引文格式:韦豪东,易尧华,余长慧,等.融合注意力与序列单元的文本超分辨率[J].武汉大学学报(信息科学版),2024,49(7):1120-1129.DOI:10.13203/j.whugis20220158

Citation: WEI Haodong, YI Yaohua, YU Changhui, et al. Text Super-Resolution Method with Attentional Mechanism andSequential Units[J]. Geomatics and Information Science of Wuhan University, 2024, 49(7): 1120-1129. DOI: 10.13203/j.whugis20220158

融合注意力与序列单元的文本超分辨率

韦豪东[1] 易尧华[1] 余长慧[1] 林立宇[2]

1 武汉大学遥感信息工程学院,湖北 武汉,430079

2 武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

摘 要:街景影像中的文本信息是感知与理解场景的关键线索,低分辨率街景影像文本区域细节缺乏导致文本识别准确率降低。文本超分辨率通过增强文本区域边缘及纹理细节提高文本识别准确率,提出了融合注意力与序列单元的街景影像文本超分辨率方法。首先,采用混合残差注意力结构提取影像文本区域空间信息、通道信息并融合特征,序列单元通过双向门控循环结构提取影像中文本间的序列先验信息;然后利用梯度先验知识作为约束条件,重构街景影像文本区域。采用TextZoom真实场景影像及合成文本影像进行对比分析,实验结果表明,超分辨率重构的街景影像文本区域边缘清晰、纹理细节丰富,可以提高街景影像文本识别准确率。

关键词:街景影像;超分辨率;注意力机制;序列信息;梯度先验损失

中图分类号:P237 **文献标识码**:A **收稿日期**:2022-07-11

DOI:10.13203/j.whugis20220158 **文章编号**:1671-8860(2024)07-1120-10

Text Super-Resolution Method with Attentional Mechanism and Sequential Units

WEI Haodong[1] *YI Yaohua*[1] *YU Changhui*[1] *LIN Liyu*[2]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

2 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: **Objectives**: The text in street view images is the clue to perceive and understand scene information. Low-resolution street view images lack details in the text region, leading to poor recognition accuracy. Super-resolution can be introduced as pre-processing to reconstruct edge and texture details of the text region. To improve text recognition accuracy, we propose a text super-resolution network combining attentional mechanism and sequential units. **Methods**: A hybrid residual attentional structure is proposed to extract spatial information and channel information of the image text region, learning multi-level feature representation. A sequential unit is proposed to extract sequential prior information between texts in the image through bidirectional gated recurrent units. Using gradient prior knowledge as the constraint, a gradient prior loss is designed to sharpen character boundaries. **Results**: In order to verify the effectiveness of the proposed method, we use real scene text images in TextZoom and synthetic text images to carry out comparative analysis experiments. Experimental results show that compared with the baseline and state-of-the-art general super-resolution algorithm, our model reconstruct sharper text edges and clearer texture details in visual perception, and achieve higher recognition accuracy. **Conclusions**: Our method can make better use of the prior knowledge of text areas in images, which help reconstruct text details, improving accuracy of the text recognition task.

基金项目:国家重点研发计划(2021YFB2206200)。

第一作者:韦豪东,硕士,主要从事图像超分辨率重建及相关应用的理论与方法研究。WHUweihaodong@163.com

通讯作者:易尧华,博士,教授。yyh@whu.edu.cn

Key words: street view images; super-resolution; attentional mechanism; sequential information; gradient prior loss

街景影像中的文本包含丰富语义信息,是感知与理解场景的关键线索。街景影像中的文本通常形状不规则、背景复杂且分辨率低,这些都直接导致了街景影像文本识别准确率降低[1]。超分辨率重建算法对影像文本区域进行超分辨率预处理,通过提高影像文本区域清晰度提升文本识别准确率[2]。

基于浅层卷积神经网络(convolutional neural networks, CNN)的文本超分辨率算法学习低分辨率(low resolution, LR)文本到高分辨率(high resolution, HR)文本的映射。Dong 等[3]在 2015 年国际文档分析与识别比赛[4]文本超分辨率赛道上首次提出基于CNN的文本超分辨率重建方法(super-resolution CNN, SRCNN)[5]。文献[6]使用转置卷积和亚像素卷积对网络提取的特征图进行上采样,处理黑白文档超分辨率问题。文献[7]分别训练文本 SRCNN 分支与影像 SRCNN 分支,融合双分支提升场景图像超分辨率重建质量。这些方法卷积层数较少,无法充分学习LR文本和HR文本之间复杂的映射关系,也没有充分利用文本区域先验知识。

近年来,深层CNN成功应用于影像文本超分辨率任务。文献[8]引入自然场景图像抠图算法对提取的文本、前景和背景层在两路分别重建文本区域边缘与颜色信息,并使用文本内容作为监督标签训练网络。文献[1]提出使用文本内容识别结果构建文本感知损失训练网络。文献[2]构建了真实自然场景影像文本数据集 Text-Zoom,并提出了融合序列残差块与中央对齐模块的文本超分辨率重建网络(text super-resolution network, TSRN)验证其有效性。上述方法从不同角度使网络模型专注于处理场景影像的文本区域,增强LR文本细节,提升文本识别准确率。这些网络普遍采用通用图像超分辨率重建网络结构,没有充分利用图像中文本区域与背景差异、文本之间语义序列关系等先验信息。

本文提出了融合混合残差注意力机制与序列单元的文本超分辨率重建网络(hybrid sequential residual attentional text super-resolution network, HSRATN)。为了使模型关注场景影像文本区域以充分学习文本区域先验知识,本文提出了混合残差注意力结构,融合空间注意力机制和拉普拉斯通道注意力机制,利用特征图在空间和通道上各自的依赖关系,学习文本部分的多层次特征表示。此外,为了利用影像中文本字符间的序列先验信息,本文进一步提出了混合序列残差注意力模块,在混合残差注意力结构中融合包含双向门控循环结构的序列单元,提取字符序列关系;采用梯度先验损失函数衡量HR文本图像与超分辨率重建文本图像在梯度场上的差距,并对网络的中间特征图进行可视化。

