计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(8):278−285 [doi: 10.15888/j.cnki.csa.009111]

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

工业场景下高斯引导的非显著性字符抹除①

姚　超, 庞雄文

(华南师范大学 计算机学院, 广州 510631)
通信作者: 庞雄文, E-mail: augepang@163.com

摘　要: 图像文本信息在日常生活中无处不在, 在传递信息的同时, 也带来了信息泄露的问题. 近年来文本擦除模型很好地解决了这个问题. 然而, 在工业场景下, 图像会出现高光, 对比度较大的非字符区域, 模型往往很容易其影响发生注意力偏移的现象, 从而忽略了字符区域导致不理想的文本抹除效果. 为了克服这一局限性, 基于注意力提出了一种新的文本擦除网络, 即在网络中嵌入了一层额外的特征层用以给生成图中存在字符的区域进行评分. 同时, 引入了高斯热力图并将其作为基础设计损失函数, 采用监督的方式纠正模型的注意力, 将模型注意力引导至正确的字符区域. 通过在 4 种不同的数据集上进行对比, 本文所提方法总体上拥有更好的抹除效果. 同时, 该方法在图像存在复杂的背景情况下, 其在图像抹除任务中仍然具有较高的灵活性.

关键词: 字符抹除; 注意力漂移; 高斯引导; 区域评分

引用格式: 姚超,庞雄文.工业场景下高斯引导的非显著性字符抹除.计算机系统应用,2023,32(8):278–285. http://www.c-s-a.org.cn/1003-3254/9111.html

Gaussian-guided Non-saliency Character Erasure under Industrial Scenarios

YAO Chao, PANG Xiong-Wen

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Image text messages are ubiquitous in everyday life, and while conveying information, they also bring the problem of information leakage. In recent years, text erasure models have solved this problem very well. However, in industrial scenarios where images are highlighted and non-character areas with high contrast, the models are often susceptible to their influence of attentional drift, thus neglecting the character areas and resulting in unsatisfactory text erasure. In order to overcome this limitation, this study proposes a new text erasure network based on attention. Specifically, an additional feature layer is embedded in the network to score the areas where characters are present in the generated image. At the same time, the study introduces a Gaussian heat map and uses it as the basis for designing a loss function that corrects the model's attention and guides it to accurate character areas in a supervised manner. Through comparison on four different datasets, the proposed method has better erasure results overall. In addition, the method has the same high flexibility for the text erasure task in the presence of complex backgrounds in images.

Key words: character erasure; attentional drift; Gaussian-guided; region score

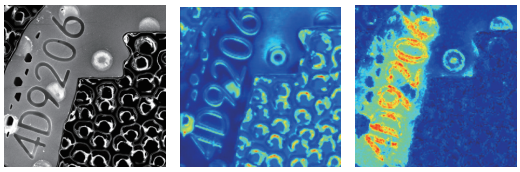
图像文本抹除技术是指在图像中通过用合理的语义内容替换场景中存在的文本并保留纹理细节, 起到擦除文本的作用, 其广泛应用在隐私保护[1]、图像编辑[2]、图像恢复[3] 等领域.

近年来, 基于生成对抗网络 (GAN)[4] 的方法广泛地应用在图像转换的任务中. 一些工作[5] 利用其对图像特殊的翻译能力来解决字符抹除的问题, 模型通过端到端的方式进行训练, 在学习的过程中总结归纳图

① 收稿时间: 2022-11-12; 修改时间: 2022-12-23; 采用时间: 2023-01-06; csa 在线出版时间: 2023-06-09
CNKI 网络首发时间: 2023-06-13

像信息与语义信息将语义上合理的内容替换掉文字部分以实现图像文字的抹除. 这些工作的一个主流策略是明确地从给定的输入图像和目标图像中分离出内容和表征, 采用循环一致性的方式[6] 在二者之间建立高维映射关系从而使得图像之间的转换更为平滑并且效果更好. 此后, 为了更好地学习到图像中关键的特征与语义信息, 在编码和解码的过程中会加入注意力机制[7,8], 结合高维度映射而创建的可学习参数权重, 使得模型在学习过程中会自发地关注一些重要区域, 在风格迁移或者图像生成的任务中都有着较好的效果.

