E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)05-1364-13

论文引用格式:Zhu Z J, Zhang L, Li P, Tu R W, Bai Y Q and Wang Y E. 2025. Text image super-resolution via structure perception enhancement and cross modal fusion. Journal of Image and Graphics, 30(5):1364-1376(朱仲杰, 张磊, 李沛, 屠仁伟, 白永强, 王玉儿. 2025. 结构感知增强与跨模态融合的文本图像超分辨率. 中国图象图形学报, 30(5):1364-1376)[DOI:10.11834/jig.240559]

结构感知增强与跨模态融合的文本图像超分辨率

朱仲杰*,张磊,李沛,屠仁伟,白永强,王玉儿

浙江万里学院宁波市工业视觉与工业智能重点实验室,宁波　315100

摘　要:**目的**　场景文本图像超分辨率是一种新兴的视觉增强技术,用于提升低分辨率文本图像的分辨率,从而提高文本可读性。然而,现有方法无法有效提取文本结构动态特征,导致形成的语义先验无法与图像特征有效对齐并融合,进而影响图像重建质量并造成文本识别困难。为此,提出一种基于文本结构动态感知的跨模态融合超分辨率方法以提高文本图像质量和文本可读性。**方法**　首先,构建文本结构动态感知模块,通过方向感知层和上下文关联单元,分别提取文本的多尺度定向特征并解析字符邻域间的上下文联系,精准捕获文本图像的结构动态特征;其次,设计语义空间对齐模块,利用文本掩码信息促进精细化文本语义先验的生成,并通过仿射变换对齐语义先验和图像特征;最后,在此基础上,通过跨模态融合模块结合文本语义先验与图像特征,以自适应权重分配的方式促进跨模态交互融合,输出高分辨率文本图像。**结果**　在真实数据集TextZoom上与多种主流方法进行对比,实验结果表明所提方法在ASTER(attentional scene text recognizer)、CRNN(convolutional recurrent neural network)和MORAN(multi-object rectified attention network)3种文本识别器上的平均识别精度为62.4%,较性能第2的方法有2.8%的提升。此外,所提方法的峰值信噪比(peak signal-to-noise ratio,PSNR)和结构相似性(structural similarity index,SSIM)指标分别为21.9 dB和0.789,分别处于第1名和第2名的位置,领先大多数方法。**结论**　所提方法通过精准捕获文本结构动态特征来指导高级文本语义先验的生成,从而促进文本和图像两种模态的对齐和融合,有效提升了图像重建质量和文本可读性。

关键词:场景文本图像超分辨率(STISR);文本结构动态特征;多尺度定向特征;语义空间对齐;跨模态融合

Text image super-resolution via structure perception enhancement and cross modal fusion

Zhu Zhongjie*, Zhang Lei, Li Pei, Tu Renwei, Bai Yongqiang, Wang Yuer

Key Laboratory of Ningbo Industrial Vision and Industrial Intelligence, Zhejiang Wanli University, Ningbo　315100, China

Abstract: **Objective** Scene text image super-resolution (STISR) is an emerging visual enhancement technology aimed at enhancing the quality of low-resolution text images and improving text readability. This technology is extensively applied in

收稿日期:2024-09-18;**修回日期**:2024-10-25;**预印本日期**:2024-11-01

*通信作者:朱仲杰　zhongjiezhu@yeah.net

基金项目:宁波市科技创新2025重大专项(2022Z076);国家自然科学基金项目(61671412);浙江省自然科学基金项目(LZ24F010004);宁波市"科创甬江2035"关键技术突破项目(2024Z295, 2024Z125, Z024Z122);宁波市"科创甬江2035"重大应用示范项目(2024Z023);浙江省基础公益项目(LGN22F010002)

Supported by: Ningbo Municipal Major Project of Science and Technology Innovation 2025(2022Z076); National Natural Science Foundation of China (61671412); Natural Science Foundation of Zhejiang Province, China (LZ24F010004); Key Technology Breakthrough Project of Ningbo "Yongjiang Sci-Tech Innovation 2035"(2024Z295, 2024Z125, Z024Z122); Major Application Demonstration Project of Ningbo "Yongjiang Sci-Tech Innovation 2035"(2024Z023); Basic Public Welfare Project of Zhejiang Province(LGN22F010002)

fields such as autonomous driving, document retrieval, and text recognition. Existing STISR techniques can be categorized into traditional and deep learning methods. Traditional methods offer some image enhancement but heavily rely on manual feature extraction and complex prior knowledge. On the contrary, deep learning methods showcase considerable advantages due to their robust automatic feature extraction and learning capabilities. Therefore, deep learning STISR methods have become increasingly popular within the academic community. Recent deep learning-based STISR methods leverage rich semantic priors from text images to guide image reconstruction and text recovery. Unlike methods that solely rely on visual feature extraction, such as convolutional neural networks, multilayer perceptrons, and Transformers, these emerging methods integrate semantic priors with visual feature analysis for more efficient image reconstruction. However, they often overlook dynamic features of text structure, such as the contextual connections between neighboring characters and directional features. Therefore, the text semantic priors generated are not effectively aligned or fused with image features, limiting the quality of the reconstructed images. To overcome these limitations, a new cross-modal fusion super-resolution method, which incorporates text structure dynamic perception, is proposed to enhance the quality of low-resolution text images and text readability. **Method** In this study, an innovative cross-modal fusion STISR method that leverages text structure dynamic perception is proposed. The process begins with low-resolution text images being corrected using a spatial transform network module and a thin plate splines module. These modules adjust the spatial positions of irregular characters, preparing the images for further processing. The corrected images are then processed through a text structure dynamic perception module and a semantic space alignment module to extract image modal and text modal features. The text structure dynamic perception module, which includes a direction sensing block and a context linkage unit, captures multiscale directional features and identifies the contextual relationships between character neighboring characters, respectively, accurately capturing the dynamic features of the text structure within the image modal. The semantic space alignment module processes the low-resolution images using recognition rendering and binarization to derive text semantic priors and mask priors. These priors are then combined through feature addition to generate advanced text semantic priors, which are aligned with image features through affine transformations, guided by the image modal. Finally, the developed cross-modal fusion module employs an adaptive weight distribution strategy to enhance the interactive integration of text and image features across modals, producing the final super-resolved text image. **Result** The proposed method was evaluated against 13 mainstream methods, with a primary focus on the text recognition accuracy of the reconstructed text images (this metric is critical, given the unique nature of text images, where readability and recognizability are paramount). Secondary evaluation metrics included peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), which, despite their limitations in fully capturing image quality due to the misalignment of low-resolution and high-resolution images in real datasets, supplemented the evaluation. Experiments conducted on the real dataset Texezoom demonstrated that the proposed method achieved recognition accuracies of 67.1%, 56.6%, and 63.6% on three standard text recognizers, namely, attentional scene text recognizer(ASTER), convolutional recurrent neural network(CRNN), and multi-object rectified attention network(MORAN), respectively, outperforming the existing representative method, PERMR, by 3.0%, 4.6%, and 3.1%, respectively. In terms of image quality, the proposed method achieved PSNR and SSIM values of 21.9 dB and 0.789, ranking first and second, respectively, among all compared methods. Additionally, visual comparisons further highlighted the superior quality and readability of the text images reconstructed using the proposed approach. **Conclusion** In this study, a novel STISR method that effectively generates advanced text semantic priors by accurately capturing dynamic text structure features is proposed. This method facilitates the alignment and integration of text semantic priors with image features, drastically improving the quality of reconstructed text images and enhancing text readability. Experimental results demonstrate that the proposed method outperforms other mainstream methods, delivering notable improvements in image quality and text readability.

