.53	0.900 6	0.691 1	0.770 3	0.793 0
DPGSR[17]	23.36	18.76	19.77	20.77	0.871 1	0.671 9	0.750 7	0.769 8
TPGSR[14]	23.73	18.68	20.06	20.97	0.880 5	0.673 8	0.744 0	0.771 9
TPGSR-3[14]	24.35	18.73	19.93	21.18	0.886 0	0.678 4	0.750 7	0.777 4
Ours	24.76	**19.98**	**20.39**	**21.88**	**0.901 3**	**0.697 6**	**0.778 0**	**0.797 7**

3.4.2　不同方法对比

为了更直观地展示本文方法的优势，进行了可视化操作。其中，TBSRN，MTSR 方法未公开相关资源，没有进行可视化对比。本文从 TextZoom 测试集的 3 个子集中分别选取 2 张图片进行可视化效果的对比，可视化结果如图 4 所示。

图 4　不同方法在 TextZoom 数据集上的可视化结果

Fig. 4　Visualization result of different methods on TextZoom dataset

其中，红色字符表示识别错误的字符（彩图见期刊电子版）。

Bicubic方法重建的图像存在过度平滑的问题，导致文本图像整体视觉效果模糊，无法获得清晰的字符边缘。SCRNN，HAN方法的视觉效果没有明显的提升，边缘完整性较差。尽管TSRN，PCAN，TBSRN和TG方法能够获得相对较好的文本图像重建效果，但在细节处理上仍存在问题，字符间界限模糊，存在相邻字符之间的粘连现象。与前几种方法相比，虽然TPGSR，TATT方法可以生成较为清晰的文本图像，但仍会重建出错误的字符信息，缺乏对文本区域细节信息的重建并伴有伪影的文本边缘。本文提出的方法能更好地重建语义正确的文本图像，恢复清晰的字符边缘，提高场景文本图像的视觉效果，更适用于STISR重建任务。

4　结　论

本文提出多模态语义交互的文本图像超分辨率重构网络，用于重建语义正确的文本图像。考虑到文本内容包含丰富的语义信息，使用语义推理模块获取语义正确的文本内容信息，引导网络进行图像重构。通过设计的多模态语义交互块作为特征提取的主干网络，其中，利用视觉双流集成块有效整合来自局部-全局不同层次信息，跨模态自适应融合块动态学习两种模态之间的语义交互，正交双向门控循环单元捕捉多模态特征在水平和垂直方向信息。在TextZoom数据集上的实验结果表明，与先进方法相比，本文方法在PSNR和SSIM上均有提升，在ASTER，MORAN，CRNN 3种识别器下的平均识别率相比TPGSR模型分别提高了2.9%，3.6%和3.7%。本文方法重建的文本图像更利于人眼辨识。现有的文本图像超分辨率重构方法多集中于提升英文和数字字符的分辨率，为满足中文字符的超分辨率重构需求，下一步研究着重于中文字符的超分辨率重构处理。

作者贡献声明：

韩玉兰：重构方法的提出，论文构思；

罗轶宏：论文撰写，实验设计及数据整理；

崔玉杰：论文审核及投稿指导；

兰朝凤：测量实验数据分析。

参考文献：

[1] GUAN T K, SHEN W, YANG X, *et al*. Self-supervised character-to-character distillation for text recognition [C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023. *Paris*, *France*. IEEE, 2023: 19473-19484.

[2] LI M H, LV T C, CHEN J Y, *et al*. TrOCR: transformer-based optical character recognition with pre-trained models [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37 (11): 13094-13102.

[3] DONG C, LOY C C, HE K M, *et al*. *Learning a Deep Convolutional Network for Image Super-resolution* [M]. Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 184-199.

[4] NIU B, WEN W L, REN W Q, *et al*. *Single image Super-resolution Via a Holistic Attention Network* [M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 191-207.

[5] 寇旗旗，李超，程德强，等. 基于注意力和宽激活密集残差网络的图像超分辨率重建[J]. 光学 精密工程，2023，31(15)：2273-2286.
KOU Q Q, LI CH, CHENG D Q, *et al*. Image super-resolution reconstruction based on attention and wide-activated dense residual network [J]. *Opt. Precision Eng.*, 2023, 31(15): 2273-2286. (in Chinese)

[6] 周颖，裴盛虎，陈海永，等. 基于多尺度自适应注意力的图像超分辨率网络[J]. 光学 精密工程，2024，32(6)：843-856.
ZHOU Y, PEI SH H, CHEN H Y, *et al*. Image super-resolution network based on multi-scale adaptive attention [J]. *Opt. Precision Eng.*, 2024, 32 (6): 843-856. (in Chinese)