然而, 我们通过实验发现, 这些方法着重解决了如何抹除图像中的文本, 而忽略了背景对字符的影响. 与自然场景不同, 在工业场景中, 工件表面通常包含高光和高对比度区域, 这会误导模型做出错误的字符定位. 因此, 当使用上述算法进行工业文本抹除时, 字符区域很容易被忽略, 尤其是在通过弱监督方式获得文本定位的方法上. 模型缺乏针对性的引导, 在学习过程中往往容易受到显著性非字符区域的影响, 使其发生自注意力漂移的情况 (attention drift), 如图 1 所示.



(a) 原图 (b) 未经引导的注意力热力图 (c) 经过引导的注意力热力图

图 1 注意力热力表示图

在本文中, 我们提出了一种工业场景下的文本抹除方法, 其采用了端到端的形式实现图像文本的抹除. 该算法通过高斯编码图来引导模型更多地关注字符区域. 具体来说, 在模型上采样的过程中我们给生成图额外增添了一层图像特征层, 即为关注域层, 来对字符区域进行评分, 所得到的区域分数表示为给定像素是字符中心的概率. 同时, 为了使模型学习到字符区域的特征, 避免受到背景图像显著性特征的影响, 我们利用标准 2D 高斯分布图对图像字符区域进行编码得到高斯热力图[9–11], 为模型提供图像文本定位的强引导信息. 与离散识别每个像素的二值分布图相比, 高斯热力图在处理没有严格约束的字符区域方面具有很高的灵活性. 在此基础上, 我们将关注域和高斯热力图相结合, 设计了一个新的损失函数, 来纠正模型注意力的区域, 避免使模型产生注意力漂移.

此外, 为了保留更多的背景信息, 实现更好的擦除效果, 我们将关注域、生成图和原图进行加权融合. 最终使得模型不仅可以更加关注生成的擦除图像的文本区域, 还可以保留目标图像的背景区域.

本文的主要贡献如下.

(1) 我们设计了一种新的字符抹除模型, 该模型通过得分域为其提供精确的字符指导, 以解决由于背景的显著性区域引起的注意力漂移问题, 并最终拥有更好的图像擦除效果.

(2) 我们引入了高斯热力图, 通过对字符区域进行高斯编码, 并将其作为损失函数的基础来指导模型在抹除字符的过程中更加关注字符区域.

(3) 我们提出了一个端到端的文本抹除方法, 该方法在工业场景的 4 个数据集上都证明了方法的有效性.

1 相关工作

现有的图像文本抹除方法主要分为两类: 传统的非学习方法和基于深度学习的机器学习方法. 传统方法, 例如 Khodadadi 等人[12] 提出的方法通常使用颜色直方图或者阈值去提取图像的文本区域, 然后利用匹配修复算法, 高效地重构文本区域的图像信息. 此外, Wagh 等人[13] 通过对字符特征的信息找到文字区域, 然后使用最邻近匹配算法, 对文字去除后的区域进行填充, 但进行文字区域填充需要反复迭代, 因此算法的效率不高. 同时, 传统方法比较适用于简单的场景, 在背景较为单一的情况下, 性能良好, 但是在复杂的背景环境下则表现不佳, 无法满足常规性的任务.

随着深度神经网络的发展, 采用神经网络的方法在抹除文字的任务上有了重大的突破. Nakamura 等人[2] 最先提出了一种基于 CNN 的滑动窗口方法进行场景文本擦除. 但是采用滑动窗口的卷积神经网络, 无法使得模型学习到图像整体的语义信息, 从而不可避免地破坏了擦除结果的一致性和连续性. 随后 Zhang 等人设计了一个端到端的网络模型 EnsNet[14], 将生成对抗网络和 U-Net 结构相结合, 整体地进行擦除训练. 这种方法解决了模型学习图像整体语义的问题, 但是会经常出现文本定位不准确和文本抹除不彻底的问题. 为了克服这个问题, Tursun 等人提出了一种二阶段的抹除方法 MTRNet[15], 通过引入辅助掩码提供关于文本所在位置的信息, 从而使得模型能够更好地关注文本抹除. 之后, Tursun 等人对 MTRNet 进行了优化提出了

一个扩展版本 MTRNet++[16], 它引入了一个掩模细化分支, 将粗糙的区域掩模转化为像素级掩模. 用作一个精细修饰分支的输入, 以提供额外的文本信息. 然而, 受限于文字检测模块, 如果不能分割出好的掩模, 文字的抹除就无法达到理想的效果修复, 同时, 模型更加臃肿, 效率更低. 此后, Tang 等人[17] 采用笔画遮罩和背景修复相结合的方式, 从裁剪的文本图像中提取文本笔画作为相对较小的孔, 以保持更多的背景内容, 从而获得更好的修复效果. Bian 等人[18], 将文本去除问题解耦为文本笔划检测和笔划去除并设计了单独的网络来解决这两个子问题,