Key words: scene text image super resolution (STISR); dynamic features of text structure; multi-scale orientation feature; semantic space alignment; cross modal fusion

0 引 言

自然场景中的文本图像蕴含丰富的语义信息，对人类认识和理解世界起着至关重要的作用，在自动驾驶（Hao等，2024；Pourkeshavarz等，2024）、文档检索（乔梁 等，2023；秦海 等，2023）等领域广泛应用。然而，受到采集环境和设备性能的限制，采集到的文本图像经常出现运动模糊、失真和鬼影等问题，严重影响文本可读性，从而妨碍后续的文本识别、关键信息提取等任务。因此，如何提高场景文本图像的质量，确保文本信息的清晰可读，成为亟需解决的问题。

场景文本图像超分辨率（scene text image super-resolution，STISR）是一种新兴的视觉增强技术，专门用于增强低分辨率（low-resolution，LR）文本图像的分辨率，以提高文本可读性。通过学习LR与高分辨率（high-resolution，HR）文本图像之间的映射关系，STISR可以有效提高LR图像的质量，同时增强文本清晰度。然而，与普通图像的超分辨率（super-resolution，SR）处理不同，文本图像SR涉及字符连续性、复杂先验信息等独特挑战，如图1所示。这些挑战具体包括：字符的多尺度特征使得细节恢复更加困难；字符与背景之间缺乏关联，背景噪声对文本恢复产生影响；此外，保持字符之间的连续性和整体布局也是巨大挑战之一。

图1 STISR面临的挑战（如字符连续性、多尺度细节等）

Fig. 1 Challenges faced by STISR (such as character continuity, multi-scale details, and so on)

为此，研究人员提出了面向文本图像的超分辨率方法以同时增强图像质量和文本可读性，可以分为传统方法（Glasner等，2009；Farsiu等，2004；Chang等，2004）和深度学习方法（Wang等，2020；Chen等，2021，2022；Zhu等，2024）。传统方法包括基于插值的方法（Li和Orchard，2001）和基于统计的方法（Yang等，2010）等，由于依赖手工设计的特征提取器和复杂的先验知识，该类方法难以应对复杂场景中的多样化挑战。相比之下，深度学习方法凭借其强大的特征自动提取和学习能力，能够通过端到端的方式对大规模数据进行训练，在效果上显著优于传统方法，因此受到更广泛的关注和应用。

深度学习STISR方法高度依赖于大规模、高质量的训练数据。目前，训练数据主要分为合成图像数据集和真实图像数据集两种。合成文本图像可以通过高斯模糊、Bicubic等插值方法快速生成大量数据，但由于生成图像与真实场景中的图像在数据分布上存在差异，这些合成数据无法精确模拟现实世界中文本的复杂性，限制了模型的性能表现。相反，真实文本图像较合成图像在视觉感知上更为模糊，能更准确反映真实环境中的复杂噪声和文本细节，如图2所示，因此在训练中能更好地提升模型的准确性、鲁棒性和泛化能力。目前大多数STISR方法优先选择使用真实图像进行训练，以期获得更佳的模型效果。

然而，真实LR文本图像通常存在严重的质量退化，仅通过一般的图像特征提取与处理难以有效恢复文本细节。因此，研究者逐渐开始探索并利用文本图像中的丰富语义先验信息指导图像重建与文本恢复。相较于传统的卷积神经网络（convolutional neural network，CNN）、多层感知机（multilayer perceptron，MLP）和Transformer（Vaswani等，2017）等仅依赖视觉特征的方法，这些新兴的STISR方法通过结合语义先验进行文本内容学习，辅以视觉分析，大幅提升了图像重建的效率以及文本细节的清晰度，尤其在复杂背景和高噪声环境下表现更为出色。

图 2　真实 LR、合成 LR 和 HR 的对比

Fig. 2　Comparison of real LR, synthetic LR and HR

尽管现有研究已认识到文本语义先验在 STISR 中的重要性，但如字符定向特征和字符邻域的上下文关联等文本结构动态特征仍被大多数现有方法忽视。这些特征描述了字符的排列方向及空间位置关系，对于精确解析文本细节特征及其上下文依赖关系至关重要。然而，目前的 STISR 方法未能充分利用这些特征，导致在图像生成过程中重要的字符细节邻域信息丢失，使得生成的语义先验无法有效与图像特征对齐和融合，最终影响生成的文本图像质量，降低文本可读性。

为此，本文提出一种基于文本结构动态感知的跨模态融合 STISR 方法，通过捕获文本的动态结构特征来指导文本语义先验的对齐，从而促进文本和图像的交互融合，实现精细化图像重建。本文贡献总结如下：1）构建文本结构动态感知模块，通过方向感知层和上下文关联单元，分别提取文本的多尺度定向特征并解析字符邻域间的上下文联系，精准捕获文本图像的结构动态特征；2）构建语义空间对齐模块，利用文本掩码先验信息促进精细化文本语义先验的生成，并通过仿射变换对齐文本语义先验与图像特征；3）在上述基础上，通过跨模态融合模块结合文本语义先验与图像特征，以自适应权重分配的方式促进跨模态交互融合，实现高效的图像重建和文本恢复，如图 3 所示。

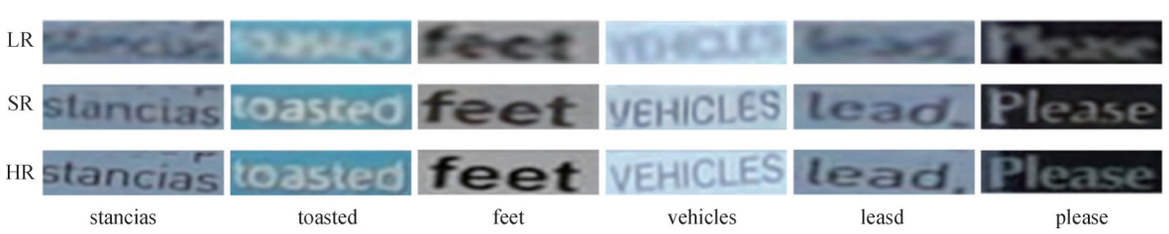

图 3　所提方法的重建文本图像

Fig. 3　Reconstructed text images rendered by the proposed method

1　相关工作

1.1　场景文本图像超分辨率

早期的深度学习 STISR 方法未能有效区分文本图像和普通图像，而是将文本图像视做一般图像进行处理，导致文本高频细节的提取能力不足，尤其在重建复杂低分辨率图像时表现不佳。Dong 等人（2014）在 SR 领域首次应用 CNN，较传统方法取得了一定进步，但由于网络深度较浅，仍难以捕捉复杂的高频细节。为此，Tran 和 Ho-Phuoc（2019）引入深度拉普拉斯金字塔网络，通过多尺度处理提升了文本高频细节特征的捕捉能力。然而，上述的早期深度学习方法大多基于合成数据训练，未能充分模拟真实世界文本图像的复杂性，导致其在实际应用中效果有限。

近年来，Wang 等人（2020）创建 STISR 领域首个真实数据集 TextZoom，推动了 STISR 的快速发展。与此同时，他们对文本结构动态特征初步尝试利用，通过结合 BiLSTM（bi-directional long short-term memory）和 CNN 的方式来感知文本字符的上下文依赖信息和局部结构特征。尽管基于真实数据集进行训练的模型取得了显著进步，但由于缺乏明确的文本语义先验，超分辨率性能依然有限。为了进一步

