基于动态排列自回归的场景文本识别网络

王嘉宝[1] 陈宏辉[1] 陈平平[1]
WANG Jiabao CHEN Honghui CHEN Pingping

摘 要 随着计算机视觉广泛渗透到生产和生活中的各个领域，场景文本识别面临着愈发复杂的考验。纯视觉的场景文本识别模型侧重于构建有效的视觉特征提取网络，而缺乏对文本语义的理解，因此在处理遮挡或模糊文本图像时常遇到瓶颈。针对该问题，提出了一种利用语义信息辅助识别任务的场景文本识别算法。首先通过 Transformer 视觉编码器 ViT 提取特征，其次利用双分支结构的特征交互模块增强视觉特征，接着联合动态排列语言模型实现自回归解码。所提出的算法充分利用视觉特征和语义特征，有效地减少了遮挡等复杂文本的识别难度，实现了对场景文本的鲁棒性识别。实验结果表明，所提出的算法在 6 个基准数据集上实现了 96.65% 的平均识别精度，展现了显著的竞争力。

关键词 深度学习；场景文本识别；动态排列语言模型；自回归

doi：10.3969/j.issn.1672-9528.2024.05.001

0 引言

场景文本识别（scene text recognition，STR）的目标是在自然场景图像中准确识别文本序列，该技术已广泛应用于图像搜索、自动驾驶、增强现实等领域。然而，由于场景文本图像具有文本多样、背景复杂、图片遮挡或失真等挑战，仅依赖视觉模型（vision model，VM）产生的视觉特征限制了文本识别网络的性能。

为了克服这些挑战，当前的研究致力于将语义信息巧妙地融入 STR 模型中。例如，ASTER[1] 采用循环神经网络（recurrent neural network，RNN）和注意力机制相结合的方式来捕获文本的语义信息。使用 Transformer[2] 的 Bi-STET[3]，通过自回归（autoregressive，AR）训练两个单向解码器学习内部语言模型（language model，LM）。而这些方法仅限于单向 AR 解码，即从左到右解码、从右到左解码或者两者简单相结合的方式，因此这些方法只能获得有限的上下文语义。为了解决这个限制，SRN[4] 和 ABINet[5] 结合 VM 和外部 LM 进行双向的上下文预测和细化。尽管 VM 和外部 LM 结合的场景文本识别取得了一定的成效，但是存在这两个模型之间独立性的影响因素，这可能会存在错误纠正问题。Parseq[6] 引入排列语言模型（permutation language modeling，PLM）[7]，在使用内部 LM 的同时实现外部 LM 的细化能力。由于 PLM 在训练过程中是固定的，这可能导致模型在学习中无法动态调整以适应不同的排列情况。因此，本文提出了一种基于动态排列语言模型（dynamic permutation language modeling，DPLM）的 STR 算法，通过引入注意力机制，DPLM 能够使模型更主动地学习排列的生成方式，以便更好地适应 STR 任务的需求。同时，考虑到场景文本识别数据集样本不均衡的问题 [8]，BatchformerV2[9] 的双分支结构被引入特征交互模块（feature interaction module，FIM）中，以探索样本之间的关系。

本研究的核心是充分利用和结合视觉特征和语义特征来实现 STR 任务。首先将图片输入编码器中提取视觉特征；其次将视觉特征输入双分支结构的 FIM 中，从而增强特征表示；同时利用 DPLM 动态生成不同排列，以学习不同排列下的文本特征表示；然后通过注意力掩码实现 AR 解码；最后通过一轮迭代细化，输出最终预测文本序列。实验结果表明，本文提出的文本识别网络在 6 个基准数据集上达到 96.65% 的平均识别精度，具备一定的优势。

1 相关工作

1.1 场景文本识别

场景文本识别作为计算机视觉领域中一个长期存在的任务，经受了广泛的研究和深入探讨。Shi 等人 [10] 提出的基于连接主义时间分类（connectionist temporal classification，CTC）的 STR 方法，首先采用卷积神经网络（convolutional neural network，CNN）提取视觉特征，其次采用 RNN 对视觉特征进行序列建模，最后送入 CTC 解码器对 CNN 和 RNN 进行端到端训练，实现结果的预测。为了提升 STR 算法在不规则场景文本下的性能，诸如 RARE[11] 和 MORAN[12] 等受到自然语言处理相关理论和算法的启发，采用注意力机制对文本序列进行建模和解码 [13]。以 CA-FCN[14] 和 Textscanner[15]

