Software Engineering and Applications 软件工程与应用, 2023, 12(2), 345-353
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.122035

Hans汉斯

基于上下文注意的场景文本识别

董田荣

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2023年3月17日；录用日期：2023年4月19日；发布日期：2023年4月29日

摘　要

作为计算机视觉领域的研究热点，自然场景中不规则文本的识别是一项具有挑战的任务。本文提出了一种简单有效的方法来识别不规则文本。所提出的方法采用薄板样条变换将不规则文本转换为规则文本，采用融合空间多尺度感知模块的ResNet34提取文本特征，然后将文本特征通过Bi-LSTM编码为上下文特征。整个模型分别使用上下文感知模块和文本特征增强模块进行监督。上下文感知模块关注于文本特征与上下文特征构成的新的特征空间，文本特征增强模块重点关注单个字符本身以处理无上下文语义的文本行。与其他的文本识别模型相比，所提出的方法对于不规则文本的识别能力有较大的提高，同时保持了对于常规文本的识别能力。在通用的场景文本数据集上通过大量的实验验证了模型对于不规则文本识别的有效性。

关键词

文本识别，不规则文本，薄板样条变换，Bi-LSTM，多尺度感知

Context Attention Network for Scene Text Recognition

Tianrong Dong

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 17th, 2023; accepted: Apr. 19th, 2023; published: Apr. 29th, 2023

Abstract

As a research hotspot in the field of computer vision, the recognition of irregular text in natural scenes is a challenging task. In this paper, we propose a simple and effective method to recognize irregular text. The proposed method uses Thin Plate Spline to convert irregular text into regular text, ResNet34 with fused spatial multiscale perception module to extract text features, and then

文章引用: 董田荣. 基于上下文注意的场景文本识别[J]. 软件工程与应用, 2023, 12(2): 345-353.
DOI: 10.12677/sea.2023.122035

encodes text features into contextual features by Bi-LSTM. The whole model is supervised using a context-aware module and a text feature enhancement module, respectively. The context-aware module focuses on a new feature space composed of text features and contextual features, and the text feature enhancement module focuses on individual characters to handle text lines without contextual semantics. Compared with other text recognition models, the proposed approach has a large improvement in the recognition of irregular text while maintaining the recognition capability for regular text. The effectiveness of the model for irregular text recognition is verified by extensive experiments on scene text datasets.

Keywords

Text Recognition, Irregular Text, Thin Plate Spline, Bi-LSTM, Multi-Scale Perception

1. 引言

文字是人类历史上最具影响力的发明之一，在我们日常生活中扮演着重要的角色。场景文本识别指的是在自然场景中阅读文本，在工业界具有广泛的应用，例如自动驾驶、盲人辅助技术、图像检索。目前，光学字符识别系统已经成功地应用于文档识别中并取得了显著成效。但是，与文档中的文本不同，自然场景中的文本具有复杂背景和任意成像的特点，这些特点会导致文本出现模糊、失真、低分辨率、低对比度等现象。此外，多变的字体类型也增加了准确识别场景文本的难度。因此对于自然场景中的文本，尤其是不规则文本的精确识别具有重要的研究意义。

2. 相关工作

近年来，文本识别在深度学习领域的驱动下取得了一定的进展。目前的文本识别方法将文本识别视为序列学习问题，大多数基于序列学习的方法基于由编码器和解码器组成。编码器使用卷积神经网络或者长短期记忆网络将输入图像编码为固定维度的向量。解码器使用连接主义时间分类(CTC)或注意力机制将编码器编码的特征序列解码为目标字符串。

上下文无关的方法通常将场景文本识别视为单纯的视觉分类任务。CRNN [1]使用卷积神经网络(CNN)和递归神经网络(RNN)提取给定文本图像的序列视觉特征，然后直接输入 CTC 解码器预测每个时间步的字符类别。为了减轻 CTC 损失的反向传播的计算压力，Xie 等人[2]提出了聚集交叉损失，沿时间维度优化每个字符的概率，大大提高了效率。Liao 等人[3]通过像素级分类来预测每个位置的字符类别，然后通过启发式规则将字符汇集成文本行，但是该方法需要昂贵的字符级注释。对于不规则文本，为了消除失真和曲率带来的负面影响，Shi 等人[4]在序列识别之前加入了具有多个控制点对的校正模块。Cheng 等人[5]从四个方向提取场景文本识别特征，并设计了一个滤波门来控制各个方向的贡献。Li 等人[6]采用二维注意力机制提升不规则文本识别的准确率。STAN [7]提出了一种基于序列转换注意力的网络。以上这些工作均专注于如何提取更有效的视觉特征，忽略了语义上下文的重要性。