1 相关工作

1.1 超分辨率重建

图像超分辨率重建是指从退化的LR图像重建生成对应的HR图像[9],广泛应用于文档图像[6]、街景影像[2]、人脸图像[10]、遥感影像[11]重建等。基于学习的超分辨率重建从样本数据中学习先验知识,模型一方面通过加深网络来扩大网络感受野,提升性能[5];另一方面则通过残差连接、密集连接等结构提取并保存各层次特征图的信息,提高模型训练效率[12]。

1.2 注意力机制

注意力机制[13]在自然语言处理领域取得成功应用后,也开始应用于计算机视觉语义分割[14]、图像分类[15]和图像识别[16]等任务。文献[17]首次将注意力机制应用于图像超分辨率重建,利用通道注意力机制学习通道间相互关系来为各通道赋予权重,忽略低频信息而加强高频信息。文献[18]提出混合残差注意力网络(hybrid residual attentional network, HRAN),融合空间注意力和通道注意力机制,学习特征图空间和通道之间的关系。文献[19]将多注意力结构应用于文本超分辨率任务。文献[20]提出密集残差拉普拉斯注意力网络(densely residual Laplacian attentional network, DRLN),在通道注意力中增加拉普拉斯金字塔结构,增强网络特征学习能力。注意力机制可以根据特征的相对重要性对其赋予权重,为了充分学习影像中文本区域特征,本文方法利用注意力机制加强文本高频信息。

2 本文方法原理

2.1 超分辨率网络结构

如图1所示,本文方法网络结构包括预处理

网络、推理网络与重建网络,选用HRAN作为基础模型进行优化。

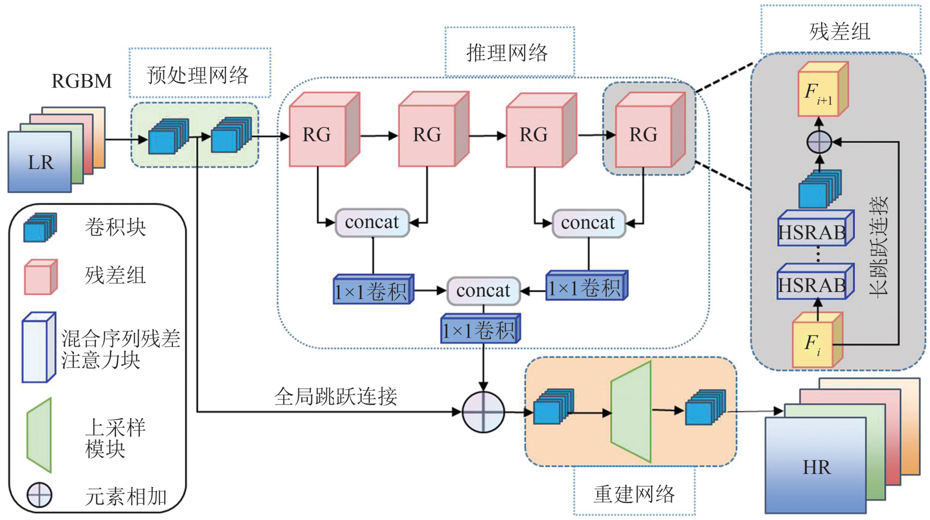

图1 本文方法网络结构

Fig. 1 The Architecture of HSRATN

预处理网络由两个核大小为3×3的卷积层组成,将输入从图像空间映射到特征空间,进行初级特征提取。通过适应性图像阈值算法生成红绿蓝(red green blue,RGB)图像的二值化语义掩膜(mask, M),将两者在通道维度上连结构成RGBM四通道特征图,将其作为网络的输入。适应性图像阈值算法首先将图像转为灰度图像,然后计算图像的平均亮度,将大于平均亮度的区域像素值变成0,而其他区域像素值变成255[19]。如图2所示,二值化语义掩膜可以看作图像文本区域和背景区域的语义分割图,作为语义先验输入网络有助于增强文本重建效果,并提高模型学习效率。

图2 二值化语义掩膜

Fig. 2 Demonstration of Binary Semantic Masks

推理网络由多个残差组(residual group, RG)和特征融合结构组成。残差组包括多个混合序列残差注意力模块(hybrid sequential residual attentional block, HSRAB)和一个3×3卷积层,提取特征图中的空间信息、通道信息和文本序列信息。特征融合结构使用1×1卷积融合相邻RG输出的特征图,融合提取的低、中与高层次特征。网络中全局跳跃连接保存输入图像的初级特征,使其只需学习残差映射关系以恢复丢失的高频细节,不必学习完整图像之间复杂的映射关系。

重建网络将信息从特征空间映射到图像空间,包括两个3×3卷积层和一个用像素洗牌层[21]实现的上采样模块。

2.2 HSRAB

如图3所示,HSRAB由混合残差注意力结构和序列单元(sequential unit, SU)组成。HSRAB中加入短跳跃连接,通过建立恒等映射通道提高输入特征的复用率,解决训练过程中出现的梯度异常和网络性能退化问题[22]。第i个RG中第j个HSRAB可以表示为:

$$F_{i,j+1}=H_{\mathrm{SUv}}(H_{\mathrm{SUh}}(H_{\mathrm{SA}}(F_{i,j})\cdot H_{\mathrm{LCA}}(F_{i,j}))+F_{i,j}) \tag{1}$$

式中,$F_{i,j}$、$F_{i,j+1}$分别为输入、输出特征;H_{SA}、H_{LCA}表示空间注意力(spatial attention, SA)函数和拉普拉斯通道注意力(Laplacian channel attention, LCA)函数;H_{SUh}和H_{SUv}表示处理图像水平行和竖直列的两个序列单元。

2.2.1 混合残差注意力结构

混合残差注意力结构融合了SA和LCA机制。SA通过提取感受野大小不同的特征图获得多层次空间特征,引入的LCA通过学习多频率子带特征,适应性地调整特征依赖关系来为特征通道赋权。融合二者能使模型更好地关注图像的文本区域,学习文本部分的多层次特征表示。