1. 福州大学物理与信息工程学院 福建福州 350108
[基金项目] 国家自然科学基金面上项目“基于物理层网络编码的随机多址接入技术研究”（61871132）

为代表的基于分割的 STR 算法将 STR 任务视为像素级分类任务，其中每个字符都是目标类别，这类算法在一定程度上推进了 STR 任务的发展。随着 Transormer 算法的兴起，ViT[16] 尝试在图像分类任务中引入 Transformer 算法，并取得了一定成效。在此基础上，ViTSTR[17] 将 ViT 架构应用于 STR 任务中，ViTSTR 主要关注其中的 VM，直接解码 ViT 编码器学习到的视觉特征。然而，这些方法主要依赖于输入图像的特征进行预测，缺乏对文本语义的理解，因此当遇到遮挡或者文本图像失真的情况时，算法的鲁棒性较差。

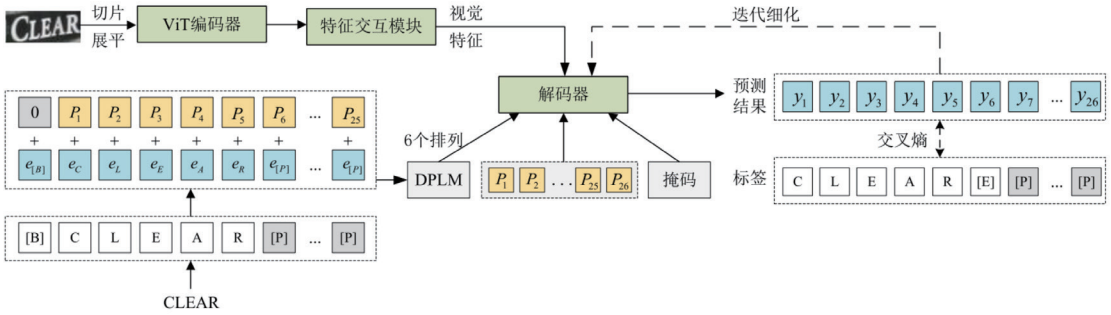

图 1 整体网络结构

1.2 基于语义信息的 STR

基于语义信息的 STR 方法在文本识别领域已经成为主流。这类方法致力于通过利用语义信息来辅助识别任务，以提升整体的识别性能。SEED[18] 采用预训练好的 LM 的词嵌入，通过监督预测语义信息和全局语义信息来引导解码过程。SRN 引入全局语义推理模块，通过多路并行传输捕获全局上下文语义，以在没有 AR 操作的情况下细化输出序列。在此基础上，ABINet 阻止视觉和语言模型之间的梯度流来显式 LM 建模，同时采用 LM 迭代修正视觉预测，特别在处理低质量图片时表现出色，但是显式 LM 也存在对 VM 正确的预测结果进行错误纠正的问题。VisionLAN[19] 和 MATRN[20] 在这一领域进行了更深入的研究，致力于使用结合 VM 和 LM 的方法。VisionLAN 通过对字符进行选择性掩码，在训练阶段引入被掩码的字符获得相应特征图，从而引导视觉模型同时使用字符的视觉纹理特征和上下文的语义特征。而 MATRN 则通过多模态 Transformer 和视觉的掩码策略更新视觉和语义特征，实现 VM 和 LM 之间的交叉引用，从而增强了视觉和语义特征的表达能力。与前述方法不同，Parseq 采用了内部 LM 和 AR 解码结合的方法。它引入 PLM 学习，生成内部自回归 LM 集合，通过联合处理图像和上下文特征来执行解码过程和迭代细化过程，避免了 VM 和 LM 融合的需求。这种独特的方法为 STR 任务带来了新的思路和效果提升，然而训练过程中固定的排列语言模型无法灵活适应不同排列情况，在面对多样化的文本场景时，模型精度会受到一定的影响。

2 主要方法

2.1 网络结构

本算法整体模型遵循 Parseq 和 BatchFormerV2 二者相结合的方式，主要分为 4 个部分：编码器、特征交互模块、动态排列语言模型和解码器。本文提出的算法框架如图 1 所示。首先，通过 ViT 编码器从输入的图像中提取视觉特征；其次，采用特征交互模块在 batch 维度隐式探索样本之间的关系；随后，联合动态排列语言模型学习权重矩阵，利用注意力机制调整输入排列的顺序；最后，解码器将这些特征解码为具有上下文嵌入的文本预测结果。