[7] 夏振平，陈豪，张宇宁，等. 基于混合时空卷积的轻量级视频超分辨率重建[J]. 光学 精密工程，2024，32(16)：2564-2576.
XIA ZH P, CHEN H, ZHANG Y N, *et al*. Lightweight video super-resolution based on hybrid spatio-temporal convolution [J]. *Opt. Precision Eng.*,

2024, 32(16): 2564-2576. (in Chinese)

[8] ZHU S P, ZHAO Z Y, FANG P F, *et al*. Improving scene text image super-resolution *via* dual prior modulation network[C]. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. ACM, 2023: 3843-3851.

[9] CHEN X Y, WANG X T, ZHOU J T, *et al*. Activating more pixels in image super-resolution transformer[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023. *Vancouver*, *BC*, *Canada*. IEEE, 2023: 22367-22377.

[10] WANG W J, XIE E Z, SUN P Z, *et al*. TextSR: content-aware text super-resolution guided by recognition[EB/OL]. 2019: 1909.07113. https://arxiv.org/abs/1909.07113v4.

[11] WANG Y 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*). *July* 8-12, 2019. *Shanghai*, *China*. IEEE, 2019: 1024-1029.

[12] MOU Y Q, TAN L, YANG H, *et al*. *PlugNet: Degradation Aware Scene Text Recognition Supervised by a Pluggable Super-resolution Unit*[M]. Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 158-174.

[13] 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.

[14] MA J Q, GUO S, ZHANG L. Text prior guided scene text image super-resolution[J]. *IEEE Transactions on Image Processing*, 2023. DOI:10.1109/TIP.2023.3237002.

[15] 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*). *June* 18-24, 2022. *New Orleans*, *LA*, *USA*. IEEE, 2022: 5911-5920.

[16] YANG H, ZHOU H B. Degradation prior guided scene text image super-resolution[C]. 2022 4*th International Conference on Frontiers Technology of Information and Computer* (*ICFTIC*). *December* 2-4, 2022. *Qingdao*, *China*. IEEE, 2022: 170-175.

[17] MA J Z, JIN L W, ZHANG J X, *et al*. TextSRNet: scene text super-resolution based on contour prior and atrous convolution[C]. 2022 26*th International Conference on Pattern Recognition* (*ICPR*). *August* 21-25, 2022. *Montreal*, *QC*, *Canada*. IEEE, 2022: 3252-3258.

[18] FU X Y, CH′NG E, AICKELIN U, *et al*. CRNN: a joint neural network for redundancy detection[C]. 2017 *IEEE International Conference on Smart Computing* (*SMARTCOMP*). *May* 29-31, 2017. *Hong Kong*, *China*. IEEE, 2017: 1-8.

[19] 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*). *June* 20-25, 2021. *Nashville*, *TN*, *USA*. IEEE, 2021: 7098-7107.

[20] LIU Z, LIN Y T, CAO Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021. *Montreal*, *QC*, *Canada*. IEEE, 2021: 10012-10022.

[21] LI J F, WEN Y, HE L H. SCConv: spatial and channel reconstruction convolution for feature redundancy[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023. *Vancouver*, *BC*, *Canada*. IEEE, 2023: 6153-6162.

[22] 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.

[23] 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: 109-118.

[24] ZHAO C R, FENG S Y, ZHAO B N, *et al*. Scene text image super-resolution *via* parallelly contextual attention network[C]. *Proceedings of the* 29*th ACM International Conference on Multi-*

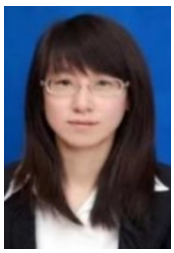
media. *October* 20-24, 2021, *Virtual Event*, *China*. ACM, 2021: 2908-2917.

[25] 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*). *June* 20-25, 2021. *Nashville*, *TN*, *USA*. IEEE, 2021: 12026-12035.

[26] 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.

[27] HONDA K, KUREMATSU M, FUJITA H, *et al*. Multi-task learning for scene text image super-resolution with multiple transformers[J]. *Electronics*, 2022, 11(22): 3813.

作者简介:



韩玉兰(1984—),女,黑龙江大庆人,博士,讲师,硕士生导师,主要从事图像重构、计算机视觉和机器学习的研究。E-mail: hanyulan@hrbust.edu.cn

罗轶宏(1998—),女,黑龙江海林人,硕士研究生,主要从事图像超分辨率重建方面的研究。E-mail: luoy1h@163.com