提升LR图像质量和文本可读性，Ma等人(2023)引入预训练文本识别器获取文本语义先验来指导文本图像重建。由于该方法仅简单地将文本语义先验嵌入到文本图像中，未能充分关注字符邻域间的结构动态特征和特征对齐问题，导致文本与图像两种模态特征的融合欠佳，从而影响重建图像质量。于是，为了增强文本结构动态特征，Chen等人(2021)基于字符空间位置关系，结合文本语义先验，并应用字符级别注意力机制，提高了模型对字符邻域上下文依赖性的关注度。然而，语义先验与图像特征之间的特征不一致性仍然存在，使得文本与图像模态的融合不够充分，文本细节的恢复空间仍待提升。为此，Zhu等人(2023)联合视觉和语言模型共同促进精细化文本语义先验的生成，通过增强语义先验准确性的方式指导图像重建。但该方法忽视了连续字符间的上下文依赖性和文本结构动态特征，且在跨模态特征融合时依旧出现跨特征不对齐问题，影响LR图像的重建。

针对上述问题，本文提出一种基于文本结构动态感知的跨模态融合超分辨率方法，有效提升真实LR图像的质量及其文本可读性。

1.2 场景文本识别

场景文本识别(scene text recognition，STR)旨在从文本图像中识别文本信息，是STISR的重要后续任务。STR方法(Shi等，2016，2019；Luo等，2019)主要包括基于CTC(connectionist temporal classification)机制(Graves等，2006)和基于注意力机制(Vaswani等，2017)的方法。CTC机制识别方法的代表性工作为CRNN(convolutional recurrent neural network)(Shi等，2016)，该方法整合CNN和RNN分别用于特征提取和序列建模，并通过CTC损失函数学习字符位置，实现端到端识别。尽管该方法在规则文本中展现出优势，但面临复杂背景和不规则文本布局时识别效果不佳。

基于注意力机制的方法则能够聚焦关键信息，在复杂场景中展现出更高的识别精度和鲁棒性。MORAN(multi-object rectified attention network)(Luo等，2019)结合文本的视觉和语言信息，通过多模态特征融合提高了识别准确性，并增强了模型对不规则布局和多样字体的泛化能力。然而，这种多模态方法的训练和推理过程较为复杂，且计算成本高。为此，ASTER(attentional scene text recognizer)(Shi等，2019)引入空间变换网络来矫正文本图像的几何形变，简化了识别网络架构并降低了计算成本。此外，ASTER通过字符级注意力机制，进一步提升了模型对复杂场景文本图像的识别能力。

尽管这些STR方法一定程度上提高了识别性能，但在处理真实LR图像时仍面临挑战。因此，利用STISR技术增强图像再识别文本成为解决复杂STR任务的一种有效方法。

2 所提方法

本文提出一种基于文本结构动态感知的跨模态融合STISR方法，整体框架如图4所示。对于输入的LR文本图像，首先经空间变换网络和薄板样条插值矫正图像，便于后续处理。随后输入图像分别通过文本结构动态感知模块和语义空间对齐模块获取图像模态和文本模态特征。文本结构动态感知模块通过方向感知层和上下文关联单元捕获文本结构动

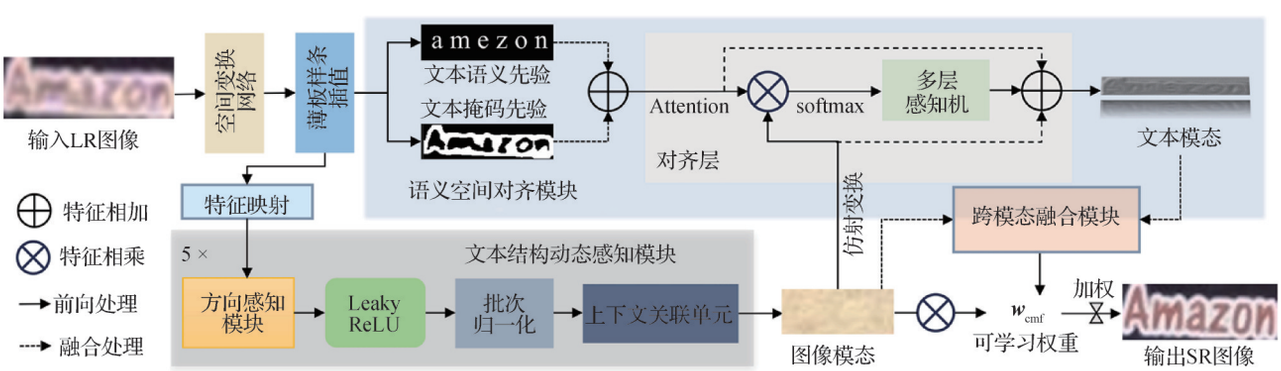

图4 所提方法的整体架构

Fig. 4 Overall architecture of the proposed method

态特征，促进图像模态的生成。语义空间对齐模块对LR图像识别渲染、二值化操作等处理得到文本语义先验。随后通过仿射变换，并在图像模态的指导下对齐并生成文本模态。最后，文本和图像模态通过跨模态融合模块加权输出最终的SR文本图像。

2.1 文本结构动态感知模块

文本图像中的字符通常以连续排列的形式呈现，在水平方向上具有强烈的上下文关联性、方向性和多尺度等结构动态特征。然而，现有方法多以常规 CNN 和 ViT（vision Transformer）（Dosovitskiy 等，2020）提取文本信息的局部和全局特征，难以兼顾水平方向上的文本结构动态特征，在处理连续字符时受到限制。为此，设计文本结构动态感知模块，采用 3 × 5、1 × 5、3 × 7 和 1 × 3 多个非对称卷积核，通过方向感知的方式动态捕捉文本的多尺度定向特征，同时通过上下文关联单元加强字符间的依赖性与上下文联系。相较于常规的 3 × 3 和 5 × 5 卷积核，方向感知模块以多尺度条状卷积为基础，在水平方向上具有更大的感受野。该策略能够更有效地保留字符邻域间的局部细节特征，有助于促进字符邻域间的上下文理解，其框架如图 5 所示。结合上下文关联单元，文本结构动态感知模块能够充分捕捉连续字符之间的空间关系与排列特征，从而更好地理解图像中的文本特征。

具体而言，方向感知层由 1 × 3、1 × 5、3 × 5 和 3 × 7 这 4 种不同尺度的条状卷积分支组成。矫正图像 $\boldsymbol{I} \in \mathbf{R}^{H\times W\times C}$ 首先通过特征映射层快速捕获图像特征，表示为 $\tilde{\boldsymbol{I}} \in \mathbf{R}^{H\times W\times C}$。特征映射层包含初始特征提取网络（Guo 等，2023）、归一化层、3 × 3 卷积和 LeakyReLU 激活函数。随后，映射特征 $\tilde{\boldsymbol{I}}$ 被输入 4 条多尺度卷积分支并得到 4 种特征，采用 LeakyReLU 激活函数增强特征的非线性表征能力，表示为 $\tilde{\boldsymbol{I}}'_{k\in[1,4]}$。上述流程可表示为

$$\tilde{\boldsymbol{I}} = \sigma_1\left\{f^{c}_{3\times 3}\left[LN(\boldsymbol{I})\right]\right\} \tag{1}$$

$$\begin{cases} \tilde{\boldsymbol{I}}'_1 = \sigma_1\left\{f^{sc}_{1\times 3}\left[\tilde{\boldsymbol{I}}\right]\right\} \\ \tilde{\boldsymbol{I}}'_2 = \sigma_1\left\{f^{sc}_{1\times 5}\left[\tilde{\boldsymbol{I}}\right]\right\} \\ \tilde{\boldsymbol{I}}'_3 = \sigma_1\left\{f^{sc}_{3\times 5}\left[\tilde{\boldsymbol{I}}\right]\right\} \\ \tilde{\boldsymbol{I}}'_4 = \sigma_1\left\{f^{sc}_{3\times 7}\left[\tilde{\boldsymbol{I}}\right]\right\} \end{cases} \tag{2}$$

