计算机科学
COMPUTER SCIENCE

基于多层次嵌套Transformer的船名识别网络

王腾, 冼允廷, 徐浩, 谢宋祺, 邹全义

引用本文

王腾, 冼允廷, 徐浩, 谢宋祺, 邹全义. 基于多层次嵌套Transformer的船名识别网络[J]. 计算机科学, 2025, 52(6): 179-186.

WANG Teng, XIAN Yunting, XU Hao, XIE Songqi, ZOU Quanyi. Ship License Plate Recognition Network Based on Pyramid Transformer in Transformer [J]. Computer Science, 2025, 52(6): 179-186.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500064

基于多层次嵌套 Transformer 的船名识别网络

王 腾[1] 冼允廷[1] 徐 浩[1] 谢宋祺[1] 邹全义[2]
1 华南理工大学计算机科学与工程学院 广州 510006
2 华南理工大学新闻与传播学院 广州 510006
(cswangteng@mail.scut.edu.cn)

摘 要 船舶身份识别在水上目标监管中具有重要意义和广泛应用。船名是船舶身份识别的重要组成部分,准确识别船名可以弥补传统 AIS 身份识别方法的不足,提高船舶身份识别的准确率。与传统的中文文本识别相比,水上环境复杂,光照变化大,船体受腐蚀严重,船名字体不规范,导致船名图像存在清晰度低、文字残缺、字体样式不一致等问题,进而使船名识别困难且准确率低。文中设计了一种基于多层次嵌套 Transformer 的轻量级识别网络,以解决船名识别中存在的问题。首先,通过空间变换网络对输入图片进行处理,纠正船名倾斜的情况;然后利用嵌套 Transformer 有效提取图像的多粒度特征;最后对文字和部首进行不同尺度的识别。实验结果显示,相比其他文字识别方法,所提算法在船名识别中表现优异;在 CSLD 数据集上,准确率达到了 92.68%;在 SCSLD 数据集上,准确率达到了 94.50%;在 DCSLD 数据集上,准确率达到了 66.34%;同时,该方法具有低参数量和高帧率的特点。

关键词:中文文本识别;船名识别;深度学习;场景文本识别;Transformer
中图分类号 TP183

Ship License Plate Recognition Network Based on Pyramid Transformer in Transformer

WANG Teng[1],XIAN Yunting[1],XU Hao[1],XIE Songqi[1] and ZOU Quanyi[2]
1 School of Computer Science and Engineering,South China University of Technology,Guangzhou 510006,China
2 School of Journalism and Communication,South China University of Technology,Guangzhou 510006,China

Abstract Ship identification is of great significance and widely used in the regulation of waterborne targets. As one of the important components of ship identification,accurate identification of ship name can make up for the shortcomings of traditional AIS identification methods and improve the accuracy of ship identification. Compared with the traditional Chinese text recognition,due to the complex water environment,large changes in light,serious corrosion of ship hulls,and non-standardized ship names,ship name images have low clarity,text mutilation,inconsistent font styles and other problems,which make ship name recognition difficult and low accuracy. In this paper,a lightweight recognition network based on Pyramid Transformer in Transformer is proposed to solve the problems in ship name recognition. Firstly,the input image is processed by a spatial transform network to correct the tilt of the ship name. Then,the Transformer in Transformer module is utilized to efficiently extract the multi-granularity features of the image. Finally,the text and radical are recognized at different scales. Experimental results show that the proposed algorithm has excellent performance in ship name recognition compared with other text recognition methods. The accuracy reaches 92.68% on CSLD dataset,94.50% on SCSLD dataset,and 66.34% on DCSLD dataset. At the same time,this method is characterized by a low number of parameters and a high frame rate.