2.2 编码器

编码器的构造参照 ViT，是一个 12 层的 ViT 编码器，由多头注意力（Multi-Head Attention，MHA）、层归一化（Layer Normalization，LN）、多层感知器（Multi-layer Perceptron，MLP）和残差连接构成。以场景文本图像输入，通过编码器提取视觉特征。

首先，将输入图像 $x \in \mathbb{R}^{H\times W\times C}$ 均匀地分成大小为 $P_H \times P_W$ 的块，数量为：

$$S=(H\times W)/(P_H\times P_W) \tag{1}$$

式中：H 表示长度，W 表示宽度，C 表示通道数。

其次，将它们展平，线性映射到指定维度 d 获得 Token map。

接着，将相同维度的位置编码添加到每个 Token 中。

最后，送入 ViT 进行处理得到特征 Z：

$$Z=Enc(x)\in\mathbb{R}^{\frac{H\times w}{P_H\times P_W}\times d} \tag{2}$$

2.3 特征交互模块

考虑到识别数据集样本不均匀的问题，本文引入 FIM 结构从 batch 维度隐式地探索样本之间的关系。受文献 [9] 启发，采用如图 2 所示的双分支结构的 FIM 处理特征，整个过程包含拆分、处理和连接操作，实现了对输入视觉张量的特征交互处理。具体流程如下：首

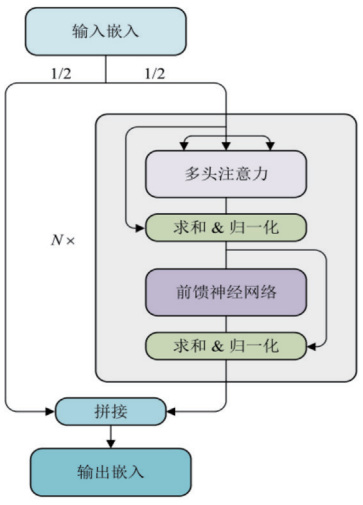

图 2 FIM 结构

先，通过拆分操作，从 batch 维度将经过编码器处理后的视觉特征 Z 平均分为两部分；其次，拆分后的第一部分特征不经过任何处理，而第二部分特征循环应用 Encoder 操作，通过多次的自注意力和前馈神经网络层，对这部分的特征进行进一步的交互和增强；接着，将两部分特征在 batch 维度上拼接形成新的特征张量 F。经过 FIM 结构处理后的特征维度与处理之前特征维度保持一致。

2.4 动态排列语言模型

对于给定的场景文本图像 x，本研究希望在模型参数 θ 的集合下，最大化其文本标签 $y=[y_1,y_2,..,y_T]$ 的概率。令 F_T 表示所有索引 [1,2,⋯,T] 可能排列的集合，采用如图 3 所示的 DPLM 生成排列，通过最大化以下公式的概率来训练解码器：

$$\log p(y \mid x)=\mathrm{E}_{f\sim F_T}\left[\sum_{t=1}^{T}\log p_\theta\left(y_{f_t}\mid y_{f<t},x\right)\right] \tag{3}$$

式中：f_t 表示特定排列 f 的第 t 个字符，$f<t$ 表示特定排列 f 的前 t-1 个字符。

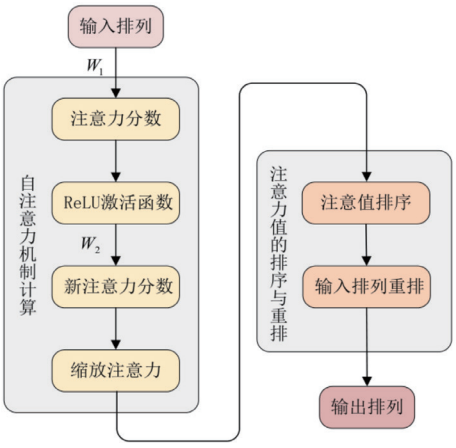

图 3 DPLM 结构

为了得到最佳的训练结果，需要遍历所有的序列排列组合情况，即形成 $T!$ 的排列组合，这在实际系统中是不可行的。同时，考虑到文本识别传统上依赖于从左到右的或从右到左单向序列建模的特征，通过多方向序列建模来改善输入和输出之间的相关性，即实验中对 DPLM 算法生成的排列进行 6 次掩码，如表 1 所示，包括 AR 掩码和随机掩码，并将添加掩码后的 6 个序列输入到解码器中预测文本。