式中，σ_1 表示 LeakyReLU 激活函数，f^{c} 表示普通卷积，f^{sc} 表示条状卷积，LN 表示层归一化。

在方向感知层中，大卷积核分支能够捕获更广泛的上下文信息，而小卷积核分支由于其较小的感受野，可以更精细地捕获局部细节。为了充分利用这些多尺度特征并增强字符之间的多尺度关联性，对不同分支的特征进行加权融合，表示为 $\tilde{\boldsymbol{I}}''_{k\in[1,4]}$。随后降维压缩通道信息，动态捕获多尺度感受野特征并去除特征冗余，具体为

$$\begin{cases} \tilde{\boldsymbol{I}}''_1 = f^{d}\left\{Cat\left[\tilde{\boldsymbol{I}}'_1, \alpha\tilde{\boldsymbol{I}}'_2, \beta\tilde{\boldsymbol{I}}'_3, \gamma\tilde{\boldsymbol{I}}'_4\right]\right\} \\ \tilde{\boldsymbol{I}}''_2 = f^{d}\left\{Cat\left[\alpha\tilde{\boldsymbol{I}}'_1, \tilde{\boldsymbol{I}}'_2, \beta\tilde{\boldsymbol{I}}'_3, \gamma\tilde{\boldsymbol{I}}'_4\right]\right\} \\ \tilde{\boldsymbol{I}}''_3 = f^{d}\left\{Cat\left[\alpha\tilde{\boldsymbol{I}}'_1, \beta\tilde{\boldsymbol{I}}'_2, \tilde{\boldsymbol{I}}'_3, \gamma\tilde{\boldsymbol{I}}'_4\right]\right\} \\ \tilde{\boldsymbol{I}}''_4 = f^{d}\left\{Cat\left[\alpha\tilde{\boldsymbol{I}}'_1, \beta\tilde{\boldsymbol{I}}'_2, \gamma\tilde{\boldsymbol{I}}'_3, \tilde{\boldsymbol{I}}'_4\right]\right\} \end{cases} \tag{3}$$

式中，f^{d} 表示降维，Cat 表示特征拼接，α、β 和 γ 表示可学习权重系数。

随后，4 条分支再次通过多尺度卷积和 LeakyReLU 激活函数，以充分捕捉文本图像的结构动态特征，表示为 $\tilde{\boldsymbol{I}}'''_{k\in[1,4]}$。在方向感知模块的末端，拼接 4 层不同尺度特征，采用 1 × 1 卷积降维，增强跨通道的空间特征交互，并通过残差连接映射特征 $\tilde{\boldsymbol{I}}$ 得到输出 $\bar{\boldsymbol{I}}$，具体为

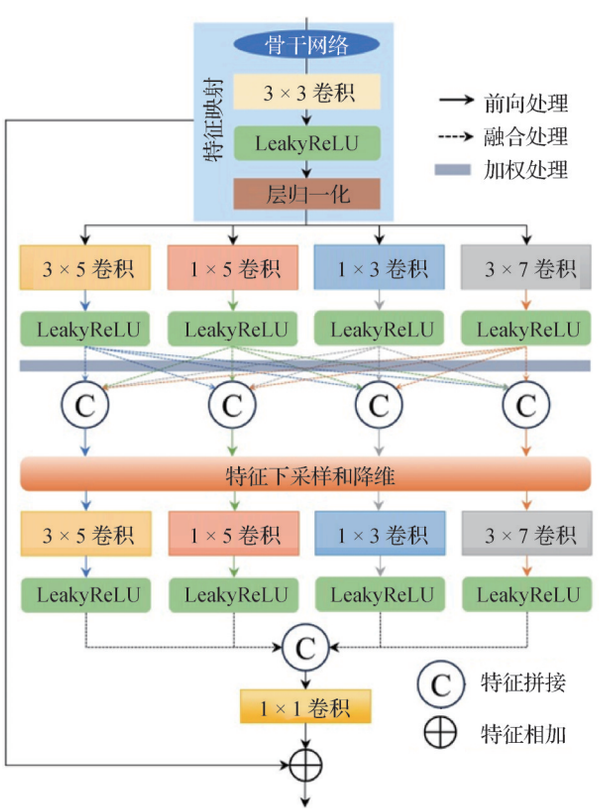

图 5 方向感知层的流程图

Fig. 5 Flowchart of the direction sensing block

$$\begin{cases}\tilde{\boldsymbol{I}}_1''' = \sigma_1\left\{f_{1\times3}^{\mathrm{sc}}\left[\tilde{\boldsymbol{I}}_1''\right]\right\}\\ \tilde{\boldsymbol{I}}_2''' = \sigma_1\left\{f_{1\times5}^{\mathrm{sc}}\left[\tilde{\boldsymbol{I}}_2''\right]\right\}\\ \tilde{\boldsymbol{I}}_3''' = \sigma_1\left\{f_{3\times5}^{\mathrm{sc}}\left[\tilde{\boldsymbol{I}}_3''\right]\right\}\\ \tilde{\boldsymbol{I}}_4''' = \sigma_1\left\{f_{3\times7}^{\mathrm{sc}}\left[\tilde{\boldsymbol{I}}_4''\right]\right\}\end{cases} \tag{4}$$

$$\bar{\boldsymbol{I}} = \tilde{\boldsymbol{I}} + f_{1\times1}^{\mathrm{c}}\left\{Cat\left[\tilde{\boldsymbol{I}}_1''', \tilde{\boldsymbol{I}}_2''', \tilde{\boldsymbol{I}}_3''', \tilde{\boldsymbol{I}}_4'''\right]\right\} \tag{5}$$

上下文关联单元由GRU(gated recurrent unit)层和MLP层组成,用于学习字符邻域间的长距离依赖关系,以及上下文关联信息。GRU层的输入首先经过激活函数和归一化层,以降低计算复杂度并提升特征的非线性表达能力。同时GRU层能够捕获邻近字符区域的长距离依赖关系,并通过MLP层增强表征能力,输出图像模态特征 $\boldsymbol{I}_V \in \mathbf{R}^{H\times W\times C}$, 具体为

$$\bar{\boldsymbol{I}} = LN\left\{\sigma_1\left[\bar{\boldsymbol{I}}\right]\right\} \tag{6}$$

$$\boldsymbol{I}_v = MLP\left\{GRU\left[\bar{\boldsymbol{I}}\right]\right\} \tag{7}$$

2.2 语义空间对齐模块

文本图像中的语义先验是重要的先验知识,对图像恢复起重要指导作用。然而,现有的STISR方法在利用文本语义先验时往往缺乏精细化处理,通常仅将语义先验简单渲染,随后嵌入图像特征。这类方法忽略了不同特征之间存在的不一致性,导致文本图像在恢复过程中易出现语义混淆和信息丢失的问题。为此,构建语义空间对齐模块,整合多重先验信息来增强语义先验的准确性,并在图像模块的引导下对语义先验仿射变换,以增强文本和图像模态的特征一致性。