此外, 我们通过实验发现, 图像翻译模型 Attention-GAN[19], U-GAT-IT[20] 同样可以进行图像字符的抹除, 通过引入注意力机制对于图像信息进行高维编码, 使得模型在学习过程中自主地关注字符区域实现图像的字符抹除. 整体上来说, 基于半监督学习方式的注意力有助于解决掩模对于模型学习产生的弊端. 但是, 如果图片中出现非文字的显著性区域时, 模型在学习的过程中会赋予该区域更高的权重, 从而忽略了字符区域, 最终无法达到满意文字擦除的效果.

2 方法

字符擦除模型由两个生成器$G_{s\to t}$和$G_{t\to s}$以及两个鉴别器D_s和D_t组成. 其中G又由编码器ε_s, 解码器σ_s以及辅助分类器η_s组成. 整体架构如图 2 所示, X_s和X_t分别代表原始图片和抹除文字后的图片集合. 将$x\in\{X_s,X_t\}$作为一组用于训练的源域样本和目标域样本传入到模型中, 传入的图像首先通过ε_s来获得编码特征图f_c. 接着, 基于注意机制的辅助分类器会从编码后的特征图中提取高语义信息, 得到的高语义信息特征图传入到σ_s进行解码并生成的预测图y. 其中y包含两部分, 生成图像 (generating images, GI), 区域得分 (region score, RS). RS与GI结合可以有效地获得字符擦除的区域, 剩余的区域由原始输入图像填充.

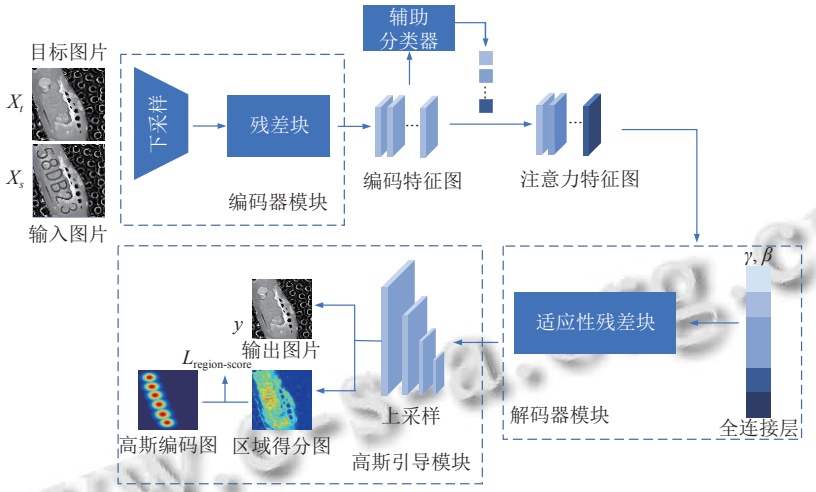


图 2 字符擦除模型的整体架构

2.1 辅助分类器

图 3 为辅助分类器模块的流程示意图, 该模块受到 CAM[21,22] 的启发, 其核心在于通过全局最大池化 (global max pooling) 和全局平均池化 (global average pooling)[23] 来学习特征层的权重, 以无监督学习方式产生注意力.

通过使用一个全连接层将维数降低到一维, 从而可以确定特征层是源域部分还是目标域的一部分. 与此同时, 通过映射到更高维数而产生的可学习参数权重ω_m以及ω_a与编码特征图f_c点积相乘, 给每个通道分配一个权重, 确定该层通道相应特征的重要性, 从而生成注意力特征图.

2.2 解码器

解码过程如图 4 所示, 为了解码注意力特征图中的注意力信息, 我们使用参数γ和β加权 AdaLin (黄色虚线框)[24] 归一化的结果. AdaLIN 将 AdaIN[25] 和 LN[26] 结合选择性地保留或更改内容信息, 并在更改图像特征的同时维护原始域的内容结构.