表 1 注意掩码 m 示例

（a）AR 掩码

	[B]	y_1	y_2	y_3
y_1	1	0	0	0
y_2	1	1	0	0
y_3	1	1	1	0
[E]	1	1	1	1

（b）随机掩码

	[B]	y_1	y_2	y_3
y_1	1	0	1	1
y_2	1	0	0	0
y_3	1	0	1	0
[E]	1	1	1	1

DPLM 生成排列流程如下：将标签编码得到目标序列的张量表示，输入 DPLM 模型中动态调整排列顺序，其中 DPLM 算法旨在结合可学习的权重矩阵和注意力机制，获得更有意义的排列结果。在生成排列时，首先，通过参数 $W_1\in\mathbb{R}^{1\times 26}$ 将输入排列进行线性映射，引入可学习的关系权重使得模型能够动态调整排列顺序。接着，应用 ReLU 激活函数，以引入非线性变换，有助于模型更好地捕捉排列之间的复杂关系。随后，再次借助参数 $W_2\in\mathbb{R}^{26\times 1}$ 进行线性映射来和输入排列保持维度一致，并通过归一化处理获得缩放后的注意力值。这一系列处理的目的是动态地学习每个排列的重要性。最后，利用学到的注意力值对原始排列进行排序，生成新的排列。

2.5 解码器及解码策略

解码器的作用是从 FIM 处理后的视觉特征 F 中提取字符信息，输出预测文本。解码器由 1 层 Transformer 的 Decoder 构成，其中注意力头 d 是 12 个，数量为编码器中注意力头数量的两倍，解码器的输入由位置、带有位置信息的上下文词嵌入、掩码和视觉特征四个部分构成。解码阶段可以表示为：

$$F^{'}=Dec(p,c,m,F) \tag{4}$$

式中：$p\in\mathbb{R}^{(T+1)\times d}$ 是位置 Token，$c\in\mathbb{R}^{(T+1)\times d}$ 是带有位置信息的上下文词嵌入，$m\in\mathbb{R}^{(T+1)\times(T+1)}$ 是如表 2 所示的掩码。同时，为了简化模型，省略了 LN 层和 Dropout 层。解码器的第一个 MHA 计算公式为：

$$M_1=p+Soft\max(m+\frac{pc^T}{\sqrt{d}})c \tag{5}$$

解码器的第二个 MHA 计算公式为：

$$M_2=M_1+Soft\max(\frac{m_1F^T}{\sqrt{d}})F \tag{6}$$

式中：第二个 MHA 的输出 $W_2\in\mathbb{R}^{(T+1)\times d}$，用于预测文本字符序列：

$$F^{'}=Linear\left(MLP\left(M_2\right)+M_2\right) \tag{7}$$

表 1 是注意掩码 m 示例。以序列 [1, 2, 3] 为例，[B] 表示序列开始，[E] 表示序列结束。1 表示输出对相应的输入标记存在依赖性，0 表示输出对相应的输入标记不存在依赖性。

本研究采用如表 2 所示的从左到右的 AR 解码，解码过程中，每个字符都依赖于先前的字符，而不依赖于后面的字符，在后续的迭代过程中，前面的输出在 [E] 处截断作为迭代过程输入的上下文。

表 2 是从左到右 AR 解码示例。[B] 表示序列开始，[E] 表示序列结束。1 表示输出对相应的输入标记存在依赖性，0 表示输出对相应的输入标记不存在依赖性。

表 2 从左到右 AR 解码示例

	[B]	y_1	y_2	...	y_T
y_1	1	0	0	0	0
y_2	1	1	0	0	0
...	1	1	1	0	0
y_T	1	1	1	1	0
[E]	1	1	1	1	1