受Zhu等人(2024)的启发,通过文本结构掩码先验辅助语义先验的生成,以增强文本语义先验的准确性。一方面,矫正图像通过二值化处理得到文本的结构掩码先验 $\boldsymbol{I}_m \in \mathbf{R}^{H\times W\times C}$;另一方面,矫正的LR图像 $\boldsymbol{I} \in \mathbf{R}^{H\times W\times C}$通过预训练的文本识别器,得到识别文本 $t \in \mathbf{R}^{L\times D}$,然后将其渲染至相同特征图大小得到语义先验 $\boldsymbol{I}_s \in \mathbf{R}^{H\times W\times C}$。随后将两重先验特征逐元素相加获取精细化的文本语义先验 $\boldsymbol{I}_t \in \mathbf{R}^{H\times W\times C}$。

由于图像模态特征源于真正的LR文本图像,包含更丰富的原始特征,因此对其进一步利用。通过对齐层将语义先验和图像模态对齐,以确保特征一致。对齐层通过Attention机制计算先验特征和图像模态的相似性矩阵,表示为 $\boldsymbol{D}$。将矩阵 $\boldsymbol{D}$ 与语义先验相乘得到初始对齐文本先验,表示为 $\boldsymbol{I}_t'$,具体表达为

$$\boldsymbol{D} = \sigma_2\left\{Att\left[\boldsymbol{I}_t, \boldsymbol{I}_V\right]\right\} \tag{8}$$

$$\boldsymbol{I}_t' = \boldsymbol{I}_t \odot \boldsymbol{D} \tag{9}$$

式中,σ_2表示sigmoid激活函数,$\odot$表示矩阵相乘。

随后,通过仿射变换进一步对齐特征。仿射变换由多层感知机生成变换参数,使得语义先验在变换过程逐渐适应图像模态特征。具体流程如下,首先将初始对齐文本先验 $\boldsymbol{I}_t'$与图像模态特征 $\boldsymbol{I}_V$ 拼接,表示为 $\boldsymbol{M}$。特征 $\boldsymbol{M}$ 综合了两种模态特征,具有更强的特征表达能力。在对齐层的末端,以 $\boldsymbol{M}$ 作为学习参数通过MLP进行学习,生成变换矩阵 $\boldsymbol{W}$ 和偏移向量 $\boldsymbol{B}$。最后,应用仿射变换实现文本语义先验对图像特征的对齐,表示为 $\boldsymbol{I}_T'$,具体表达为

$$\boldsymbol{M} = Cat\left[\boldsymbol{I}_t', \boldsymbol{I}_V\right] \tag{10}$$

$$\begin{cases}\boldsymbol{W} = f_a\left[\boldsymbol{M}\right]\\ \boldsymbol{B} = f_b\left[\boldsymbol{M}\right]\end{cases} \tag{11}$$

$$\boldsymbol{I}_T' = \boldsymbol{I}_t'\boldsymbol{W} + \boldsymbol{B} \tag{12}$$

式中,f_a和f_b表示不同的MLP。

为了减少在对齐层中的信息丢失,对初始对齐文本先验和图像模态再次利用,输出文本模态,具体表达为

$$\boldsymbol{I}_T = \boldsymbol{I}_T' + \mu_1\boldsymbol{I}_V + \mu_2\boldsymbol{I}_t \tag{13}$$

式中,$\boldsymbol{I}_T$ 表示对齐的文本模态,μ_1和μ_2表示融合系数,为学习参数。

2.3 跨模态融合模块

跨模态融合模块包含全局分支、局部分支和可学习权重分支3个分支,如图6所示。该模块融合文本特征的全局和局部信息,能够自适应地调整不同模态特征的贡献,实现有效的跨模态特征整合。全局分支首先通过全局平均池化将特征图每个通道的所有像素值进行平均化,得到一个全局特征向量。通过降低空间维度,使得后续的卷积层能够更专注于通道之间的关系。局部分支则通过多层卷积直接处理特征图,保留了更多的局部空间细节。可学习权重分支结合了全局与局部特征,通过综合两种特征得到更加丰富的特征表示,从而确定各模态的权重占比。

具体来说,跨模态融合模块首先将两种模态特征进行拼接,表示为 $\boldsymbol{I}_{T,V}$,作为后续学习样本。随后将 $\boldsymbol{I}_{T,V}$ 传入全局分支和局部分支,全面捕捉样本信

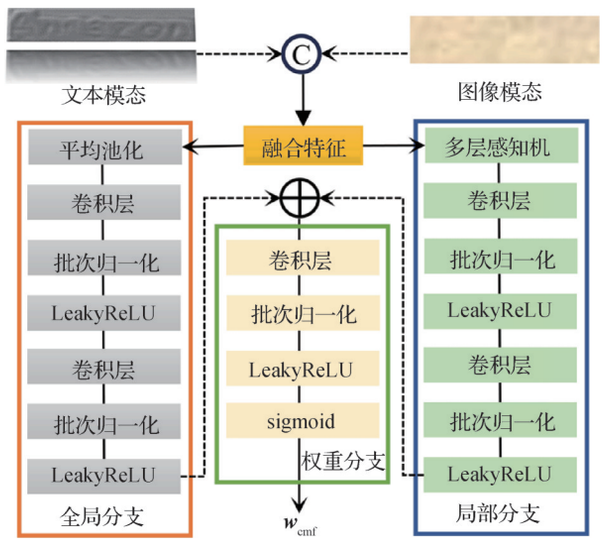

图6　跨模态融合模块的流程图

Fig. 6　Flowchart of the cross modal fusion module

息。全局分支通过全局平均池化压缩特征到1×1的大小，随后通过卷积变换网络学习精细化特征表示，表示为 $\boldsymbol{I}_{T,V}^{g}$。局部分支则直接通过卷积变换网络学习样本的局部结构特征，表示为 $\boldsymbol{I}_{T,V}^{l}$。接下来，通过可学习权重分支生成融合权重。该分支接收 $\boldsymbol{I}_{T,V}^{g}$ 和 $\boldsymbol{I}_{T,V}^{l}$ 作为输入，输出融合权重 $\boldsymbol{w}_{cmf}$。最后，将文本和视觉模态信息加权融合。考虑到文本和图像两种模态信息并非简单的线性关系，因此采用相似性融合，具体表达为

$$\boldsymbol{I}_{T,V}=Cat\left[\boldsymbol{I}_{T},\boldsymbol{I}_{V}\right] \tag{14}$$

$$\begin{cases}\boldsymbol{I}_{T,V}^{g}=f^{g}\left[\boldsymbol{I}_{T,V}\right]\\ \boldsymbol{I}_{T,V}^{l}=f^{l}\left[\boldsymbol{I}_{T,V}\right]\end{cases} \tag{15}$$

$$\boldsymbol{w}_{cmf}=f^{w}\left[\boldsymbol{I}_{T,V}^{g},\boldsymbol{I}_{T,V}^{l}\right] \tag{16}$$

$$\boldsymbol{I}_{out}=\sigma_{2}\left[\boldsymbol{w}_{cmf}\boldsymbol{I}_{V}\right]\odot\boldsymbol{I}_{V}+\sigma_{2}\left[\boldsymbol{w}_{cmf}\boldsymbol{I}_{T}\right]\odot\boldsymbol{I}_{T} \tag{17}$$

式中，f^{g}、f^{l} 和 f^{w} 分别表示全局分支、局部分支和可学习权重分支。

2.4　损失函数

本文通过像素损失、识别损失和微调损失对模型进行优化。像素损失通过 L_2 损失函数优化模型，具体计算为

$$\boldsymbol{L}_{pix}=\left\|\boldsymbol{X}_{HR}-\boldsymbol{X}_{SR}\right\|_{2} \tag{18}$$