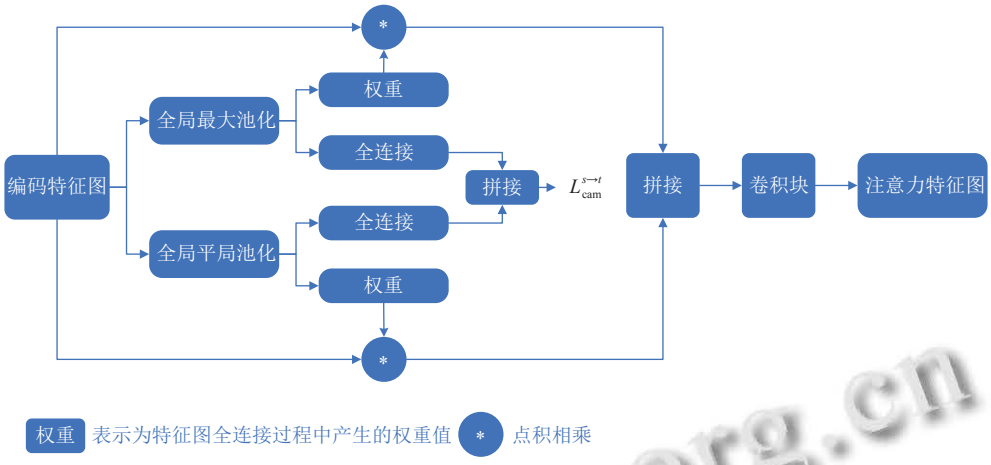


图 3 辅助分类器的流程示意图

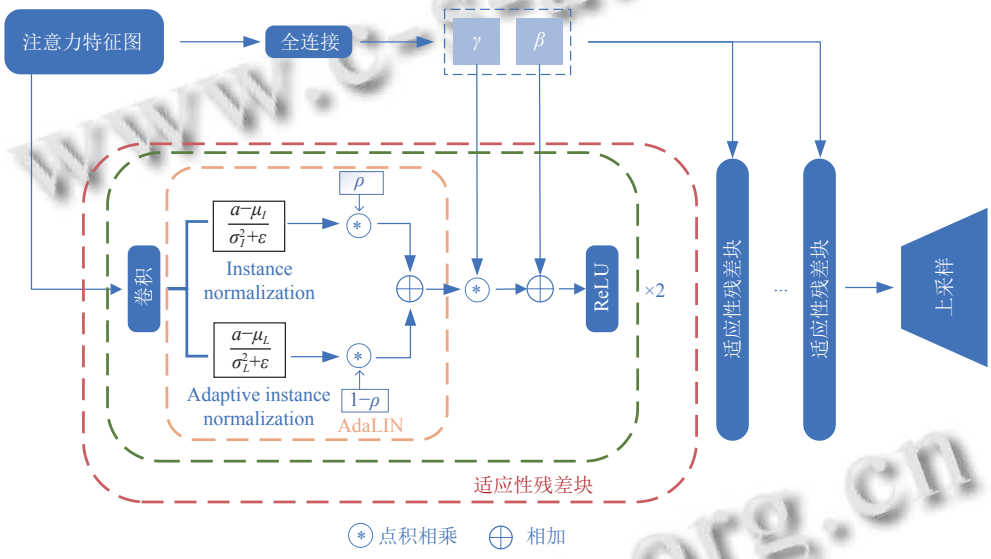


图 4 解码器的流程示意图

图 4 中, μ_I、μ_L和σ_I、σ_L分别表示通道和特征层的均值和标准值, 此外, γ 和 β 是由注意力图中的完全连接的层动态计算得到, ρ的值在解码器的残差块中初始化为 1, 在解码器的上采样块中初始化为 0.

2.3 高斯引导

在高斯导引模块中, 我们设置了一层额外的特征层, 即关注域层 (region score layer, RSL), 以给出字符区域的得分, 该得分表示给定像素是字符中心的概率. 同时我们引入高斯热力图对字符区域进行编码. 通过计算编码后的高斯热力图与 RSL 之间的最小均方差来纠正模型产生的偏差. 这种指向性的纠正措施能够使模型能够不断地聚焦于字符的特征区域从而解决注意漂移的问题. 基于高斯热力图的编码形式在与没有明确界定边界的目标区域合作时, 它提供了高度的灵活性, 已经广泛应用于各个领域, 如动作识别工作[27,28], 具体流程如图 5 所示. 由于图像上的字符框通常重叠在一起, 造成图形的变形, 因此我们将二维各向同性标准高斯映射曲折到字符框选择区域.

区域得分在训练过程进行更新, 如图 6 所示, 在早期阶段, 模型对图像中的文本区域中并不熟悉, 因此区域得分相对较低. 随着训练的进行, 模型能够更精确地处理字符区域, 并且预测的得分值也逐渐提高.