Keywords Chinese text recognition,Ship license plate recognition,Deep learning,Scene text recognition,Transformer

1 引言

随着深度学习技术的快速发展,水上交通的监控与管理也朝着信息化、智能化的方向迈进,例如图像处理技术和视频跟踪技术在航运领域的应用。船名作为船舶身份标识之一,具有唯一性的特征。因此,通过深度学习技术对船名进行有效识别,可以确定船舶的身份。

中文船名通常由中文、数字和字母组成,然而所处水上环境复杂、岸边抓拍距离远、抓拍角度固定、船舶文体腐蚀、船名字体不规范等不可避免因素,导致抓拍到的船名图像存在模

到稿日期:2024-05-16 返修日期:2024-09-05
基金项目:广东省哲学社会科学规划项目(GD24YXW02);广东省高校青年创新人才类项目(2023KQNC005)
This work was supported by the Guangdong Philosophy and Social Science Foundation Regular Project(GD24YXW02) and Youth Innovative Talent Projects of Guangdong Universities(2023KQNC005).
通信作者:冼允廷(xianyt@scut.edu.cn)

糊、倾斜、字体不完整、字体样式多等问题。因此,可以将船名识别看作更复杂、困难的中文文本识别任务(Chinese Text Recognition,CTR)。目前很多文字识别都聚焦于英文识别任务,中文识别的工作相对较少。许多中文识别的具体应用,都是将英语识别的网络模型进行简单的迁移。然而,这种简单迁移的做法难以达到如英语识别般的预期效果,原因在于中文与英文之间存在较大差异,中文字符内部结构更加复杂,且字符类别更多,根据 GB18030-2005 标准,中文字符共有 70244 个,而英文字符仅有 26 个。

中文文本识别在许多领域都有广泛应用,如自动驾驶、图片检索等。在深度学习方法兴起之前,对中文文本识别的方法主要是基于形态学方法[1-3]来对中文字符做手工特征提取,最后利用特征对字符进行分类,实现中文文本识别。随着深度学习的快速发展,越来越多的网络模型被应用到文字识别领域中。在中文识别领域,Yu 等[4]提出了一个中文文本数据集,包括文档数据集、场景数据集、网站信息数据集,并选择多个经典网络模型进行对比,包括 CRNN[5],ASTER[6],MASTER[7],ABINet[8]等。实验结果表明,许多在英文数据集上达到 Sota(state-of-the-art)的网络模型,在中文数据集上却表现不佳。

中英文之间存在较大差异,主要在于中文字符结构复杂多变和中文字符个数远多于英文字符个数这两方面。这导致中文字符识别难度大,且在实际应用中容易出现"zero-shot"或"few-shot"的问题(即在训练时部分中文字符没有出现或出现次数较少)。针对上述问题,许多研究开始使用更细粒度的识别方法来识别中文字符,如在基于首部识别的方法中,Wang 等[9]提出具有紧密连接结构的中文字符部首分析网络 DenseRAN 来分析汉字部首及其二维结构,有效提高了手写汉字识别的准确率;Wang 等[10]提出了一种部首聚合网络 RAN,实验证明该方法能够有效地识别网络未学习过的手写汉字;Deng 等[11]提出了 RRecT 网络,利用中文字符部首分解作为辅助监督信号,并在中文文本数据集[4]上证明了该方法的有效性;Cao 等[12]提出了一种层次分解嵌入的方法,将中文字符部首树的布局编码为语义向量,利用该向量来学习字符的基本结构和部首信息,有效解决了识别过程中出现的"zero-shot"问题。在基于笔画识别的方法中,Chen 等[13]提出了一种基于笔画的方法,将每个字符分解成一系列笔画,并采用基于匹配的策略,将预测的行程序列转换为一个特定的字符,解决了汉字识别"zero-shot"的问题;Liu 等[14]提出 SSDCNN 网络,它利用笔画序列信息和汉字的 8 个方向特征进行汉字识别(OLHCCR),有效提高了手写汉字识别的精度。也有研究尝试结合大模型来解决中文字符个数多的问题,例如 Yu 等[15]提出预先训练一个类似 Clip 的模型,获得每个汉字的规范表示,利用预训练得到的模型对 CTR 网络进行监督,最终实现网络能在不进行微调的情况下具有识别文本图像中的"zero-shot"的中文字符的能力。