式中，$\boldsymbol{X}_{HR}$ 和 $\boldsymbol{X}_{SR}$ 表示高分辨率图像和超分辨率图像。

识别损失通过文本焦点损失（Chen 等，2021）来优化文本语义先验，具体计算为

$$\boldsymbol{L}_{text}=\varphi_{1}\left\|\boldsymbol{A}_{HR}-\boldsymbol{A}_{SR}\right\|+\varphi_{2}WCE\left(t_{p},t_{l}\right) \tag{19}$$

式中，$\boldsymbol{A}$ 表示文本注意力分布图，WCE 表示加权交叉熵，φ_1 和 φ_2 表示可学习权重系数，t_p 和 t_l 表示预测文本和标签文本。

研究证明（Zhu 等，2023；Ma 等，2022），预训练文本识别器通过微调参数的性能更优，因此采用微调损失来优化预训练模型，具体计算为

$$\boldsymbol{L}_{ft}=CE\left(t_{p},t_{l}\right) \tag{20}$$

式中，CE 表示交叉熵。

总体损失表示为

$$\boldsymbol{L}_{total}=\boldsymbol{L}_{pix}+\theta_{1}\boldsymbol{L}_{text}+\theta_{2}\boldsymbol{L}_{ft} \tag{21}$$

式中，θ_1 和 θ_2 表示可学习权重系数。

3　实验结果与分析

3.1　实验细节

为了验证所提方法的有效性，本文在真实数据集 TextZoom 上进行了大量实验，并与多种主流方法对比。实验硬件环境为单个 24 GB 显存的 NVDIA TITAN RTX GPU 和 Xeon Silver 4214 CPU。软件环境基于 Ubuntu16.04 操作系统、Python3.8 和 PyTorch 1.8 深度学习框架。预训练的文本识别器采用 ABINet（autonomous，bidirectional and iterative network）（Fang 等，2021）。超参数设置如下：实验均训练 600 轮，批次大小为 64，采用 Adam 优化器，学习率设置为 0.001，并在 400 轮之后衰减 0.5 倍。

3.2　数据集和评价标准

TextZoom 是目前唯一针对真实场景的 STISR 数据集，与合成数据相比，更贴近真实环境中的文本细节，识别难度更高。该数据集包含 17 367 对 LR-HR 训练图像对和 4 373 对验证图像对，尺寸分别为 16×64 像素和 32×128 像素。验证集按照难度分为简单（1 619 对）、中等（1 411 对）和困难（1 343 对）3 个等级，以评估不同复杂度场景下的文本图像恢复效果。由于合成数据和真实数据的巨大差异性，如图 1 所示，因此仅在该数据集上进行训练和测试。

为了验证实验结果的有效性和代表性，遵循 STISR 领域的公认标准（Wang 等，2020），使用 ASTER、CRNN 和 MORAN 3 种文本识别器进行评估。以文本识别准确率（accuracy，Acc）为主要指标，同时补充峰值信噪比（peak signal-to-noise ratio，PSNR）和结构相似性（structural similarity index，

SSIM)等SR领域的常用指标,全面评估性能。由于STISR的根本任务是提升文本可读性,因此以识别准确率作为最重要的评价标准。而PSNR和SSIM受到其自身局限性,在复杂场景下无法完全反映人眼的感知质量,特别对关键点集中在文本区域的文本图像而言,表现出较明显的不足,因此将这两项指标作为辅助评价标准。

3.3 对比实验与分析

将所提方法与多种主流方法对比分析,包括经典方法SRCNN(super-resolution using deep convolutional network)(Dong等,2014)、LAPSRN(Laplacian pyramid network for fast and accurate super-resolution)(Tran和Ho-Phuoc,2019)、BICUBIC,先进方法TSRN(text super-resolution network)(Wang等,2020)、TBSRN(Transformer-based super-resolution network)(Chen等,2021)、TG(text gestalt)(Chen等,2022)、TPGSR(text prior guided scene text image super-resolution)(Ma等,2023)、TATT(text attention network)(Ma等,2022)、DPMN(dual prior modulation network)(Zhu等,2023)、MSPIE(multiscale prior interaction enhancement)(Zhu等,2024)、PAM(pixel adapter module)(Zhang等,2023)、TEAN(text enhanced attention network)(Shu等,2023)和PERMR(perceiving multiple representations)(Shi等,2023),文本识别精度对比结果如表1所示。同时,将PSNR和SSIM作为补充评价标准,对比结果如表2表示。图7展示了不同方法的SR的可视化效果。可以看出,与其他优秀方法相比,所提方法在文本细节重建方面表现出更优越的性能。

从上述对比实验可以看出,所提方法较其他先进方法具有较大优势,在整体识别精度上高于次优方法2.8%。这主要归因于先前研究在文本结构动态特征方面的关注不足,导致文本的内在特性挖掘不够充分,缺乏足够的文本细节信息,进而影响了重建图像的质量。尤其自TPGSR首次利用文本的语义先验知识取得一定成功后,大量方法在TPGSR的基础上对语义先验开发利用。这些方法过多地关注文本语义先验,忽视了图像中内在文本结构动态特征,导致字符特征缺乏与其邻域之间的上下文依赖关系,从而影响了模型对连续字符特征的理解。此

表1 TextZoom数据集上不同方法识别准确率的比较结果

Table 1 Comparison of recognition accuracy rates for different methods on the TextZoom dataset

/%

方法	ASTER				CRNN				MORAN				平均值
	简单	中等	困难	均值	简单	中等	困难	均值	简单	中等	困难	均值	
BICUBIC	64.7	42.4	31.2	47.2	36.4	21.1	21.1	26.8	60.6	37.9	30.8	44.1	39.3
SRCNN	69.0	43.4	32.1	49.4	38.7	21.6	20.8	27.7	63.2	39.0	30.1	45.2	40.7
LAPSRN	71.5	48.6	35.2	53.0	46.1	27.9	23.6	33.3	64.6	44.9	32.2	48.3	44.8
TSRN	75.1	56.3	40.1	58.2	52.5	38.2	31.4	41.4	70.1	53.3	37.9	54.7	51.5
TBSRN	75.7	59.9	41.6	60.1	59.6	47.1	35.3	48.1	74.1	57	40.8	58.3	55.5
TG	77.9	60.2	42.4	61.2	61.2	47.6	35.5	48.9	75.8	57.8	41.4	59.4	56.5
TPGSR	78.9	62.7	44.5	63.1	63.1	52.0	38.6	52.0	74.9	60.5	44.1	60.7	58.6
TATT	78.9	63.4	45.4	63.6	62.6	53.3	39.8	52.6	72.5	60.2	43.1	59.5	58.5
DPMN	79.1	63.2	45.1	63.5	63.4	54.1	39.5	53.0	73.2	61.0	43.8	60.2	58.9
MSPIE	80.4	63.4	46.3	64.4	64.5	54.2	39.6	53.5	74.0	61.4	44.4	60.8	59.6
PAM	79.8	61.2	47.0	63.7	65.1	55.5	38.8	53.9	75.5	60.8	44.3	61.1	59.6
TEAN	80.4	64.5	45.6	64.5	63.7	52.5	38.1	52.2	76.8	60.8	43.4	61.3	59.3
PERMR	80.8	62.9	45.5	64.1	65.1	50.4	37.8	52.0	76.7	58.9	42.9	60.5	58.9
本文	**82.6**	**67.1**	**48.6**	**67.1**	**66.9**	**59.1**	**41.6**	**56.6**	**77.9**	**65.2**	**44.8**	**63.6**	**62.4**