2.4 Mask 加权融合

基于注意力机制与关注域的非字符区域内容仍然会发生变化. 为了解决这个问题, 生成器生成了 RSL 和生成图 *GI*. RSL 可以作为每个像素的权重, 定义了像素的重要性, 以确保在生成器生成的最终结果中只有特定区域的内容发生变化, 而不影响其他区域的内容.

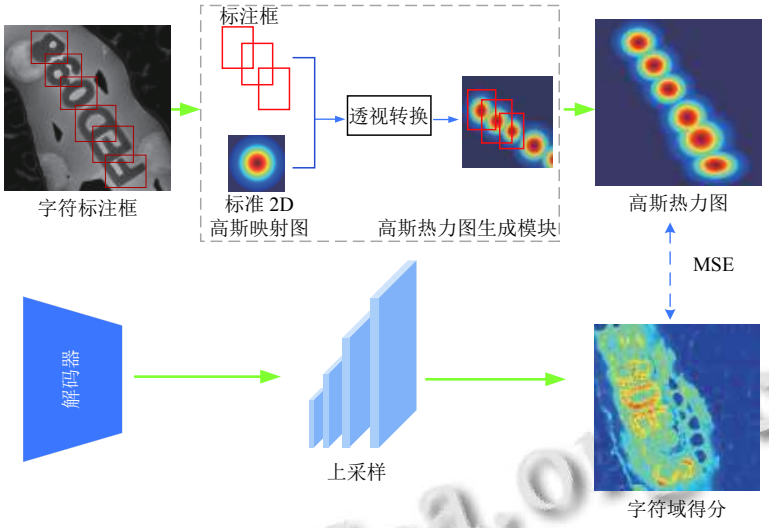


图 5 高斯引导的流程图

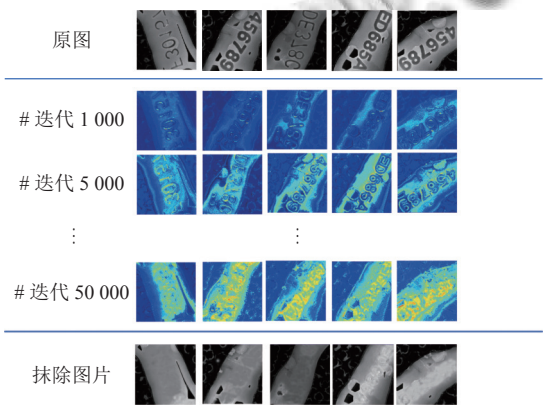


图 6 训练阶段注意力变化示意图

关注域层 RSL 以及内容 C_t 组成. 最终生成的图像通过以下融合公式得到:

$$A_t, C_t = G_{s\to t}(x) \tag{1}$$

$$F(G_{s\to t}(x)) = A_t \times C_t + (1 - A_t) \times x \tag{2}$$

这样可以保证在字符区域的内容改变擦除字符时, 字符以外的区域不会发生变化.

2.5 损失函数

对抗损失: 对抗性损失用于匹配输入图像的分布与目标图像的分布.

$$L_{\text{lsgan}}^{s\to t} = -E_{x\sim X_t}[(D_t(x))^2] + E_{x\sim X_s}[(1 - D_t(G_{s\to t}(x)))^2] \tag{3}$$

其中, x、X_t、X_s 分别表示输入图片、原图片集合及目标图片集合, $G_{s\to t}$、D_t 各自表示为生成器、判别器.

循环损失: 为了学习过程中缓解模式崩溃的问题. 即给定图像 $x \in X_s$, 在 x 从 X_s 到 X_t 以及从 X_t 到 X_s 的顺序转换之后, 图像应该成功地转换回原始域.

$$L_{\text{cycle}}^{s\to t} = E_{x\sim X_s}[|x - G_{t\to s}(G_{s\to t}(x))|_1] \tag{4}$$

其中, x、$G_{t\to s}$、$G_{s\to t}$ 分别表示为输入图片、目标图片至原图的生成器及原图至目标图片的生成器

一致性损失: 为了确保输入图像和输出图像的颜色分布相似, 我们对生成器应用了一致性约束. 给定图像 $x \in X_s$, 在使用 $G_{s\to t}$ 对 x 进行平移之后, 图像应该不会改变.

$$L_{\text{identity}}^{s\to t} = E_{x\sim X_t}[|x - G_{s\to t}(x)|_1] \tag{5}$$

其中, x、$G_{s\to t}$、D_t 分别表示为生成器、判别器.