目前,针对船名识别类别的研究相对较少。一方面是因为缺乏相应的数据集;另一方面,船名识别主要应用于特定的船舶监管等场景,因此其扩展性相对较低。Liu 等[16]针对船名图像倾斜问题,提出了一种倾斜校正方法;Liu 等[17]结合了 CRNN 与 STN 方法,提出了一种带矫正网络的船名识别算法;此外,Liu 等[18]还提出了 QSLPD 和 RTRNet 网络,用于实现对船名进行检测、矫正和识别。然而,上述 3 种方法只针对船名倾斜问题进行了改进,未考虑其他复杂情况,且在计算上耗时较长,准确率较低。另外,Zhang 等[19]使用 SSD 方法来检测船舶和船名的位置,并结合船名图像特征和 AIS(Automatic Identification System)信息,通过分类方法得出最终结果。然而,这种方法依赖于船舶 AIS 信息,且使用分类器进行船名分类,存在拓展性差等问题。最后,Zhou 等[20]提出了基于多模态和多注意力的融合网络 M3ANET 用于船名识别,但整体模型较为复杂,识别速度慢。

综上所述,在船名图像环境复杂,以及实际应用中计算资源受限等情况下,识别算法的鲁棒性、实时性面临挑战。同时,船名中的中文字符数量庞大、字体和样式多样等问题也增加了识别难度。为克服这些困难,需要研究改进船名识别技术,包括构建更大规模、更具代表性的船名数据集,设计更加鲁棒、轻量化的深度学习网络。

为了深入研究船名识别算法,本文收集并标注了在不同环境、角度和光照条件下的船名图像。这些图像涵盖了各种船舶类型和不同地区的船名。根据这些船名图像,本文创建了多个船名数据集,包括 CSLD(Chinese Ship Licence Dataset),SCSLD(Simple Chinese Ship Licence Dataset) 和 DCSLD(Difficult Chinese Ship Licence Dataset)。

另外,本文提出了一种基于多层次嵌套 Transformer 的轻量级识别网络,该网络整体包括预处理的空间变化网络、主干网络和多个预测头网络。空间变化网络的主要目标是通过学习一个几何变换来提升神经网络对输入图像的处理能力;主干网络分为 3 个阶段,通过采用金字塔结构逐步压缩特征信息,使得网络在减少计算量的同时具有多尺度信息融合、多层次特征表示的作用[21],每个阶段通过堆叠多个嵌套 Transformer[22],每一个嵌套 Transformer 包含两个不同粒度的自注意力模块,实现多粒度特征提取;最后,网络包含 3 个预测头,分别为字符个数预测头(Word Count Predict Head,WCPH)、中文字符预测头(Chinese Character Predict Head,CCPH)和中文部首预测头(Chinese Radical Predict Head,CRPH)。

本文的主要贡献如下:

1)构建了多个船名数据集,包括 CSLD,SCSLD 和 DCSLD,涵盖了不同环境、角度、光照条件以及各种船舶类型和地区的船名图像,为深入研究船名识别算法提供了数据基础。

2)针对船名识别的特点,设计了创新性的网络架构,引入嵌套 Transformer、空间变化网络、多任务预测头等结构,探索解决复杂场景下船名识别的挑战性问题。

2 基于多层次嵌套 Transformer 的船名识别网络

本文提出了一种基于多层次嵌套 Transformer 的船名识别网络,整体网络模型结构如图 1 所示,假定输入图像大小为 $H\times W\times 3$,首先将图像输入到空间变化网络(STN)中,STN 网络用于对输入图像进行几何变换,以便更好地适应文字的

形状和方向。矫正后的输入图片大小仍为 $H\times W\times 3$，随后通过图像分块嵌入向量网络（Patch embedding）将图像划分为 $\frac{H}{S}\times\frac{W}{S}$ 个嵌入向量，维度为 D_0。然后，将嵌入向量输入到主干网络中，主干网络分为 3 个阶段，每一个阶段堆叠多个嵌套 Transformer 模块，通过多尺度自注意力机制，网络既能捕捉局部模式，又能捕捉全局依赖关系。最终将 class token 输入到字符个数预测头（WCPH）预测目标字符个数，将外部 Transformer 的特征图输入到中文字符预测头（CCPH）预测最终结果，而内部 Transformer 和外部 Transformer 的特征图被输入到中文部首预测头（CRPH）用于预测字符部首序列。