注:加粗字体表示各列最优结果。

表 2　TextZoom 数据集上 Acc 及补充指标的比较结果

Table 2　Comparison of the Acc and supplementary metric of different methods on the TextZoom dataset

方法	年份	Acc/%	PSNR/dB	SSIM
BICUBIC	–	39.3	20.4	0.696
SRCNN	ECCV'2014	40.7	20.8	0.723
LAPSRN	CVPR'2017	44.8	21.3	0.744
TSRN	ECCV'2020	51.5	20.8	0.759
TBSRN	CVPR'2021	55.5	20.9	0.763
TG	AAAI'2022	56.5	18.8	0.660
TPGSR	TIP'2023	58.6	21.2	0.762
TATT	CVPR'2022	58.5	21.2	0.783
DPMN	AAAI'2023	58.9	21.3	0.784
MSPIE	TAI'2024	59.6	21.5	0.789
PAM	MM'2023	59.6	–	–
TEAN	TCSVT'2023	59.3	21.7	0.785
PERMR	EAAI'2023	58.9	21.6	**0.796**
本文	JIG'2025	**62.4**	**21.9**	0.789

注：加粗字体表示各列最优结果，"–"表示未提及。

外，大多数方法忽视了语义先验和图像模态之间的特征不一致性问题，使得语义先验和文本图像的跨模态融合不够充分，最终导致重建图像质量较低。而所提方法在一定程度上恰当地缓解了上述问题，在有效提取并利用文本结构动态特征的同时，着重于利用图像特征引导语义先验的对齐，并促进二者的交互融合，实现精细化的 LR 文本图像重建。

3.4　消融实验

为了探索所提方法中不同模块的有效性，在 TextZoom 数据集上对跨模态融合方式、文本结构感知模块中的方向感知块卷积核大小等进行多次消融实验。

3.4.1　融合方式的消融实验

在文本模态和图像模态的融合方面，采用跨模态交互融合。对不同融合方式进行消融实验，包括特征逐元素相加和特征拼接等融合方式。对于跨模态融合模块的输出 $\boldsymbol{w}_{\mathrm{cmf}}$，视其为融合权重，以进一步拟合两种模态特征，在最终输出阶段通过相似度加权融合处理得到 SR 图像。与此同时，$\boldsymbol{w}_{\mathrm{cmf}}$ 也可直接视为文本和图像模态的融合输出，因此也将该方式的输出加入消融实验。此外，在融合过程中，采用了自适应的学习权重用于跨模态融合。因此，也对学习权重进行消融实验。表 3 的实验结果表明，所提的融合方式能够达到最优效果。

3.4.2　多尺度条状卷积的消融实验

在文本结构动态感知模块的方向感知层中，设计了 4 条多尺度卷积分支，以感知字符的方向性特征。4 条分支所采用的卷积核大小分别为 1 × 3、1 ×

图 7　与其他主流方法的可视化对比

Fig. 7　Visualization comparison with other mainstream methods

表3 关于融合方式的消融实验

Table 3 Ablation experiments on fusion approaches

融合方式	Acc/%		PSNR/dB		SSIM	
	w/	w/o	w/	w/o	w/	w/o
特征相加	60.8	60.6	21.6	21.2	0.784	0.781
特征拼接	59.7	59.1	21.7	21.2	0.786	0.783
$\boldsymbol{w}_{\mathrm{cmf}}$(权重)	**62.4**	61.6	**21.9**	21.9	0.789	0.788
$\boldsymbol{w}_{\mathrm{cmf}}$(输出)	61.4	61.5	21.7	21.5	**0.792**	0.788

注:加粗字体表示各组评价指标最优结果。w/、w/o分别表示使用、不使用可学习权重。

5、3×5和3×7,均为条状卷积。为了验证这些卷积分支的有效性,以至少两个卷积分支为组合进行消融实验(经验表明单个卷积分支效果不理想)。实验结果如表4所示,当采用4层条状卷积时,方法性能达到最佳。

3.4.3 文本结构动态感知模块的迭代次数

文本结构动态感知模块通过连续迭代循环方式提炼文本模态的结构动态特征。对该模块的迭代次数进行消融实验,实验结果如表5所示。可以看出,当迭代次数为5时效果最佳。

3.4.4 先验信息的使用

在语义空间对齐模块中,文本掩码先验用于辅助文本语义先验的生成。对结构先验和语义先验进行消融实验以证明所提炼先验信息的有效性。从表6看出,当两个先验同时使用并采取特征相加的方式获取语义先验时,所提方法能够达到最优效果。

3.5 局限性

尽管所提方法在大多数情况下表现优异,但在处理极端模糊的LR文本图像时仍存在局限性。由于这些图像的分辨率极低,信息严重丢失,无法提供足够的先验知识,进而限制了模型的学习能力,使其难以有效重建图像。这一限制也普遍存在于其他STISR方法,目前尚未有良好的解决方案。

表4 多尺度条状卷积的消融实验

Table 4 Ablation experiments on multi-scale strip convolution

卷积分支数	卷积核大小				Acc/%	PSNR/dB	SSIM
	1×3	1×5	3×5	3×7			
2	√	–	–	√	58.3	21.4	0.773
	√	–	√	–	58.7	21.3	0.762
	√	√	–	–	59.4	20.9	0.765
	–	√	–	√	59.4	21.2	0.784
	–	√	√	–	61.1	21.5	0.784
	–	–	√	√	60.4	21.4	0.774
3	√	√	√	–	61.9	21.7	0.766
	√	√	–	√	62.1	21.7	0.762
	√	–	√	√	60.5	21.8	0.779
	–	√	√	√	61.5	21.7	**0.790**
4	√	√	√	√	**62.4**	**21.9**	0.789

注:加粗字体表示各列最优结果,“√”表示使用,“–”表示未使用。

表5 文本结构感知模块的迭代次数

Table 5 The number of iterations of the text structure sensing module

迭代次数	Acc/%	PSNR/dB	SSIM
2	49.3	17.8	0.687
3	55.6	19.4	0.714
4	59.9	21.3	0.783
5	**62.4**	**21.9**	**0.789**
6	61.4	21.3	0.788

注:加粗字体表示各列最优结果。

表6 文本先验信息的使用

Table 6 The use of text prior information

	文本语义先验	文本掩码先验	Acc/%	PSNR/dB	SSIM
	×	√	59.1	20.9	0.714
	√	×	60.7	21.6	0.724
	×	×	56.9	19.3	0.702
特征逐元素相加	√	√	**62.4**	**21.9**	0.789
特征逐元素拼接	√	√	61.7	21.7	**0.791**

注:加粗字体表示各列最优结果,“√”表示使用,“×”表示不使用。

4 结 论

本文提出一种基于文本结构动态感知的跨模态融合的STISR方法，旨在解决现有方法在特征提取中难以有效捕捉文本结构动态特征的问题，并克服文本和图像模态特征不一致及难以有效融合的不足。所提方法使用文本结构动态感知模块和语义空间对齐模块分别促进图像和文本模态的生成，并通过跨模态融合模块实现高效的信息交互融合，显著提高了文本图像的重建质量和文本可读性。其中，文本结构动态感知模块通过方向感知模块和上下文关联单元生成具有文本结构动态特征的图像模态。语义空间对齐模块结合文本结构掩码生成精细化文本语义先验，并在图像模态的引导下实现跨模态对齐。在真实数据集TextZoom上的实验结果表明，本文方法在3种标准文本评估器上的平均精度为62.4%，超越性能第2的方法2.8%，证明了所提方法在恢复图像质量和文本可读性方面具有显著优势。