本文计算预测文本 F' 和真实标签 $\overline{y}$ 之间交叉熵的平均值，用于衡量识别的损失：

$$L=\frac{1}{K}\sum_{k=1}^{K}L_{ce}\left(F_k^{'},\overline{y}\right) \tag{8}$$

式中：L_{ce} 表示交叉熵。

3 实验结果分析

3.1 数据集

训练数据集：由真实数据集 ArT、COCO-Text、LSVT、MLT19、OpenVINO、RCTW17、ReCTS、TextOCR 和 Uber-Text 构成。

测试数据集：由数据集 IIIT5K、SVT、ICDAR 2013（IC13）、ICDAR 2015（IC15）、SVTP 和 CUTE80 构成。为了方便对比，将测试数据集分为总数 7248（IIIT5K、SVT、IC13_857、IC15_1811、SVTP、CUTE80） 和 总 数 7672（IIIT5K、SVT、IC13_1015、IC15_2077、SVTP、CUTE80）两组。

3.2 实验细节

模型配置：编码器采用 12 层的 ViT 结构，具有 6 个注意力头和 384 维嵌入。输入图像的大小为 128×32，每一个 patch 的大小是 8×4。FIM 中的 Encoder 设置为 384 维嵌入和 8 个注意力头，层数设置为 2。字符序列的最大长度为 25，但是考虑到额外的 [B] 或者 [E] 令牌，实验中设置为 26。在训练过程中，识别字符数量为 94，包括大小写字母、数字和标点字符。在推理过程中，识别字符数量为 36，只包括大小写字母和数字。实验的迭代细化轮次为 1。

模型训练和测试：模型在两张 3090 的 GPU 上进行训练，batch size 设置为 384，训练 20 个 epoch。采用 Adam 优化器和 1cycle 学习率调度器训练，当训练到 15 个 epoch 时，学习调度器更换为 SWA。模型测试在单张 GPU 上进行，采用迭代细化一次的 AR 解码方案。

3.3 消融实验

为了验证 FIM 和 DPLM 模块对实验的影响，在实验过程中，依次删除各个模块。当删除 FIM 时，编码器提取的视觉特征直接输入解码器中；当删除 DPLM 时，特征序列采用 PLM 生成排列。实验结果如表 3 所示，本文所提出的 FIM 和 DPLM 模块对实验提升都是有效的。并且，当网络模型中同时具有这两个模块时，文本识别精度最高。

表 3 FIM 和 DPLM 的消融实验结果

FIM	DPLM	平均精度 /%（总数 7248）	平均精度 /%（总数 7672）
×	×	96.37	95.88
√	×	96.50	96.05
×	√	96.48	96.02
√	√	96.65	96.19

3.5 对比实验

本研究与其它较为先进的 5 种识别算法进行对比，在表 4 中显示了在多个基准数据集上的实验结果，表格中加粗为本列数据中的最优数据，下划线为本列数据中的次优数据。

表 4 不同基准数据集上的识别精度对比

方法	IIIT5K	SVT	IC13_857	IC13_1015	IC15_1811	IC15_2077	SVTP	CUTE80
ViTSTR-S	98.1	95.8	97.6	97.7	88.4	87.1	91.4	96.1
CRNN	94.6	90.7	94.1	94.5	82.0	78.5	80.6	89.1
TRBA	98.6	97.0	97.6	97.6	89.8	88.7	93.7	97.7
ABINet	98.6	97.8	<u>98.0</u>	97.8	90.2	88.5	93.9	97.7
PARSeq	<u>99.1</u>	<u>98.3</u>	98.3	98.3	<u>90.3</u>	<u>89.1</u>	<u>95.2</u>	98.6
本文	99.3	98.8	98.3	<u>98.2</u>	90.7	89.7	96.3	<u>98.3</u>

由表 4 可知，本文所提算法在 IIIT5K、SVT、IC13_857、IC13_857、IC15_1811、IC15_2077 和 SVTP 数据集上均达到最优结果，特别是在 SVT、IC15_2077 和 SVTP 数据集上，识别精度分别提升了 0.5%、0.6% 和 1.1%，在 IC13_1015 和 CUTE80 数据集上达到了次优识别效果。图 4 展示了本文所提算法的部分识别结果。图片下方第一排为文本标签，第二排为预测结果，加粗位置表示错误的预测字符。