为了提高模型性能，多尺度上下文信息在视觉识别中发挥重要的作用。递归神经网络(RNN)具有捕获长距离依赖的能力，被广泛应用于上下文建模，但是 RNN 的建模不考虑上下文的二维空间性，而且需要较高的计算能力。为了获取上下文信息，李等人[8]提出了一种多级特征选择的文本识别方法，采用堆

叠块的体系结构逐步细化文本特征和上下文特征。SEED [9]提出使用预训练的语言模块来获取上下文信息。然而现有的方法只能捕获单个维度的上下文信息，相比之下，所提出的方法采用基于空洞卷积改进的特征提取网络进行特征提取，可以在两个维度上处理上下文建模。

3. 方法

3.1. 网络结构

所提出方法的网络结构图如图 1 所示。首先使用变换网络对输入的文本图像进行修正。然后，使用基于空洞卷积改进的特征编码器对变换后的图像进行文本特征提取，文本特征通过双层 Bi-LSTM 处理得到上下文特征。采用文本特征增强模块对文本特征进行细化，上下文感知解码器对接收的文本特征和上下文特征进行解码并输出一维字符序列。文本特征增强模块作为额外监督仅在训练阶段执行，预测阶段被移除。

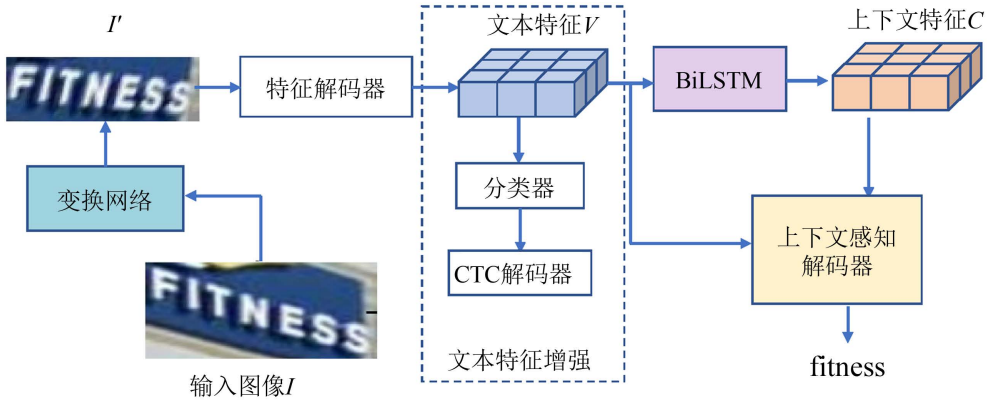

Figure 1. Overall structure
图 1. 整体结构

3.2. 变换网络

对于自然场景中的不规则文本以及多方向文本，使用变换网络对输入图像 I 进行处理，将其变换成规范化图像 I'。受论文[10]的启发，变换网络采用薄板样条变换(TPS)进行处理。TPS 在一组基准点之间使用平滑样条插值，将文本区域调整为预定义的形状。首先在文本区域顶部和底部设置预定义数量的基准点，然后将待预测的文本区域调整为固定大小。该变换网络可以去除掉图片的非文本区域，提高文本特征提取的有效性。

3.3. 基于空洞卷积改进的特征提取网络

常规文本识别将输入图像编码为一维特征序列，不规则文本识别将校正之后的文本编码为一维特征序列。传统的识别模型通常使用图 2 所示的 ResNet34 作为特征编码器，将大小为 $H \times W$ 的输入图像 I 编码为特征地图 F。为了充分学习多尺度上下文推理，融合多尺度信息，我们提出了图 3 所示的融合了空间多尺度感知模块(SMPM)的 ResNet34 进行特征提取，从而更好地适应场景文本的内在特征。

空间多尺度感知模块(Spatial Multi-Scale Perception Module, SMPM)。空洞卷积可以扩大感受野，更有效地聚合不同尺度的信息。因此，受 Li 等人[11]的启发，采用基于空洞卷积的空间多尺度感知模块提升模型的感受野。如图 4 所示，F_1 是一个(1, 1)的空洞卷积，输出的每个元素的感受野为 3×3。F_2 是一个(3, 2)的空洞卷积，输出的每个元素的感受野为 9×7。F_3 是一个(8, 4)的空洞卷积，输出的每个元素的感受野为 25×15。