CAM 损失: 通过利用来自辅助分类器 η_s 的信息, $G_{s\to t}$ 和 D_t 了解在当前状态下需要改进的部分.

$$L_{\text{cam}}^{s\to t} = -(E_{x\sim X_s}[\log(\eta_s(x))] + E_{x\sim X_t}[\log(1 - \eta_s(x))]) \tag{6}$$

$$L_{\text{cam}}^{D_t} = E_{x\sim X_t}[(\eta D_t(x))^2] + E_{x\sim X_s}[(1 - \eta D_t(G_{s\to t}(x)))^2] \tag{7}$$

其中, η_s、η、$G_{s\to t}$、D_t 分别表示为辅助分类器、超参数、生成器、判别器.

关注域损失: 为了在复杂的背景环境下, 采用监督的方式引导模型更加关注字符区域:

$$L_{\text{region-score}} = E_{x\sim X_s}\left(\sum \left\| G_{\text{map}} - G_{s\to t}^{\text{region-score}}(x) \right\|_2^2\right) \tag{8}$$

其中, x、G_{map}、$G_{s\to t}^{\text{region-score}}$ 分别表示为输入图片、高斯编码图及关注域得分,

总损失: 最后, 我们联合训练编码器, 解码器, 鉴别器和辅助分类器, 以优化最终目标:

$$L = \lambda_1 L_{\text{lsgan}}^{s\to t} + \lambda_2 L_{\text{cycle}}^{s\to t} + \lambda_3 L_{\text{identity}}^{s\to t} + \lambda_4 L_{\text{cam}} + \lambda_5 L_{\text{region-score}} \tag{9}$$

3 实验结果与分析

3.1 数据集和评估指标

为了真实地评估本文所提出方法的有效性, 我们分别在在塑料表面 (plastic surface, PS)、充电器外壳 (charger shell, CS)、SIM 卡 (SIM card, SC) 和 SIM 卡槽 (SIM card slot, SCS) 这 4 类不同的工业数据上进行测试, 并采用 FID[29]、SSIM[30]、L2loss 等指标来量化地评估最终的结果.

3.2 与现有方法比较

此外, 我们将本文方法与现有主流方法 Attention-GAN[19], CycleGAN[6], Pix2pix[31], U-GAT-IT[20] 等进行比较. 为了保证验证的公平性, 所有的模型都会使用官方提供的源代码进行训练直至收敛并测试.

表 1 展示了各个模型在 4 个数据集上的量化测试结果. 从结果可以看出, 本文方法的生成结果无论特征相似度还是图像相似度, 表现都较为优异. 实际效果如图 7 所示, 从图中我们可以发现基于注意的 Attention-GAN 和 U-GAT-IT 由于采用半监督学习缺乏一定的指导, 在提取特征时, 这些显著区域往往占据更大的权重, 在学习过程中容易产生注意漂移, 最终影响图像的生成. 因此, 在具有显著特征的场景的 PS 和 CS 数据集中表现不佳. 尽管我们的方法同样依赖于注意机制, 但当我们引入高斯热力图时, 模型的注意力也会随着指引更加关注字符区域并产生更好的擦除结果.

表 1 4 个不同数据集上不同模型的量化对比结果

模型	PS			CS			SC			SCS		
	FID	SSIM	L2loss	FID	SSIM	L2loss	FID	SSIM	L2loss	FID	SSIM	L2loss
CycleGAN	42.99	0.8462	36.42	160.56	0.867	56.12	44.84	0.9659	29.98	73.36	0.969	16.27
Pix2pix	110.25	0.7328	115.46	259.08	0.7171	224.94	27.74	0.992	8.48	153.16	0.920	256.15
AttentionGAN	134.29	0.530	78.41	157.23	0.556	36.57	21.52	0.996	2.80	110.13	0.914	161.80
U-GAT-IT	103.49	0.753	57.18	165.69	0.798	63.03	61.22	0.966	89.99	70.31	0.961	63.61
Ours	35.22	0.951	22.20	44.95	0.9491	18.27	22.34	0.995	3.23	54.19	0.975	23.01