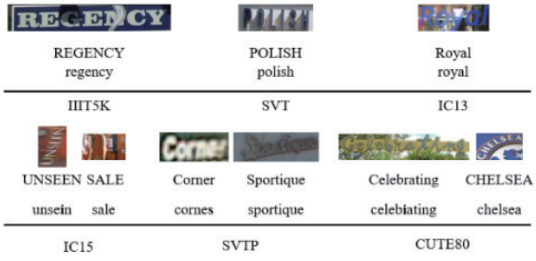

图 4 场景文本识别数据集识别结果

4 结语

本文提出了一种基于动态排列自回归的场景文本识别网络。在编码器和解码器网络结构的基础上，通过双分支设计的特征交互模块增强视觉特征。同时，采用 DPLM 动态生成排列，使得网络在不规则文本数据集上获得更好的泛化能力。实验结果表明，本文提出的算法在 6 个基准数据集上的性能优于目前大多数算法。此外，本文计算探索实现更高效的文本识别算法，以满足更复杂的实际应用需求。

参考文献：

[1]SHI B, YANG M, WANG X, et al. Aster: an attentional scene text recognizer with flexible rectification[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(9): 2035-2048.

[2]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].(2017-06-12)[2024-02-01].https://doi.org/10.48550/arXiv.1706.03762.

[3]BLEEKER M, DE RIJKE M. Bidirectional scene text recognition with a single decoder[C]//24th European Conference on Artificial Intelligence,Part 4.Amsterdam:IOS Press, 2020:2664-2671.

[4]YU D, LI X, ZHANG C, et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition.Piscataway:IEEE,2020:12110-12119.

[5]FANG S, XIE H, WANG Y, et al. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2021: 7098-7107.

[6]BAUTISTA D, ATIENZA R. Scene text recognition with permuted autoregressive sequence models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 178-196.

[7]YANG Z, DAI Z, YANG Y, et al. Xlnet: generalized autoregressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems 32,Volume 8 of 20.Red Hook:Curran Associates, 2019: 5730-5740.

[8]BAEK J, MATSUI Y, AIZAWA K. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition,[v.1]. Piscataway:IEEE,2021: 3112-3121.

[9]HOU Z, YU B S, WANG C Y, et al. Batchformerv2: exploring sample relationships for dense representation learning[EB/OL]. (2020-10-22)[2024-02-01].https://doi.org/10.48550/arXiv.2204.01254.

[10]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, 2016, 39(11): 2298-2304.

[11]SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]//29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2016: 4168-4176.

[12]LUO C, JIN L, SUN Z. Moran: a multi-object rectified attention network for scene text recognition[J]. Pattern recognition, 2019, 90: 109-118.

[13] 陈瑛 , 陈平平 , 林志坚 . 基于层次自注意力的高效场景文本识别 [J]. 无线电工程 ,2022,52(1):70-75.

[14]LIAO M, ZHANG J, WAN Z, et al. Scene text recognition from two-dimensional perspective[J].Proceedings of the AAAI conference on artificial intelligence,2019, 33(1): 8714-8721.

[14]WAN Z, HE M, CHEN H, et al. Textscanner: reading characters in order for robust scene text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto:AAAI,2020: 12120-12127.

[15]DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2024-03-01]. https://doi.org/10.48550/arXiv.2010.11929.

[16]ATIENZA R. Vision transformer for fast and efficient scene text recognition[C]//International Conference on Document Analysis and Recognition. Cham: Springer International Publishing, 2021: 319-334.

[17]QIAO Z, ZHOU Y, YANG D, et al. Seed: semantics enhanced encoder-decoder framework for scene text recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition,[v.1].Piscataway:IEEE, 2020: 13525-13534.

[18]WANG Y, XIE H, FANG S, et al. From two to one: a new scene text recognizer with visual language modeling network[C]//2021 IEEE/CVF International Conference on Computer Vision,[v.1].Piscataway:IEEE,2021: 14174-14183.

[19]NA B, KIM Y, PARK S. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 446-463.

[20]KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL].(2014-12-22)[2024-03-01].https://doi.org/10.48550/arXiv.1412.6980.

【作者简介】

王嘉宝（1999—），男，福建泉州人，硕士研究生，研究方向：计算机视觉、场景文本识别。

陈宏辉（1998—），男，福建南平人，博士研究生，研究方向：计算机视觉、场景文本识别。

陈平平（1986—），男，福建泉州人，教授，博士生导师，旗山学者，研究方向：机器学习、5G 通信、智能信息等数据传输分析及应用。

（收稿日期：2024-03-05）