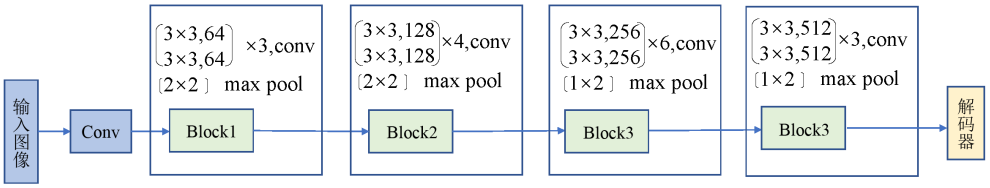

Figure 2. ResNet34encoder
图 2. ResNet34 编码器

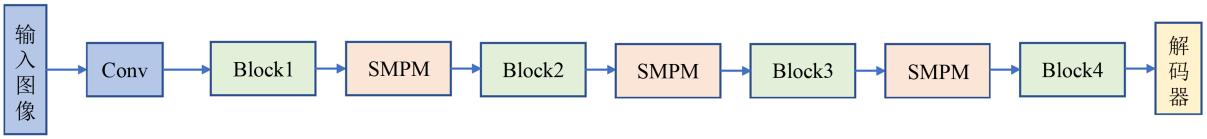

Figure 3. Modified ResNet34 encoder
图 3. 改进的 ResNet34 编码器

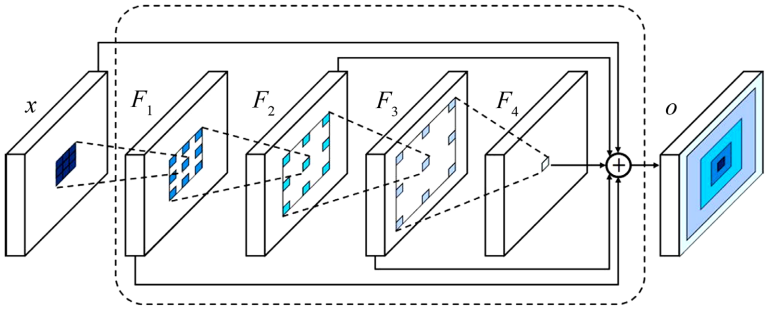

Figure 4. Spatial multi-scale perception module
图 4. 空间多尺度感知模块

3.4. 上下文特征

基于 CNN 的特征编码器提取的特征受限于其接受域，可能会由于缺乏上下文信息导致性能下降。为了改进这个缺点，在特征地图上使用双向 BLSTM 网络得到上下文特征向量，有效地捕获文本中的双向依赖，每层隐藏状态大小为 512。如图 5 所示，在每个时间步，LSTM 编码器以二维特征地图作为输入，沿垂直轴进行最大池化，并且更新隐藏状态。Zuo 等人[12]指出随着 Bi-LSTM 中编码器层数的增加，精度会下降。因此，我们仅使用两层 Bi-LSTM 进行处理。

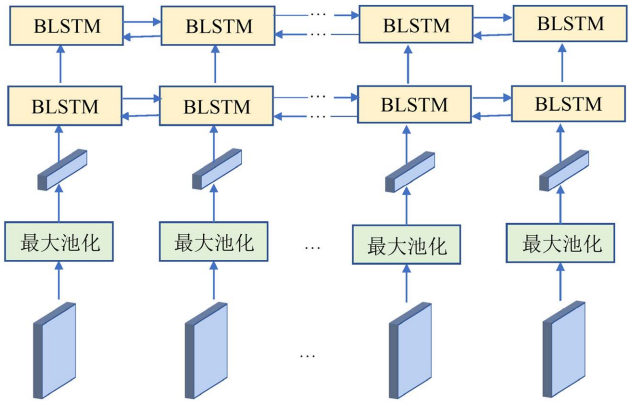

Figure 5. LSTM encoder
图 5. LSTM 编码器

3.5. 文本特征增强

采用文本特征增强对文本进行字符级建模，得到每个字符更精确的特征。文本特征增强模块不考虑特征向量之间的上下文关系，每个特征向量是相互独立的。首先通过一个全连接层将文本特征 V 转换为一个长度为 N 的序列 L，将输出序列输入到 CTC 解码器。CTC 解码器将序列 L 转化为标签序列上的条件概率分布，然后选择概率最大的字符进行输出。具体过程如下所示：