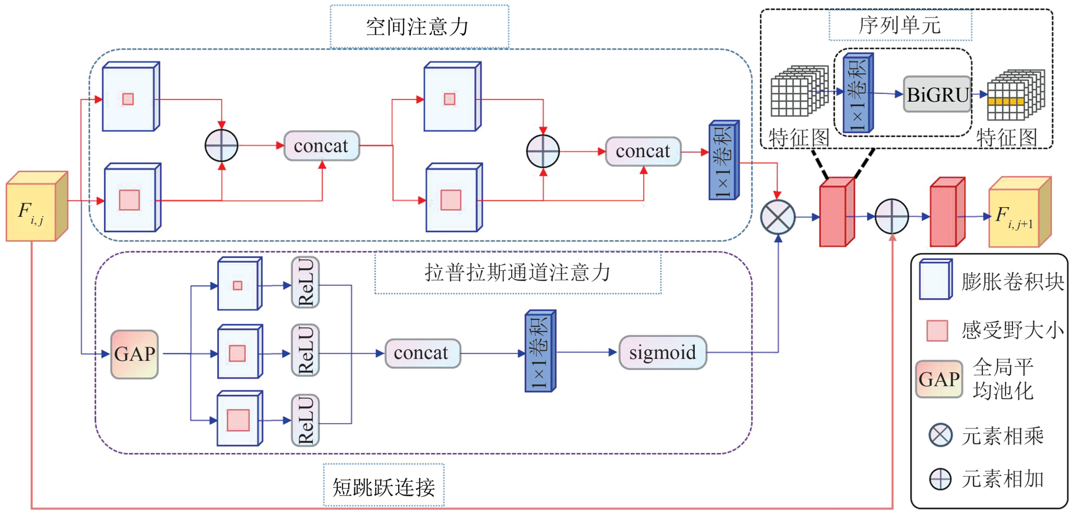

图 3　混合序列残差注意力模块结构

Fig. 3　The Proposed Hybrid Sequential Residual Attentional Block

多尺度超分辨率重建网络(multi-scale super-resolution network, MSRN)[23]利用卷积核大小不同的卷积感受野大小不同的特性,提取多尺度特征图。HRAN 使用膨胀因子大小不同的膨胀卷积获得不同的感受野,同时能减少网络参数数量。同样地,本文方法使用多个不同的膨胀卷积提取文本多尺度特征,将多尺度特征提取看作是 SA。SA 的整个过程可以表示为:

$$F_{S_1}=\tau(H_{D_1}(F_{i,j})) \tag{2}$$

$$F_{S_2}=\tau(H_{D_2}(F_{i,j})+F_{S_1}) \tag{3}$$

$$F_S=[F_{S_1}\ F_{S_2}] \tag{4}$$

$$F_{S_1}=\tau(H_{D_1}(F_S)) \tag{5}$$

$$F_{S_2}=\tau(H_{D_2}(F_S)+F_{S_1}) \tag{6}$$

$$H_{SA}(F_{i,j})=H_{1\times 1}([F_{S_1}\ F_{S_2}])=\tau(W_D[F_{S_1}\ F_{S_2}]) \tag{7}$$

式中,H_{D_1}、H_{D_2}分别表示膨胀因子为 1 和 2 的膨胀卷积;τ表示 Leaky ReLU 激活函数;H_{SA}为多尺度特征提取函数;$H_{1\times 1}$表示 1×1 卷积;W_D为其权重。首先在通道维度上连结两个膨胀卷积层的输出,然后使用 1×1 卷积进行通道压缩,保持输出与输入特征图通道数相同。

特征图各个通道所含信息的重要程度不同,通道注意力机制利用通道间相互关系来为各通道赋予权重。为了充分学习这种关系,和 DRLN 一样,本文在通道注意力结构内引入了拉普拉斯金字塔,组成 LCA。首先,LCA 使用全局平均池化(global average pooling, GAP)操作获得图像的一维统计特征 F_C,数学表达式为:

$$F_C=H_{GAP}(F_{i,j})=\frac{1}{h\times w}\sum_{a=1}^{h}\sum_{b=1}^{w}F_{i,j}(a,b) \tag{8}$$

式中,H_{GAP}为全局平均池化函数;输入特征图 $F_{i,j}$ 的高、宽和通道数分别为 h、w、C;$F_{i,j}(a,b)$表示 $F_{i,j}$ 在(a,b)位置的值。处理之后,特征 F_C 的尺寸为 1×1×C。然后,LCA 使用拉普拉斯金字塔学习多频率子带特征,其中拉普拉斯金字塔由膨胀因子不同的膨胀卷积组成,输出的多层次特征在通道维度上连结。再经过一个 1×1 卷积进行上采样,得到的一维特征经 sigmoid 激活函数处理后获得 LCA 权重。LCA 可以表示为:

$$F_{D_3}=\mathrm{ReLU}(H_{D_3}(F_C)) \tag{9}$$

$$F_{D_5}=\mathrm{ReLU}(H_{D_5}(F_C)) \tag{10}$$

$$F_{D_7}=\mathrm{ReLU}(H_{D_7}(F_C)) \tag{11}$$

$$F_{\mathrm{Laplace}}=[F_{D_3}\ F_{D_5}\ F_{D_7}] \tag{12}$$

$$H_{\mathrm{LCA}}(F_{i,j})=f(W_U F_{\mathrm{Laplace}}) \tag{13}$$

式中,H_{D_3}、H_{D_5}、H_{D_7}表示膨胀因子分别为 3、5、7 的膨胀卷积;ReLU 为激活函数;F_{D_3}、F_{D_5}、F_{D_7}、F_{Laplace}分别表示多层次特征和拉普拉斯金字塔特征;W_U、f表示上采样卷积层权重和 sigmoid 激活函数。最后,将 SA 和 LCA 分支的输出相乘进行融合,加权重要的子波段特征,即提取最有效的文本区域特征。

2.2.2　序列单元

通用图像超分辨率算法一般只考虑重建图像中纹理、边缘等高频细节,而街景影像文本中字符间有确切语义序列关系。即使某一字符模糊而无法直接辨识,人们也可以通过前后字符语义关系进行“完形填空”,从而识别模糊字符。在文本检测识别领域,许多模型使用循环神经网络及其变体提取文本语境信息[24]。在文本超分辨率领域,TSRN 使用双向长短期记忆结构提取文