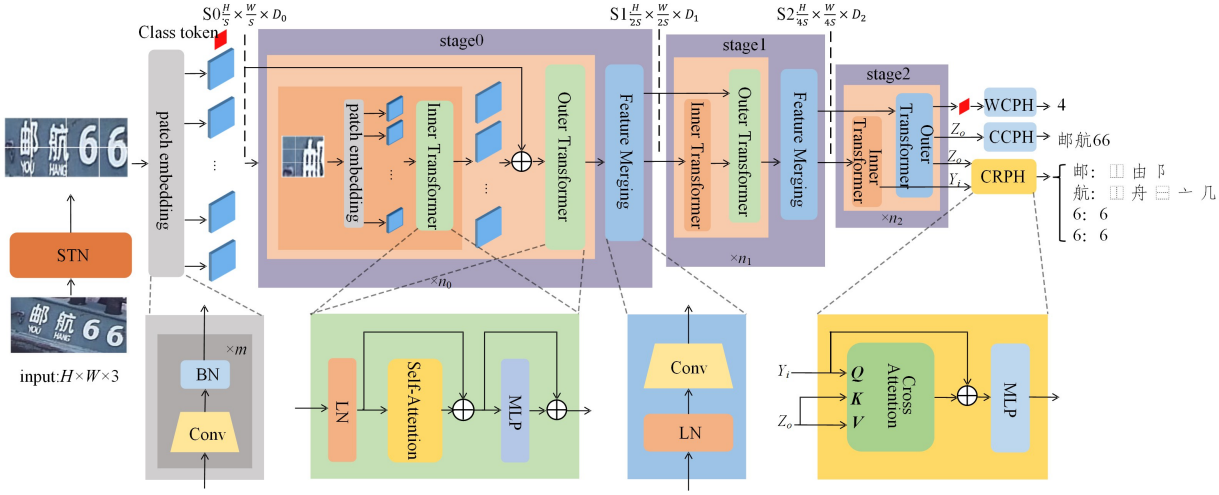


图 1　整体网络结构

Fig. 1　Overall network architecture

2.1　空间变化网络

在文字识别任务中，输入图像可能存在倾斜、扭曲或形变等问题，空间变化网络的作用可以视为通过网络自动学习如何对输入图像进行平移、缩放和旋转等几何变换，实现对文字图像进行校正、裁剪和形变处理，以适应不同的识别场景。

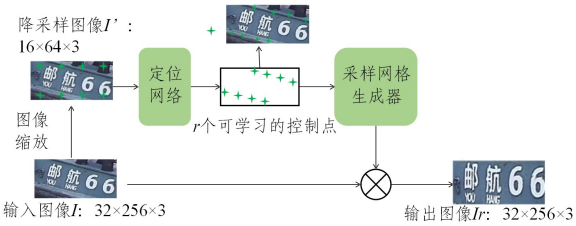


图 2　空间变化网络

Fig. 2　Spatial transformer network

本文使用的空间变化网络[6]如图 2 所示，首先对图像进行降采样，其次在降采样图像边缘初始化 r 个控制点，在训练和推理阶段通过卷积神经网络得到 r 个变化后的控制点，通过变化前后控制点的对应关系，利用薄板样条插值算法（Thin Plate Spline，TPS）计算出变化前后图像像素点的对应关系，最终得到变换后的图像。具体原理如下。

首先将文本图像大小调整为 16×64，得到图像 I'，定位网络 F 将在降采样的图像上进行计算，这样可以减少预测所需的参数量。随后在图像边缘初始化 r 个可学习的控制点 $\boldsymbol{H}$，将文本图像通过定位网络计算得到目标图像的控制点 $\boldsymbol{H}'$。

$$\boldsymbol{H}'=F(I') \tag{1}$$

$$\boldsymbol{H}'=[h_1',h_2',\cdots,h_i',\cdots,h_P']\in R^{2\times r} \tag{2}$$

其中，i 表示第 i 个控制点，且 $h_i'=[x_i',y_i']^{\mathrm{T}}$，$x_i'$ 和 y_i' 表示其归一化坐标位置。$F(\cdot)$ 为定位网络，由多个卷积层组成，卷积层之间为最大池化层，输出层是一个输出大小为 $2\times r$ 的全连接层，其中 r 代表控制点个数。