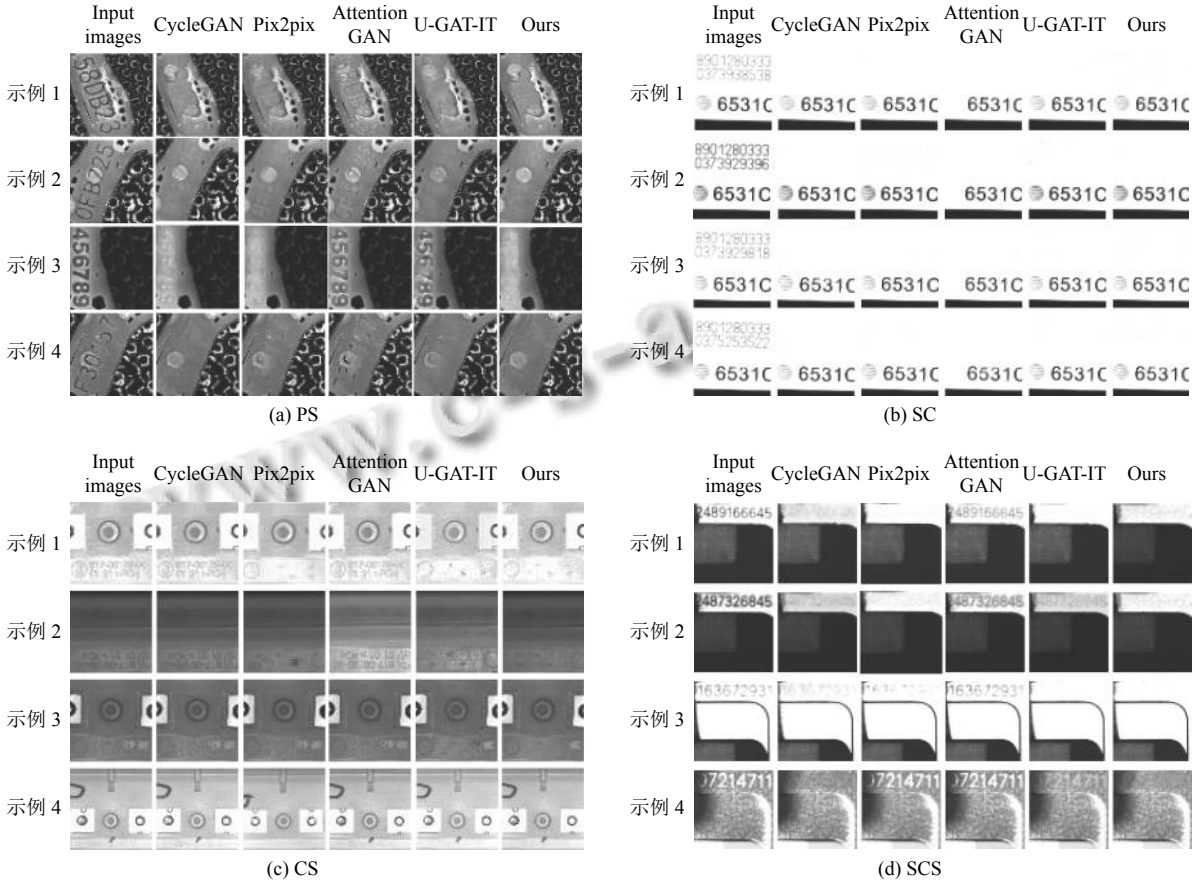


图 7 在 4 个数据集上的实际对比效果

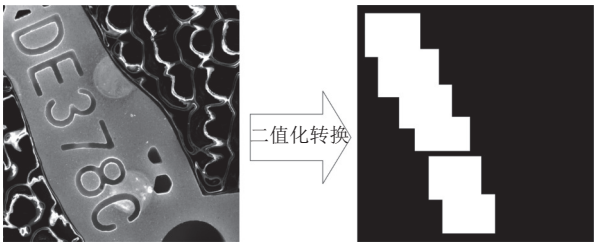
4 消融实验

在这一部分中, 我们研究了本方法中每个模块对于整体性能的影响, 量化结果如表 2 所示. 通过逐渐去除高斯地图 (Gaussian map, GM)、区域得分层 (region score layer, RSL) 和注意力 (attention) 等模块, 模型的抹除效果逐渐降低, 说明去除的模块对模型都有良好的影响.

表 2 消融实验的量化结果

Attention	RSL	GM	FID	SSIM	L2loss
√	√	√	25.22	0.951	22.20
√	√	×	44.91	0.879	32.84
√	×	×	69.78	0.878	36.51
×	×	×	103.49	0.753	57.18

为了评估 GM 对模型的影响, 我们将高斯热图换成了二值分割图, 如图 8 所示. 具体来说, 字符的目标框被选中的区域被设置为白色, 其余的被设置为黑色. 结果表明, 二值分割图可以作为字符擦除的引导, 但是这种刚性的区域约束使得模型对于区域特征的擦除效果较差, 这可能导致字符停留在局部擦除中.