本序列关系。为了简化模型，提高计算效率，本文提出采用结构更简单的双向门控循环单元（bidirectional gating cycle unit, BiGRU）提取文本字符间的序列关系。如图3所示，提出的序列单元由卷积层和BiGRU组成。首先，使用卷积层提取特征，然后将特征图水平行或者垂直列作为一维序列特征向量，输入BiGRU更新隐藏层的内部状态，学习序列特征之间的语义先验关系，该过程可以表示为：

$$\boldsymbol{S}_{t_\mathrm{h}}=\boldsymbol{H}_{\mathrm{BiGRUh}}\left(X_{t_\mathrm{h}},S_{t_\mathrm{h}-1}\right),\ t_\mathrm{h}=1,2,\cdots,H \tag{14}$$

$$\boldsymbol{S}_{t_\mathrm{v}}=\boldsymbol{H}_{\mathrm{BiGRUv}}\left(X_{t_\mathrm{v}},S_{t_\mathrm{v}-1}\right),\ t_\mathrm{v}=1,2,\cdots,W \tag{15}$$

式中，S_t表示隐藏状态；X_t表示输入特征；H_{BiGRUh}、H_{BiGRUv}分别表示处理水平文本行和竖直文本列的两个BiGRU；t_h、t_v表示沿着输入特征水平和竖直方向构造特征向量；H、W分别为特征图的宽和高。最后，SU将BiGRU输出的一维序列特征转化成特征图。

2.3 梯度先验损失

现有基于逐像素求差的损失函数无法描述HR与超分辨率重建图像在梯度轮廓上的差距，超分辨率重建的文本边缘不够清晰。Sun等[25]提出在超分辨率重建网络中使用梯度轮廓先验重建出更锐利的边缘，而TSRN将其应用在文本超分辨率重建中。Tran等[26]提出梯度差异损失，将其与像素损失相结合，增强重建图像的边缘。上述方法提出的梯度轮廓损失和梯度差异损失利用图像中的梯度先验知识指导网络梯度流动，锐化文本边缘。因此，引入梯度先验损失（gradient prior loss, GPL）重建文本边缘等细节。本文使用的GPL函数的计算公式为：

$$L_{\mathrm{GPL}}=\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla\boldsymbol{I}_{\mathrm{SR}}\left(\left[\boldsymbol{I}_{\mathrm{SR}}\ \boldsymbol{M}_{\mathrm{SR}}\right]^{i}\right)-\nabla\boldsymbol{I}_{\mathrm{HR}}\left(\left[\boldsymbol{I}_{\mathrm{HR}}\ \boldsymbol{M}_{\mathrm{HR}}\right]^{i}\right)\right\|_{1} \tag{16}$$

式中，L_{GPL}为GPL函数；$\boldsymbol{I}_{\mathrm{SR}}$表示超分辨率图像；$\boldsymbol{I}_{\mathrm{HR}}$表示高分辨率图像；$\boldsymbol{M}_{\mathrm{SR}}$表示超分辨率图像的掩膜；$\boldsymbol{M}_{\mathrm{HR}}$表示高分辨率图像的掩膜；$N$为批大小；$\nabla\boldsymbol{I}_{\mathrm{SR}}$和$\nabla\boldsymbol{I}_{\mathrm{HR}}$表示图像梯度场。

本文方法提出的联合损失约束如图4所示，首先通过逐像素计算超分辨率重建与HR图像的差距，在亮度层面上对算法进行约束，该均方差损失L_2公式为：

$$L_2=\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\left[\boldsymbol{I}_{\mathrm{SR}}\ \boldsymbol{M}_{\mathrm{SR}}\right]^{i}-\left[\boldsymbol{I}_{\mathrm{HR}}\ \boldsymbol{M}_{\mathrm{HR}}\right]^{i}\right\|_{2}\right)$$

同时，计算超分辨率重建与HR图像梯度场（像素RGB值的空间梯度）之间的差距，然后利用联合约束损失进行训练。HSRATN的损失函数$L(\boldsymbol{\Theta})$为：

$$L(\boldsymbol{\Theta})=\frac{1}{N}\sum_{i=1}^{N}\left(\left\|\left[\boldsymbol{I}_{\mathrm{SR}}\ \boldsymbol{M}_{\mathrm{SR}}\right]^{i}-\left[\boldsymbol{I}_{\mathrm{HR}}\ \boldsymbol{M}_{\mathrm{HR}}\right]^{i}\right\|_{2}+\lambda_1\left\|\nabla\boldsymbol{I}_{\mathrm{SR}}\left(\left[\boldsymbol{I}_{\mathrm{SR}}\ \boldsymbol{M}_{\mathrm{SR}}\right]^{i}\right)-\nabla\boldsymbol{I}_{\mathrm{HR}}\left(\left[\boldsymbol{I}_{\mathrm{HR}}\ \boldsymbol{M}_{\mathrm{HR}}\right]^{i}\right)\right\|_{1}\right) \tag{17}$$

式中，$\boldsymbol{\Theta}$表示网络参数；λ_1为损失权重。

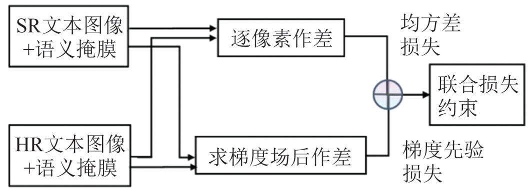

图4 HSRATN的联合损失约束

Fig. 4 Joint Constraint Loss of HSRATN

2.4 实现细节

本文方法的网络配置与HRAN相同，包含4个RG，每个RG由8个HSRAB和一个3×3卷积层组成。模型能处理彩色和灰度图像，最后一层滤波器数目相应为4或2，其余各层滤波器数目均为64。大部分卷积层采用Leaky ReLU激活函数，表达式为$y=\max(0,x)+\theta\min(0,x)$，$\theta$是很小的常数。