采样网络生成器利用薄板样条插值算法（Thin Plate Spline，TPS）计算变化前后图像像素之间的对应关系，其中变化矩阵为 $\boldsymbol{W}$：

$$\boldsymbol{W}=\begin{bmatrix} w_{1x} & w_{2x} & \cdots & w_{rx} & a_{1x} & a_{2x} & a_{3x} \\ w_{1y} & w_{2y} & \cdots & w_{ry} & a_{1y} & a_{2y} & a_{3y} \end{bmatrix}^{\mathrm{T}} \tag{3}$$

根据变化前控制点与变化后控制点，得到以下关系：

$$\begin{bmatrix} \boldsymbol{H}' \\ 0 \end{bmatrix}=\boldsymbol{L}\times\boldsymbol{W}=\begin{bmatrix} \boldsymbol{K} & \boldsymbol{Y} \\ \boldsymbol{Y}^{\mathrm{T}} & 0 \end{bmatrix}\times\boldsymbol{W} \tag{4}$$

$$\boldsymbol{Y}=[1\quad \boldsymbol{H}^{\mathrm{T}}] \tag{5}$$

$$\boldsymbol{K}=\begin{bmatrix} U(d_{11}) & \cdots & U(d_{1r}) \\ \vdots & \ddots & \vdots \\ U(d_{r1}) & \cdots & U(d_{rr}) \end{bmatrix} \tag{6}$$

其中，$U(\cdot)$ 为径向基核函数，d_{ij} 为坐标点之间的欧氏距离。

$$d_{ij}=\|(x_i,y_i)-(x_j,y_j)\| \tag{7}$$

$$U(x)=x^2\log(x) \tag{8}$$

通过求解 H 和 H' 之间对应的线性系统，计算得到 TPS 的变化矩阵 $\boldsymbol{W}$：

$$\boldsymbol{W}=\boldsymbol{L}^{-1}\times\begin{bmatrix} \boldsymbol{H}' \\ 0 \end{bmatrix} \tag{9}$$

将变换矩阵 $\boldsymbol{W}$ 应用于 I_r 中的每个像素位置，得到采样网格 $P=\{p_i\}$，其中 $p_i=[x_i,y_i]$ 表示变化后图像 I_r 上的像素点。

$$\boldsymbol{P}'=[\boldsymbol{K}'\quad 1\quad \boldsymbol{P}]\times\boldsymbol{W} \tag{10}$$

$$\boldsymbol{K}'=\begin{bmatrix} U(r_{11}) & \cdots & U(r_{1r}) \\ \vdots & \ddots & \vdots \\ U(r_{n1}) & \cdots & U(r_{nr}) \end{bmatrix} \tag{11}$$

其中，$\boldsymbol{P}'=\{p_i'\}$，$p_i'=[x_i',y_i']$ 表示变化前图像 I 上的像素点，n 为总像素个数。

由此可以得到变化后图像的像素点对应于变化前图像的

像素点的位置。最后对采样结果进行插值和裁剪，得到最终变化图像I_r。

2.2 嵌套 transformer 模块

嵌套 Transformer 模块如图 3 所示，包括内部 Transformer 和外部 Transformer。外部特征图被进一步划分为内部特征图，加上内部位置编码后输入到内部 Transformer 中，再将计算结果加到外部特征图上。外部特征图和 class token 加上对应的外部位置编码后输入外部 Transformer，得到下一层特征图。嵌套 Transformer 的核心思想是通过在每 Transformer 块内部嵌套另一个 Transformer 模块，使模型能够更好地学习到不同粒度的语义特征。内层 Transformer 块可以捕捉局部细粒度的特征，而外层 Transformer 块则关注全局的语义信息。这种层次化的特征表示有助于提高模型在图像处理任务上的性能。具体原理如下。