(a) 原图 (b) 二值化图

图 8 二值分割示意图

5 结论

实验表明, 显著性非字符区域会影响模型在擦除字符过程中的注意力, 从而产生注意力漂移. 为了防止此类现象发生, 我们提出了一种新的基于高斯引导的字符擦除模型. 基于注意机制下的模型, 通过强引导的方式成功地将注意力集中在的字符区域上, 从而获得了更好的擦除效果. 同时, 大量的实验表明, 该方法在不同场景下都拥有很高的灵活性.

参考文献

1 Inai K, Pålsson M, Frinken V, *et al*. Selective concealment of characters for privacy protection. Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm: IEEE, 2014. 333–338.

2 Nakamura T, Zhu AN, Yanai K, *et al*. Scene text eraser. Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017. 832–837.

3 Suh S, Lee H, Lukowicz P, *et al*. CEGAN: Classification enhancement generative adversarial networks for unraveling data imbalance problems. Neural Networks, 2021, 133: 69–86. [doi: 10.1016/j.neunet.2020.10.004]

4 Creswell A, White T, Dumoulin V, *et al*. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 2018, 35(1): 53–65. [doi: 10.1109/MSP.2017.2765202]

5 Liu CY, Jin LW, Liu YL, *et al*. Don't forget me: Accurate background recovery for text removal via modeling local-global context. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 409–426.

6 Almahairi A, Rajeswar S, Sordoni A, *et al*. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 195–204.

7 Chen XY, Xu C, Yang XK, *et al*. Attention-GAN for object transfiguration in wild images. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 167–184.

8 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

9 Xu YC, Fu MT, Wang QM, *et al*. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1452–1459. [doi: 10.1109/TPAMI.2020.2974745]

10 Baek Y, Lee B, Han DY, *et al*. Character region awareness for text detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9357–9366.

11 Long SB, He X, Yao C. Scene text detection and recognition: The deep learning era. International Journal of Computer Vision, 2021, 129(1): 161–184. [doi: 10.1007/s11263-020-01369-0]

12 Khodadadi M, Behrad A. Text localization, extraction and inpainting in color images. Proceedings of the 20th Iranian

Conference on Electrical Engineering (ICEE2012). Tehran: IEEE, 2012. 1035–1040.

13 Wagh PD, Patil DR. Text detection and removal from image using inpainting with smoothing. Proceedings of the 2015 International Conference on Pervasive Computing (ICPC). Pune: IEEE, 2015. 1–4.

14 Zhang ST, Liu YL, Jin LW, *et al*. EnsNet: Ensconce text in the wild. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 801–808.

15 Tursun O, Zeng R, Denman S, *et al*. MTRNet: A generic scene text eraser. Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney: IEEE, 2019. 39–44.

16 Tursun O, Denman S, Zeng R, *et al*. MTRNet++: One-stage mask-based scene text eraser. Computer Vision and Image Understanding, 2020, 201: 103066. [doi: 10.1016/j.cviu.2020.103066]

17 Tang ZM, Miyazaki T, Sugaya Y, *et al*. Stroke-based scene text erasing using synthetic data for training. IEEE Transactions on Image Processing, 2021, 30: 9306–9320. [doi: 10.1109/TIP.2021.3125260]

18 Bian XW, Wang CQ, Quan WZ, *et al*. Scene text removal via cascaded text stroke detection and erasing. Computational Visual Media, 2022, 8(2): 273–287. [doi: 10.1007/s41095-021-0242-8]

19 Tang H, Liu H, Xu D, *et al*. AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 1972–1987. [doi: 10.1109/TNNLS.2021.3105725]

20 Kim J, Kim M, Kang H, *et al*. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2019.

21 Arrieta AB, Díaz-Rodríguez N, Ser JD, *et al*. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82–115. [doi: 10.1016/j.inffus.2019.12.012]

22 Zhou BL, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929.

23 Lin M, Chen Q, Yan SC. Network in network. arXiv:1312.4400, 2013.

24 Ling J, Xue H, Song L, *et al*. Region-aware adaptive instance normalization for image harmonization. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9357–9366.

25 Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1510–1519.

26 Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450, 2016.

27 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1302–1310.

28 Duan HD, Zhao Y, Chen K, *et al*. Revisiting skeleton-based action recognition. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2969–2978.

29 Heusel M, Ramsauer H, Unterthiner T, *et al*. GANs trained by a two time-scale update rule converge to a local nash equilibrium. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

30 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

31 Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5967–5976.

(校对责编: 孙君艳)