3 文本超分辨率实验与分析

3.1 测试数据集

本文使用真实场景文本超分辨率数据集TextZoom进行实验验证。数据集包含放大倍率为2的LR（64×16像素）和HR（128×32像素）文本图像对，以及对应的文本标签。训练集约17 000张图像，测试集包含容易、一般、难3个部分，LR图像质量依次降低。因为拍摄时相机的抖动、偏移以及裁剪操作，TextZoom中文本图像更模糊，LR和HR图像对之间有偏移问题。同TSRN一样，LR图像经空间转换网络（spatial transformer networks, STN）[27]对齐处理后再输入本文网络，以减轻偏移问题对重建结果造成的影响。

3.2 训练设置

选用Adam算法作为损失函数优化器，梯度一阶和二阶矩估计的系数分别为0.9和0.999。学习率为0.000 1，训练周期为200，批量数据尺寸为25，联合约束损失中损失权重λ_1设置为0.000 1。采用PyTorch深度学习框架实现网络，代码地址为 https://github.com/Slupiter/HSRATN，硬件

参数为:Intel Xeon(R) CPU, E5-1620 3.5 GHz, GeForce GTX 1080 GPU。

3.3　消融实验

为了分析 HSRATN 中 GPL、LCA、SU 的作用,依次修改网络配置,比较重建效果的差异,验证其有效性并构建最佳网络。使用文本识别模型注意场景文本识别器(attentional scene text recognizer,ASTER)[24]的识别准确率作为评价指标,实验结果如表1所示。为了真实地反映各算法的文本重建结果在视觉感知效果上的差异,图5展示了部分结果。每个图像下方字符串为识别结果,标红表示识别错误。

表 1　不同配置 HSRATN 模型消融实验结果

Tab. 1　Ablation Study for Different Settings of HSRATN Model

方法	配置	损失函数	ASTER准确率/%			
			容易	一般	难	平均
1	HRAN	L_2	70.5	55.9	38.5	56.0
2	HRAN	L_2+L_{GPL}	71.9	55.9	37.8	56.3
3	HRAN+ LCA	L_2	71.3	56.2	38.9	56.5
4	HRAN+SU	L_2	72.8	57.1	39.6	57.5
5	HRAN+SU	L_2+L_{GPL}	72.9	57.6	39.8	57.8
6*	HRAN+SU +LCA	L_2+L_{GPL}	73.5	58.4	40.8	58.6

注:*表示最佳网络。

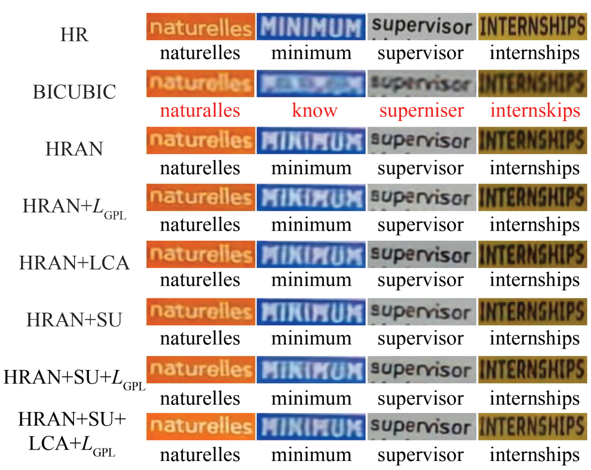

图 5　消融实验的结果

Fig. 5　Results of Ablation Study

本文分别基于像素损失与联合约束损失训练网络,表1中方法1和2的结果表明,加入GPL后,平均识别准确率提高了0.3%。虽然提升较少,但从图5第4行的结果可以看出,重建文本边缘清晰,视觉质量更好。方法1和3的结果表明,在通道注意力中加入拉普拉斯通道注意力金字塔结构后,平均识别准确率提高了0.5%。这说明混合注意力机制可以使模型更好地学习图像文本区域的多层次特征表示,重建清晰的文本细节。从方法1和4的结果可以看出,序列单元使得准确率提高了1.5%。图5第5行的结果显示,重建图像在视觉上更有辨识度,如"supervisor"中的"s、e、o"。这说明在网络中加入BiGRU可以提取文本字符间的序列先验信息,有利于提高模糊字符重建效果。

此外,本文对6种方法的中间特征图进行可视化,绘出彩色热力图,如图6所示,从蓝到红为热度增加方向,所选特征图为推理网络的输出。从图6(b)、6(c)、6(e)、6(f)可以看出,GPL增强了模型对文本边缘的关注度,有利于获得清晰锐利的文本。此外,从图6(b)~6(g)中可以看出,SU和LCA使模型更加关注文本区域,有利于提取文本区域先验知识,重建高质量文本。这进一步验证了本文方法各模块的有效性。

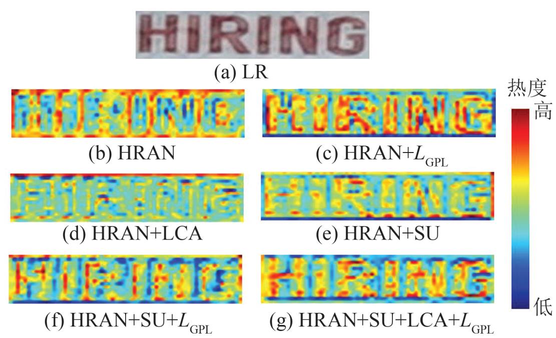

图 6　中间特征热力图

Fig. 6　Intermediate Feature Heatmap

3.4　实验结果分析

将本文方法与双三次插值(Bicubic interpolation, BICUBIC)、SRCNN、深度超分辨率重建(very deep super-resolution, VDSR)、残差超分辨率重建网络(super-resolution residual network, SRResNet)[28]、拉普拉斯超分辨率重建网络(Laplacian super-resolution network, LapSRN)[29]、增强深度残差超分辨率重建(enhanced deep residual super-resolution, EDSR)、残差密集网络(residual dense network, RDN)、MSRN、HRAN 和 DRLN 等通用超分辨率重建方法以及文本超分辨率重建 TSRN 进行对比实验。为了消除实验偏差,使用 TextZoom 和公布的代码对所有模型重新训练。SRResNet 使用的损失函数中 L_{tv} 为对抗损失,L_P 为感知损失,详细介绍见文献[28]。LapSRN 使用的损失函数为 Charbonnier 损失,详细介绍见文献[29]。绝对误差损失函数 L_1 的表