$$l = B\left(\arg\max_{\pi} p(\pi \mid L)\right) \tag{1}$$

其中 B 是一个映射函数，用于删除重复的字符和空格。CTC 计算每个字符的条件概率，并且假设不同字符是条件独立的。π 的概率表示为：

$$p(\pi \mid L) = \prod_{t=1}^{N} y_{\pi_t}^{t} \tag{2}$$

上述公式中 $y_{\pi_t}^{t}$ 是在时间步 t 时生成字符的概率。

3.6. 上下文感知解码器

最近的编码器－解码器模型依赖于基于注意力的解码器，该解码器可以在输出序列和编码器为每个文本图像生成的特征之间进行对齐。但是，基于注意力的解码器对于噪声和变化太敏感，当出现注意力漂移的问题时通常会解码失败，从而得到错位的字符序列。

与传统的注意力解码器不同，受 Wang 等人[13]的启发，模型采用了上下文感知解码器，将文本特征、上下文特征与注意力图作为输入，如图 6 所示，由一个用于获取上下文信息的 GRU 层和一个用于进行预测的线性层组成。首先，将文本特征 V 与上下文特征 C 连接形成一个新的特征空间 $S = (V, C)$。在特征空间 S 上进行自注意力操作，使用全连接层从这些特征中计算得到注意力图 M。然后，计算注意力特征向量 g_t，如公式 3 所示：

$$g_t = \sum_{x=1}^{n}\sum_{y=1}^{n} M_{t,x,y} S_{x,y} \tag{3}$$

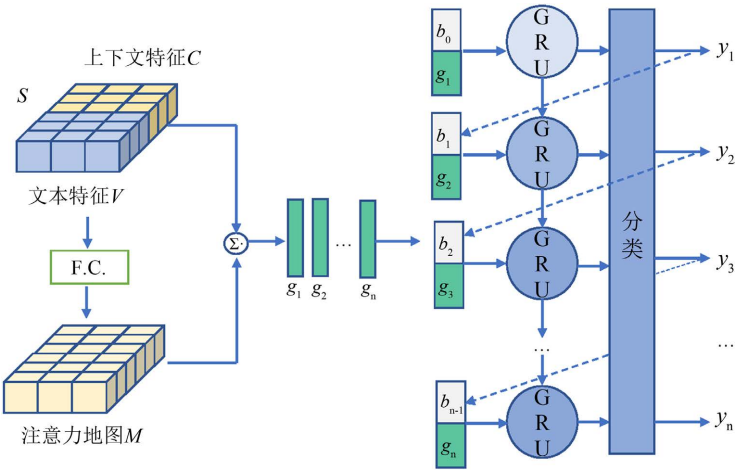

Figure 6. LSTM encoder
图 6. LSTM 编码器

在时间步 t 时，分类器输出结果 y_t：

$$y_t = wc_t + a \tag{4}$$

在公式 4 中，c_t 是 GRU 的隐藏状态，可以表示为公式 5：

$$c_t = \mathrm{GRU}\left(\left(b_{t-1}, g_t\right), c_{t-1}\right) \tag{5}$$

其中 b_t 为前一步解码结果 y_{t-1} 的嵌入向量。

4. 优化

所提出的模型采用文本特征增强损失和上下文感知损失共同监督，如公式 6 所示。

$$L = \lambda_1 \cdot L_{\mathrm{CTC}} + \lambda_2 \cdot L_{\mathrm{Con}} \tag{6}$$

上式中，λ_1 和 λ_2 是用来平衡不同监督的超参数，分别设置为 0.1 和 1.0。L_{CTC} 是 CTC 解码器的损失函数。L_{Con} 是上下文感知解码器的损失，可以表示为：

$$L_{\mathrm{Con}} = -\sum_{t=1}^{T} \log P\left(z_t \mid I, \theta\right) \tag{7}$$

其中，θ 表示模型在时间步 t 所有可训练的参数，z_t 表示在时间步 t 的真实值。

5. 实验

5.1. 实验细节

为了便于进行比较，模型训练和评估的设置与最新的方法保持一致。实验采用 Adam 优化器，学习率设置为 $1e^{-4}$，批大小设置为 185。所有实验均在 NVIDIA 3090Ti GPU (32GB RAM)上进行。在训练期间采用了数据增强技术，通过随机调整大小，失真，色彩抖动等增强 35%的输入图像。训练和测试期间，所有图像大小均调整为 32 × 128。字符集包括 26 个英文字母，10 个数字和结束符“EOS”。最大序列长度设置为 32。