首先假设有一张输入图片 $\boldsymbol{X}$，将其划分为 N 个图像分块，如式(12)所示：

$$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_i,\cdots,\boldsymbol{X}_N]\in R^{N\times S\times S\times 3} \tag{12}$$

其中，i 表示第 i 个外部图像分块，S 表示采样尺度大小。传统的 Transformation[23] 模块只是处理图像分块序列之间的关系，忽略了每一个图像分块内部的局部结构。本文使用的嵌套 Transformer 模块包括内部 Transformer(Inner Transformer)和外部 Transformer(outer Transformer)，同时关注图像的全局和局部特征。在嵌套 Transformer 中，针对每一个图像分块继续采样并划分为 M 个更小的内部图像分块，如式(13)所示：

$$\boldsymbol{X}_i=[x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^M]\in R^{M\times s\times s\times 3} \tag{13}$$

其中，j 表示第 j 个图像内部分块，s 表示采样尺度大小。随后，将每一个图像分块转化为嵌入向量，如式(14)所示：

$$\boldsymbol{Y}_i^0=[E(x_i^1),E(x_i^2),\cdots,E(x_i^j),\cdots,E(x_i^M)]+E_i \tag{14}$$

$$\boldsymbol{Y}_i^0=[y_i^1,y_i^2,\cdots,y_i^j,\cdots,y_i^M]\in R^{M\times s\times s\times d_0} \tag{15}$$

其中，$E(\cdot)$表示嵌入向量网络，y_i^j 表示第 i 个外部图像分块中第 j 个内部图像分块所得到的嵌入向量，E_i 为内部图像分块的位置编码。

随后，使用内部 Transformer 挖掘和提取内部图像分块之间的关系：

$$\boldsymbol{Y}_i^{l\prime}=\boldsymbol{Y}_i^{l-1}+A(L(\boldsymbol{Y}_i^{l-1})) \tag{16}$$

$$\boldsymbol{Y}_i^l=\boldsymbol{Y}_i^{l\prime}+F(\boldsymbol{Y}_i^{l\prime}) \tag{17}$$

其中，$A(\cdot)$表示多头自注意力机制函数，$L(\cdot)$表示层归一化函数，$F(\cdot)$为全连接层，$l=1,2,\cdots,n$ 代表第 l 层。

所有内部图像分块根据式(16)和式(17)进行计算，最后得到内部图像方块的计算结果，如式(18)所示：

$$\boldsymbol{Y}^l=[\boldsymbol{Y}_1^l,\boldsymbol{Y}_2^l,\cdots,\boldsymbol{Y}_i^l,\cdots,\boldsymbol{Y}_N^l] \tag{18}$$

对于外部图像分块，先将其转化为指定维度的嵌入向量，再将所得到的内部图像分块的计算结果加到外部图像分块的嵌入向量中，随后使用外部 Transformer 捕捉全局依赖关系。

$$\boldsymbol{Z}^0=[E(X_1),E(X_2),\cdots,E(X_i),\cdots,E(X_N)]+E_o \tag{19}$$

$$\boldsymbol{Z}^0=[\boldsymbol{Z}_c^0,\boldsymbol{Z}_1^0,\boldsymbol{Z}_2^0,\cdots,\boldsymbol{Z}_i^0,\cdots,\boldsymbol{Z}_N^0] \tag{20}$$

$$\boldsymbol{Z}^l=\boldsymbol{Z}^l+F(\boldsymbol{Y}^l) \tag{21}$$

$$\boldsymbol{Z}^{l\prime}=\boldsymbol{Z}^{l-1}+F(L(\boldsymbol{Z}^{l-1})) \tag{22}$$

$$\boldsymbol{Z}^l=\boldsymbol{Z}^{l\prime}+F(\boldsymbol{Z}^{l\prime}) \tag{23}$$

其中，$E(\cdot)$表示嵌入向量网络；E_o为外部图像分块的位置编码；$F(\cdot)$为全连接层；$L(\cdot)$表示层归一化函数，使得内外图像分块的嵌入向量维度匹配。