达式为：$L_1=\frac{1}{N}\sum_{i=1}^{N}\| I_{SR}-I_{HR}\|^i$。

表2列出了重建影像的文本识别准确率，本文使用的主流文本识别模型ASTER、多目标修正注意力网络（multi-object rectified attention network，MORAN）[30]和卷积循环神经网络（convolutional recurrent neural network，CRNN）[31]均为原作者公开的代码。可以看出，分辨率放大2倍时，HSRATN重建结果识别准确率优于其他模型。与基线算法HRAN相比，HSRATN重建结果的ASTER、MORAN和CRNN平均识别准确率分别提高了2.6%、3.2%和4.7%。与该领域领先的通用超分辨率重建算法DRLN相比，HSRATN结果的识别准确率分别提高了2.0%、2.4%和3.3%。与文本超分辨率重建算法TSRN相比，HSRATN的ASTER和CRNN平均识别准确率在×2尺度下分别获得0.3%和4.1%的提升，达到了先进水平。

表2 TextZoom真实数据集超分辨率模型重建效果/%

Tab. 2 Performance of Super-Resolution Models on Three Subsets in TextZoom/%

模型	损失函数	ASTER准确率				MORAN准确率				CRNN准确率			
		容易	一般	困难	平均	容易	一般	困难	平均	容易	一般	困难	平均
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	L_2	69.4	43.3	32.2	49.5	63.2	39.0	30.2	45.3	38.7	21.6	20.9	27.7
VDSR	L_2	71.7	43.5	34.0	51.0	62.3	42.5	30.5	46.1	41.2	25.6	23.3	30.7
SRResNet	$L_2+L_W+L_P$	69.6	47.6	34.3	51.3	60.7	42.9	32.6	46.3	39.7	27.6	22.7	30.6
EDSR	L_1	72.3	48.6	34.3	53.0	63.6	45.4	32.2	48.1	42.7	29.3	24.1	32.7
RDN	L_1	70.0	47.0	34.0	51.5	61.7	42.0	31.6	46.1	41.6	24.4	23.5	30.5
LapSRN	Charbonnier	71.5	48.6	35.2	53.0	64.6	44.9	32.2	48.3	46.1	27.9	23.6	33.3
MSRN	L_2	70.2	54.6	37.0	55.0	64.2	47.9	35.1	50.0	49.8	34.9	29.9	38.9
HRAN	L_2	70.5	55.9	38.5	56.0	64.7	48.8	36.2	50.8	52.8	36.8	30.4	40.8
DRLN	L_2	72.3	55.1	39.1	56.6	66.6	49.1	36.3	51.6	53.4	40.6	30.5	42.2
TSRN	L_2+L_{GPL}	75.1	56.3	40.1	58.3	70.1	53.3	37.9	54.8	52.5	38.2	31.4	41.4
本文方法	L_2+L_{GPL}	73.5	58.4	40.8	58.6	67.2	53.4	38.8	54.0	56.2	44.4	33.7	45.5

本文方法的主要目的是提升在测试集上的文本识别准确率。表3也展示了各算法的图像重建质量评价指标峰值信噪比（peak signal-to-noise ratio, PSNR）和结构相似性指数（structural similarity index, SSIM）结果。可以看出，本文方法的PSNR在容易和一般测试集上数值较差，有两方面原因：(1)PSNR的定义和像素损失函数高度相关，最小化像素损失等同于直接最大化PSNR值。而本文采用联合约束损失训练网络，没有为了追求较高的PSNR值采用单一的像素损失。(2)在处理真实数据集中LR/HR图像不对齐问题时，使用的STN可能造成图像像素偏移。SSIM在容易测试集上数值较差，推测本文方法更适用于较低质量图像重建。通常来说，PSNR和SSIM不能准确地反映图像视觉感知质量。因此，在本文中，文本识别准确率评价指标更为重要。

图7展示了BICUBIC、SRResNet、MSRN、HRAN、DRLN、TSRN和本文方法的部分重建图像以及ASTER文本识别结果。可以看出，由于文本特征的结构化程度较高，对某些部位的形变起到限制作用，现有通用超分辨率重建算法得到的结果过于平滑，文本边缘和纹理较为模糊，无法重建不同文本图像的细节特征，有些得到了错误的文本识别结果。HSRATN则受益于混合残差注意力结构提取的文本区域多层次特征表示、序列残差单元提取的文本序列信息和梯度先验损失，获得了丰富的文本先验知识，从而能够重建出更清晰的边缘和纹理细节，如“supervisior”中的“s”、“education”中的“a”和“national”中的“n”。此外，本文方法对倾斜文本也有效，如“access”和“musicaalta”。

图8展示了本文方法和部分主流模型在性能和参数数量方面的实验结果，均在2倍真实TextZoom数据集上进行。可以看出，HSRATN比RDN、DRLN、EDSR参数数量更少，重建效果却更好。相比于EDSR模型，参数数量少了约72%，ASTER平均识别准确率提高了5.6%。相比于基线模型HRAN，参数数量多了19%，准确率却提高了2.3%。这表明本文方法在性能和参

表 3　TextZoom 真实数据集超分辨率重建模型 PSNR 和 SSIM 值

Tab. 3　PSNR and SSIM of Super-Resolution Models on TextZoom

方法	PSNR			SSIM		
	容易	一般	难	容易	一般	难
BICUBIC	22.35	18.98	19.39	0.788 4	0.625 4	0.659 2
SRCNN	23.48	19.06	19.34	0.837 9	0.632 3	0.679 1
VDSR	24.62	18.96	19.79	0.863 1	0.616 6	0.698 9
SRResNet	24.36	18.88	19.29	0.868 1	0.640 6	0.691 1
EDSR	24.26	18.63	19.14	0.863 3	0.644 0	0.710 8
RDN	22.27	18.95	19.70	0.824 9	0.642 7	0.711 3
LapSRN	24.58	18.85	19.77	0.855 6	0.648 0	0.708 7
MSRN	23.20	19.01	20.12	0.853 8	0.658 9	0.730 1
HRAN	23.08	18.87	19.86	0.862 1	0.665 2	0.735 5
DRLN	22.91	18.98	20.00	0.862 2	0.667 6	0.742 5
TSRN	25.07	18.86	19.71	0.889 7	0.667 6	0.730 2
本文方法	23.00	18.94	20.21	0.871 0	0.675 1	0.754 2