未来计划通过利用更先进的分割技术来获取更准确的文本掩码先验，增强先验信息对重建LR图像的指导能力。同时，计划开发新的真实STISR数据集，缓解目前真实数据缺乏的问题。

参考文献（References）

Chang H, Yeung D Y and Xiong Y M. 2004. Super-resolution through neighbor embedding//Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE [DOI: 10.1109/CVPR.2004.1315043]

Chen J Y, Li B and Xue X Y. 2021. Scene text telescope: text-focused scene image super-resolution//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12026-12035 [DOI: 10.1109/CVPR46437.2021.01185]

Chen J Y, Yu H Y, Ma J Q, Li B and Xue X Y. 2022. Text gestalt: stroke-aware scene text image super-resolution//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 285-293 [DOI: 10.1609/aaai.v36i1.19904]

Dong C, Loy C C, He K M and Tang X O. 2014. Learning a deep convolutional network for image super-resolution//Proceedings of the 13th European Conference on Computer Vision——ECCV 2014. Zurich, Switzerland: Springer: 184-199 [DOI: 10.1007/978-3-319-10593-2_13]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. [2024-08-12]. https://arxiv.org/pdf/2010.11929.pdf

Fang S C, Xie H T, Wang Y X, Mao Z D and Zhang Y D. 2021. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7098-7107 [DOI: 10.1109/CVPR46437.2021.00702]

Farsiu S, Robinson D, Elad M and Milanfar P. 2004. Advances and challenges in super-resolution. International Journal of Imaging Systems and Technology, 14(2): 47-57 [DOI: 10.1002/ima.20007]

Glasner D, Bagon S and Irani M. 2009. Super-resolution from a single image//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 349-356 [DOI: 10.1109/ICCV.2009.5459271]

Graves A, Fernández S, Gomez F and Schmidhuber J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM: 369-376 [DOI: 10.1145/1143844.1143891]

Guo H, Dai T, Meng G H and Xia S T. 2023. Towards robust scene text image super-resolution via explicit location enhancement//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: ACM: 782-790 [DOI: 10.24963/ijcai.2023/87]

Hao Y D, Madani S, Guan J F, Alloulah M, Gupta S and Hassanieh H. 2024. Bootstrapping autonomous driving radars with self-supervised learning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 15012-15023 [DOI: 10.1109/CVPR52733.2024.01422]

Li X and Orchard M T. 2001. New edge-directed interpolation. IEEE Transactions on Image Processing, 10(10): 1521-1527 [DOI: 10.1109/83.951537]

Luo C J, Jin L W and Sun Z H. 2019. MORAN: a multi-object rectified attention network for scene text recognition. Pattern Recognition, 90: 109-118 [DOI: 10.1016/j.patcog.2019.01.020]

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

Ma J Q, Liang Z T and Zhang L. 2022. A text attention network for spatial deformation robust scene text image super-resolution//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 5911-5920 [DOI: 10.1109/CVPR52688.2022.00582]

Pourkeshavarz M, Zhang J R and Rasouli A. 2024. CaDeT: a causal disentanglement approach for robust trajectory prediction in autono-

mous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14874-14884 [DOI: 10.1109/CVPR52733.2024.01409]

Qiao L, Li Z S, Cheng Z Z and Li X. 2023. SCID: a Chinese characters invoice-scanned dataset in relevant to key information extraction derived of visually-rich document images. Journal of Image and Graphics, 28(8): 2298-2313 (乔梁, 李再升, 程战战, 李玺. 2023. SCID: 用于富含视觉信息文档图像中信息提取任务的扫描中文票据数据集. 中国图象图形学报, 28(8): 2298-2313) [DOI: 10.11834/jig.220911]

Qin H, Li Y J, Liang Q K and Wang Y N. 2023. AsymcNet: a document images-relevant asymmetric geometry correction network. Journal of Image and Graphics, 28(8): 2314-2329 (秦海, 李艺杰, 梁桥康, 王耀南. 2023. 针对文档图像的非对称式几何校正网络. 中国图象图形学报, 28(8): 2314-2329) [DOI: 10.11834/jig.220426]

Shi B G, Bai X and Yao C. 2016. 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, 39(11): 2298-2304 [DOI: 10.1109/TPAMI.2016.2646371]

Shi B G, Yang M K, Wang X G, Lyu P Y, Yao C and Bai X. 2019. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2035-2048 [DOI: 10.1109/TPAMI.2018.2848939]

Shi Q, Zhu Y, Liu Y T, Ye J Y and Yang D W. 2023. Perceiving multiple representations for scene text image super-resolution guided by text recognizer. Engineering Applications of Artificial Intelligence, 124: #106551 [DOI: 10.1016/j.engappai.2023.106551]

Shu R, Zhao C R, Feng S Y, Zhu L and Miao D Q. 2023. Text-enhanced scene image super-resolution via stroke mask and orthogonal attention. IEEE Transactions on Circuits and Systems for Video Technology, 33(11): 6317-6330 [DOI: 10.1109/TCSVT.2023.3267133]

Tran H T M and Ho-Phuoc T. 2019. Deep Laplacian pyramid network for text images super-resolution//Proceedings of 2019 IEEE-RIVF International Conference on Computing and Communication Technologies. Danang, Vietnam: IEEE: 1-6 [DOI: 10.1109/RIVF.2019.8713657]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Wang W J, Xie E Z, Liu X B, Wang W H, Liang D, Shen C H and Bai X. 2020. Scene text image super-resolution in the wild//Proceedings of the 16th European Conference on Computer Vision—ECCV 2020. Glasgow, UK: Springer: 650-666 [DOI: 10.1007/978-3-030-58607-2_38]

Yang J C, Wright J, Huang T S and Ma Y. 2010. Image super-resolution via sparse representation. IEEE Transactions on Image Processing, 19(11): 2861-2873 [DOI: 10.1109/TIP.2010.2050625]

Zhang W Y, Deng X, Jia B J, Yu X T, Chen Y F, Ma J, Ding Q and Zhang X M. 2023. Pixel adapter: a graph-based post-processing approach for scene text image super-resolution//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 2168-2179 [DOI: 10.1145/3581783.3611913]

Zhu S P, Zhao Z Y, Fang P F and Xue H. 2023. Improving scene text image super-resolution via dual prior modulation network. Proceedings of the AAAI Conference on Artificial Intelligence, 37(3): 3843-3851 [DOI: 10.1609/aaai.v37i3.25497]

Zhu Z J, Zhang L, Bai Y Q, Wang Y E and Li P. 2024. Scene text image superresolution through multiscale interaction of structural and semantic priors. IEEE Transactions on Artificial Intelligence, 5(7): 3653-3663 [DOI: 10.1109/TAI.2024.3375836]

作者简介

朱仲杰，男，教授，主要研究方向为2D和3D视频编码与传输、高动态图像视频编码及处理。
E-mail：zhongjiezhu@yeah.net
张磊，男，硕士研究生，主要研究方向为计算机视觉与图像处理。E-mail：zhlangl01@hotmail.com
李沛，男，硕士研究生，主要研究方向为图像视频处理及异常事件检测。E-mail：2022882023@zwu.edu.cn
屠仁伟，男，博士，讲师，主要研究方向为点云质量评价。
E-mail：renweitu@126.com
白永强，男，副教授，主要研究方向为信息隐藏、图像处理和人工智能。E-mail：byq-163@163.com
王玉儿，女，助理研究员，主要研究方向为视频编码和信息隐藏。E-mail：365401628@qq.com