5.2. 数据集

训练集由两个合成数据集组成，分别是 MJ 和 ST。测试集由两种类型的数据集组成：常规文本数据集和不规则文本数据集。

常规文本数据集包括 IIIT5k，SVT 和 ICADR2013，IIIT5k 数据集是从互联网收集的，训练集包含 2000 张图像，测试集包含 3000 张图像。SVT 数据集是从谷歌街景中收集的。测试集包含 647 张图片。ICADR2013 包含 848 个训练图像和 1015 个测试图像。

不规则文本数据集包括 ICADR2015，SVTP 和 CUTE。ICADR2015 数据集包含 2077 张测试图像，大部分图像是模糊的和多方向的。SVTP 数据集包含 639 张用于测试的图片。CUTE80 包含 80 幅自然场景中拍摄的高分辨图像，测试集包括 288 张图片，是专门为评估弯曲文本识别的性能而收集的。

5.3. 消融实验

在 3.3 节使用了文本特征增强模块作为中间监督对文本识别模型进行优化，以获取更精细的局部视觉特征。从表 1 可以看出，加入文本特征增强模块后规则文本和不规则文本的识别准确率均有所提高。

5.4. 与最新的方法进行比较

我们在六个数据集上对所提出的方法评估了所提出模型的准确性，并与最新的方法进行了比较。为了保证结果公平，与其他方法一样，在两个合成数据集上进行训练，在推理阶段不使用词典。如表 2 所

示。可以看到，所提出的方法在常规数据集上均优于其他方法，在包含不规则文本的街景文本数据集 SVTP 和 CUTE 上准确率分别达到了 80.2%和 84.2%。与其他字符级注释的方法相比，我们的方法在常规数据集和不规则数据集上均取得了不错的性能。部分数据集的识别结果如图 7 所示，与 Li 等人[6]提出的方法相比，本文的方法对于不规则文本的识别结果更准确。

Table 1. Ablation experiment of text feature enhancement module
表 1. 文本特征增强模块的消融实验

文本特征增强模块	规则文本数据集			不规则文本数据集		
	IIIT5k	IC13	SVT	IC15	SVTP	CUTE
×	92.9	92.2	89.7	76.4	78.6	78.2
√	**94.1**	**94.6**	**90.6**	**77.3**	**80.2**	**84.2**

Table 2. Comparison with state-of-the-art methods
表 2. 与最新的方法进行比较

方法	规则文本数据集			不规则文本数据集		
	IIIT5k	IC13	SVT	IC15	SVTP	CUTE
ASTER [4]	93.4	91.8	89.5	76.1	78.5	79.5
SAR [6]	91.5	91.0	84.5	69.2	76.4	83.3
MORAN [14]	91.2	92.4	88.3	68.8	76.1	77.4
ACSR [2]	82.3	89.7	82.6	68.9	70.1	68.9
SEED [9]	93.8	92.8	89.6	80	81.4	83.6
EPAN [15]	94.0	94.5	88.9	73.3	79.4	82.6
STAN [7]	94.1	92.8	90.5	76.7	82.2	83.3
ESIR [16]	93.3	91.3	90.2	76.9	79.6	83.3
本文方法	**94.1**	**94.6**	**90.6**	**77.3**	**80.2**	**84.2**

Figure 7. Recognition results display
图 7. 识别结果展示

5.5. 鲁棒性

自然场景中的文本通常受到各种各样因素的干扰，为了验证所提出的模型是否对细微干扰具有敏感

性，在经过处理的 IIT5k 和 IC13 数据集上进行了深入研究。对于 IIIT5k 数据集，通过重复边界像素，以额外 10%的垂直高度和 10%的水平宽度填充图像。对于 IC13 图像，将边界框扩展为具有额外的 10%高度和 10%宽度的矩形。将所提出的模型与在未经过处理数据集上性能比较好的方法 ASTER [4]、SEED [9]和 STAN [7]进行对比，结果如表 3 所示，IC13-d 和 IIIT5k-d 分别代表经过处理之后的数据集，“ac”代表准确率，“gap”表示与原始数据集的差距，“ratio”表示精度下降比。可以看出与其余三种方法相比，即使是存在干扰的数据集，所提出的方法仍然保持具有竞争力的性能。