数数量之间有较好的权衡。

3.5　合成 TextZoom 数据集实验

为了进一步验证本文方法的泛化性能,使用双三次插值下采样合成的 TextZoom 数据对部分算法进行训练和测试。表 4 展示了 HSRATN 与 BICUBIC、通用超分辨率重建算法 SRResNet、MSRN、HRAN 以及文本超分辨率重建算法 TSRN 的 ASTER 文本识别准确率。可以看出,分辨率放大 4 倍时,HSRATN 重建结果识别准确率优于其他模型。相比于基线模型 HRAN,平均识别准确率提高了 0.9%。这表明本文方法在其他超分辨率及退化模型上有较强的泛化能力。

3.6　ICDAR 2015 TextSR 数据集实验

受文献[19]启发,使用 ICDAR 2015 TextSR 数据集[4]进行实验,验证本文方法重建效果,该数据集源自法语视频字幕。此处采用数据集规定的 Tesseract-OCR v3.02 软件测试文本识别准确

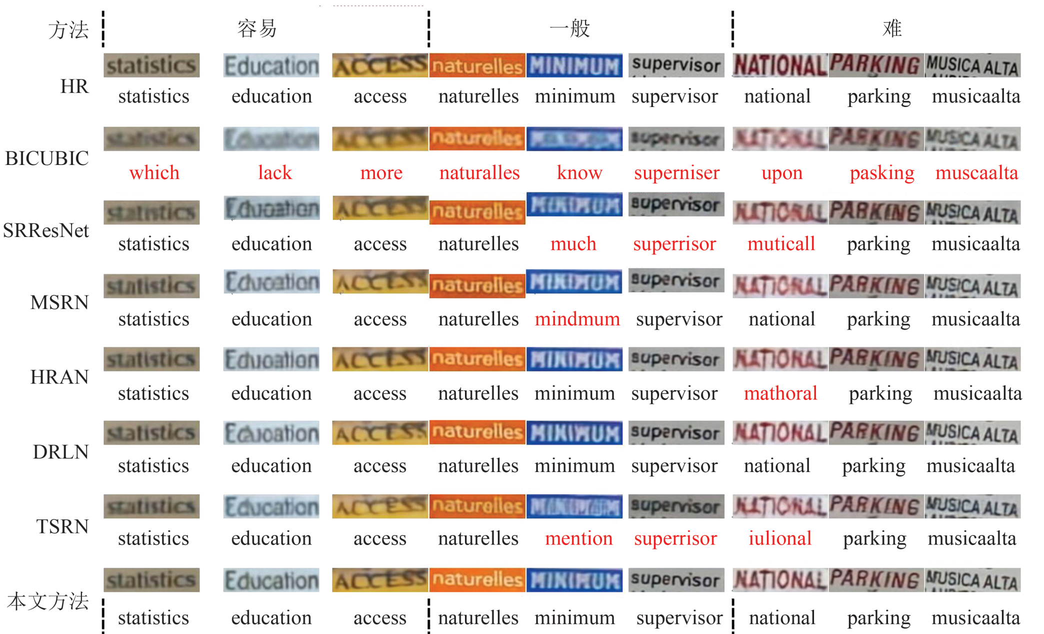

图 7　主流超分辨率模型结果

Fig. 7　Results of State-of-the-Art Super-Resolution Model

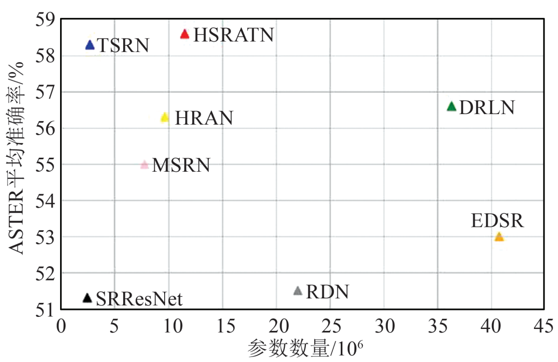

图 8　性能和参数数量比较

Fig. 8　Comparison for Performance and Number of Parameters

表 4　超分辨率重建模型在合成 TextZoom 数据集上的结果

Tab. 4　Results of Super-Resolution Models on Synthetic TextZoom

方法	放大倍率	ASTER 准确率/%			
		测试 1	测试 2	测试 3	平均
BICUBIC	4	21.7	31.5	45.9	32.3
SRResNet	4	43.1	54.9	58.4	51.6
MSRN	4	45.6	57.8	61.1	54.3
HRAN	4	51.3	59.8	62.3	57.4
TSRN	4	51.8	60.9	60.4	57.4
本文方法	4	52.8	60.7	62.4	58.3

率 O_{ac},可以表示为:

$$O_{ac}=1-\frac{1}{K}\sum_{i=1}^{N}d_i \tag{18}$$

式中,N 为测试集图像数量;K 为图像中字符总数;d_i 表示图像文本标签和识别结果字符串的编辑距离。

表5展示了部分通用超分辨率重建方法、文本超分辨率方法[19, 32]和本文方法2倍放大实验的客观评价指标结果。可以看到,本文方法在该数据集上文本识别率高于其他方法,O_{ac} 为79.04%,比HR原图识别准确率78.80%高0.24%,比HRAN和文献[33]分别提高0.41%和0.21%。图9展示了本文方法超分辨率重建的部分结果,可以看出重建图像文本边缘清晰、纹理细节丰富,接近HR图像。这表明本文方法在其他场景文本图像数据上有较强的泛化能力。

表5 ICDAR 2015 TextSR数据集上的结果

Tab. 5 Results on ICDAR 2015 TextSR Dataset

方法	放大倍率	PSNR	SSIM	O_{ac}/%
SRCNN	2	31.75	0.980	77.19
SRResNet	2	29.04	0.950	76.55
LapSRN	2	31.40	0.972	76.48
EDSR	2	32.09	0.976	77.53
RDN	2	32.61	0.978	77.67
HRAN	2	32.92	0.987	78.63
文献[19]	2	32.86	0.979	78.80
文献[32]	2	33.21	0.978	78.78
文献[33]	2	33.94	0.982	78.83
本文方法	2	32.97	0.987	79.04

图9 本文方法在ICDAR 2015 TextSR数据集上的结果

Fig. 9 Results of HSRATN on ICDAR 2015 TextSR Dataset

4 结 语

针对街景影像分辨率低导致文本识别准确率降低的问题,本文提出了一种融合混合残差注意力机制与序列单元的影像文本超分辨率重建算法。实验结果表明,本文方法可以更好地利用街景影像中文本区域的先验知识,超分辨率重建影像文本边缘清晰、纹理细节丰富,提高了文本识别准确率。

参 考 文 献

[1] Wang W J, Xie E Z, Sun P Z, et al. TextSR: Content-Aware Text Super-Resolution Guided by Recognition [EB/OL]. (2019-09-16) [2021-07-01]. http://arxiv.org/abs/1909.07113.

[2] Wang W J, Xie E Z, Liu X B, et al. Scene Text Image Super-Resolution in the Wild[C]//European Conference on Computer Vision, Glasgow, UK, 2020.

[3] Dong C, Zhu X M, Deng Y B, et al. Boosting Optical Character Recognition: A Super-Resolution Approach [EB/OL]. (2015-06-07) [2021-06-01]. http://arxiv.org/abs/1506.02211.

[4] Peyrard C, Baccouche M, Mamalet F, et al. ICDAR2015 Competition on Text Image Super-resolution[C]//The 13th International Conference on Document Analysis and Recognition, Tunis, The Republic of Tunisia, 2015.

[5] Dong C, Loy C C, He K M, et al. Image Super-Resolution Using Deep Convolutional Networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(2): 295-307.

[6] Pandey R K, Vignesh K, Ramakrishnan A G, et al. Binary Document Image Super Resolution for Improved Readability and OCR Performance [EB/OL]. (2018-12-06) [2021-10-08]. http://arxiv.org/abs/1812.02475.

[7] Nakao R, Iwana B K, Uchida S. Selective Super-Resolution for Scene Text Images[C]//International Conference on Document Analysis and Recognition, Sydney, Australia, 2019.

[8] Lin K, Liu Y B, Li T H, et al. Text Image Super-Resolution by Image Matting and Text Label Supervision [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019.

[9] Wang Z H, Chen J, Hoi S C H. Deep Learning for Image Super-Resolution: A Survey [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(10): 3365-3387.

[10] Liao Haibin, Chen Youbin, Chen Qinghu. Non-local Similarity Dictionary Learning Based Super-Resolution for Improved Face Recognition[J]. *Geomatics and Information Science of Wuhan University*, 2016, 41(10): 1414-1420. (廖海斌, 陈友斌, 陈庆虎. 基于非局部相似字典学习的人脸超分辨率与识别[J]. 武汉大学学报(信息科学版), 2016, 41(10): 1414-1420.)

[11] Chen Hang, Luo Bin. Multi-angle Remote Sensing Images Super-Resolution Reconstruction Using Dynamic Upsampling Filter Deep Network[J]. *Geomatics and Information Science of Wuhan University*, 2021, 46(11): 1716-1726.(陈行, 罗斌. 利用动态上采样滤波深度网络进行多角度遥感影像超分辨率重建[J]. 武汉大学学报(信息科学版), 2021, 46(11): 1716-1726.)

[12] Lim B, Son S, Kim H, et al. Enhanced Deep Residual Networks for Single Image Super-Resolution[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[C]//The 31st International Conference on Neural Information Processing Systems, NY, USA, 2017.

[14] Fu J, Liu J, Tian H J, et al. Dual Attention Network for Scene Segmentation [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019.

[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[C]//International Conference on Learning Representations, Vienna, Austria, 2021.

[16] Zhao H S, Jia J Y, Koltun V. Exploring Self-attention for Image Recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020.

[17] Zhang Y L, Li K P, Li K, et al. Image Super-Resolution Using very Deep Residual Channel Attention Networks[C]//The 15th European Conference on Computer Vision, Munich, Germany, 2018.

[18] Muqeet A, Iqbal M T B, Bae S H. HRAN: Hybrid Residual Attention Network for Single Image Super-Resolution [J]. *IEEE Access*, 2019, 7: 137020-137029.

[19] Wang Y Y, Su F, Qian Y. Text-Attentional Conditional Generative Adversarial Network for Super-Resolution of Text Images[C]//IEEE International Conference on Multimedia and Expo, Shanghai, China, 2019.

[20] Anwar S, Barnes N. Densely Residual Laplacian Super-Resolution[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(3): 1192-1204.

[21] Shi W Z, Caballero J, Huszár F, et al. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-pixel Convolutional Neural Network [C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[23] Li J C, Fang F M, Mei K F, et al. Multi-scale Residual Network for Image Super-Resolution[C]// European Conference on Computer Vision, Munich, Germany, 2018.

[24] 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.

[25] Sun J, Sun J, Xu Z B, et al. Gradient Profile Prior and Its Applications in Image Super-Resolution and Enhancement[J]. *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, 2011, 20(6): 1529-1542.

[26] Tran H T M, Ho-Phuoc T. Deep Laplacian Pyramid Network for Text Images Super-Resolution [C]//IEEE-RIVF International Conference on Computing and Communication Technologies, Danang, Vietnam, 2019.

[27] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial Transformer Networks[C]//The 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 2015.

[28] Ledig C, Theis L, Huszar F, et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017.

[29] Lai W S, Huang J B, Ahuja N, et al. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017.

[30] Luo C J, Jin L W, Sun Z H. MORAN: A Multi-object Rectified Attention Network for Scene Text Recognition[J]. *Pattern Recognition*, 2019, 90(C): 109-118.

[31] Shi B G, Bai X, Yao C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(11): 2298-2304.

[32] Xue M L, Huang Z H, Liu R Z, et al. A Novel Attention Enhanced Residual-in-Residual Dense Network for Text Image Super-Resolution[C]//IEEE International Conference on Multimedia and Expo, Shenzhen, China, 2021.