Table 3. Robustness analysis
表 3. 鲁棒性分析

方法	IC13	IC13-d			IIIT5k	IIIT5k-d		
	ac	ac	gap	ratio	ac	ac	gap	ratio
ASTER [4]	91.8	89.1	−2.7	3.0%	93.4	89.6	−3.8	4.0%
SEED [9]	92.8	90.1	−2.7	2.9%	93.8	90.3	−3.5	3.7%
STAN [7]	92.8	90.3	−2.5	2.7%	94.1	91.2	−2.9	3.1%
本文方法	94.6	92.7	−1.9	2.0%	94.1	91.6	−2.5	2.7%

6. 总结

本文提出了一种有效且鲁棒的上下文感知文本识别模型。采用 ResNet34 作为主干网络，与基于空洞卷积的空间多尺度感知模块融合提取文本特征，增强了模型的感受野。采用双层 Bi-LSTM 编码器将文本特征编码为鲁棒的上下文特征，文本特征增强模块和上下文特征解码器的联合监督提高了对于场景中不规则文本的识别能力，同时保持了对常规文本的识别性能。在公开数据集上进行大量的实验验证了所提出方法的有效性。在未来的工作中，我们希望提高对于艺术字的识别能力。

参考文献

[1] Shi, B., Xiang, B. and Cong, Y. (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 & Machine Intelligence*, **39**, 2298-2304. https://doi.org/10.1109/TPAMI.2016.2646371

[2] Xie, Z., Huang, Y., Zhu, Y., *et al*. (2019) Aggregation Cross-Entropy for Sequence Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 6538-6547. https://doi.org/10.1109/CVPR.2019.00670

[3] Liao, M., Zhang, J., Wan, Z., *et al*. (2019) Scene Text Recognition from Two-Dimensional Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8714-8721. https://doi.org/10.1609/aaai.v33i01.33018714

[4] Shi, B., Yang, M., Wang, X., *et al*. (2018) Aster: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 2035-2048. https://doi.org/10.1109/TPAMI.2018.2848939

[5] Cheng, Z., Xu, Y., Bai, F., *et al*. (2018) Aon: Towards Arbitrarily-Oriented Text Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 5571-5579. https://doi.org/10.1109/CVPR.2018.00584

[6] Li, H., Wang, P., Shen, C., *et al*. (2019) Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8610-8617. https://doi.org/10.1609/aaai.v33i01.33018610

[7] Lin, Q., Luo, C., Jin, L., *et al*. (2021) STAN: A Sequential Transformation Attention-Based Network for Scene Text Recognition. *Pattern Recognition*, **111**, Article ID: 107692. https://doi.org/10.1016/j.patcog.2020.107692

[8] 李利荣, 张开, 张云良, 等. 基于多级特征选择的自然场景文本识别算法[J]. 光电子·激光, 2022(5): 33.

[9] Qiao, Z., Zhou, Y., Yang, D., *et al*. (2020) Seed: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 13528-13537. https://doi.org/10.1109/CVPR42600.2020.01354

[10] Baek, J., Kim, G., Lee, J., *et al*. (2019) What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 4715-4723. https://doi.org/10.1109/ICCV.2019.00481

[11] Li, H., Yang, D., Huang, S., *et al*. (2020) Two-Dimensional Multi-Scale Perceptive Context for Scene Text Recognition. *Neurocomputing*, **413**, 410-421. https://doi.org/10.1016/j.neucom.2020.06.071

[12] Zuo, L.Q., Sun, H.M., Mao, Q.C., *et al*. (2019) Natural Scene Text Recognition Based on Encoder-Decoder Framework. *IEEE Access*, **7**, 62616-62623. https://doi.org/10.1109/ACCESS.2019.2916616

[13] Wang, T., Zhu, Y., Jin, L., *et al*. (2020) Decoupled Attention Network for Text Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12216-12224. https://doi.org/10.1609/aaai.v34i07.6903

[14] Luo, C., Jin, L. and Sun, Z. (2019) Moran: A Multi-Object Rectified Attention Network for Scene Text Recognition. *Pattern Recognition*, **90**, 109-118. https://doi.org/10.1016/j.patcog.2019.01.020

[15] Huang, Y., Sun, Z., Jin, L., *et al*. (2020) EPAN: Effective Parts Attention Network for Scene Text Recognition. *Neurocomputing*, **376**, 202-213. https://doi.org/10.1016/j.neucom.2019.10.010

[16] Zhan, F. and Lu, S. (2019) Esir: End-to-End Scene Text Recognition via Iterative Image Rectification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 16-17 June 2019, 2059-2068. https://doi.org/10.1109/CVPR.2019.00216