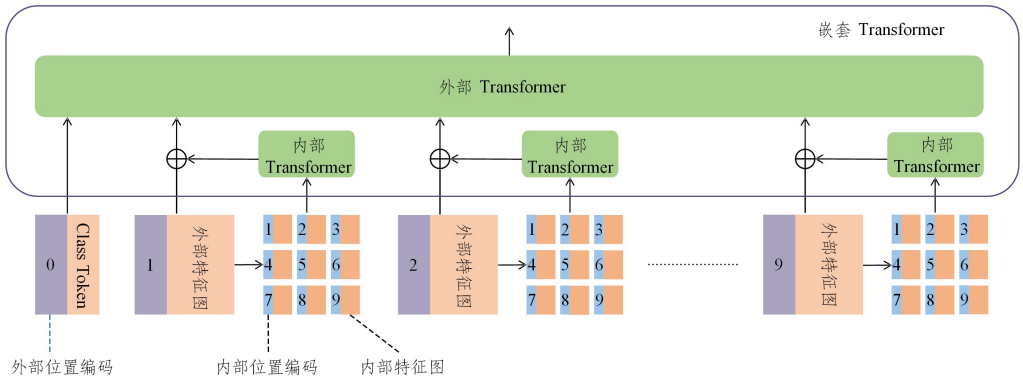


图 3 嵌套 Transformer 结构

Fig. 3 Nested Transformer structure

2.3 多任务预测头

对于每一个中文字符，可以将其划分为特定部首序列。在中文文本识别任务中，中文字符内部结构复杂，识别难度大，然而中文字符的构成又具有规律性，即大多数字符都由常见的部首组合而成，如图 4(a)所示，可以将中文字符分为 12 种基本字形结构[13]，每一个字都可以根据字形结构得到其部首序列，如图 4(b)所示。为此，本文设计了中文部首预测头(CRPH)，以引入更精细的监督，即部首，使网络能够捕捉部首感知特征，从而提高识别率。

在本文网络的主干网络中，存在两条数据流，其中一条为内部 Transformer，用于处理局部特征；另一条为外部 Transformer，用于捕获全局关系特征。相对应的中文字符也可以分为细粒度的部首特征和粗粒度的字符特征。利用嵌套 Transformer 结构，网络能够获取中文字符的多尺度特征，因此可以利用外部 Transformer 特征图预测字符结果，采用外部和内部 Transformer 特征图做交叉注意机制来捕获和融合相应的字符部首信息。

$$\boldsymbol{Z}_r=A(\boldsymbol{Y}_i,\boldsymbol{Z}_o,\boldsymbol{Z}_o)+\boldsymbol{Y}_i \tag{24}$$

$$P_r=F(\boldsymbol{Z}_r) \tag{25}$$

$$P_c=F(\boldsymbol{Z}_o) \tag{26}$$

其中，$\boldsymbol{Y}_i$ 为内部 Transformer 的特征图，$\boldsymbol{Z}_o$ 为外部 Transformer 的特征图，$A(\cdot)$ 表示多头自注意力机制函数，$F(\cdot)$ 为全连接层，P_r 为最终部首预测结果，P_c 为最终字符预测结果。

使用交叉熵损失函数计算字符与部首损失，如式(27)、式(28)所示：

$$L_c=-\sum_{I_i,c_i\in X}(\log(c_i\,|\,\boldsymbol{Z}_o)) \tag{27}$$

$$L_r=-\sum_{I_i,r_i\in X}(\log(r_i\,|\,\boldsymbol{Z}_r)) \tag{28}$$

其中，X 为训练集，I_i，c_i，r_i 分别为第 i 个图像以及该图像对应的真实字符序列标签与部首序列标签，Z_o 为外部特征图，Z_r 为部首特征图。

在主干网络中，class token 是一个可学习的参数向量，被添加到输入图像的 patch 序列中。通过与其他图像分块嵌入向量进行交互学习，能够聚合整个图像的全局特征，捕捉整个图像的整体语义信息。利用 class token 预测目标字符个数可以帮助网络模型学习字符的整体分布和排列规律，提供全局的参考信息。利用 class token 预测目标字符个数可以表示为：

$$P_w=F(\boldsymbol{Z}_c) \tag{29}$$

使用 L1 损失函数计算目标字符预测损失可以表示为：

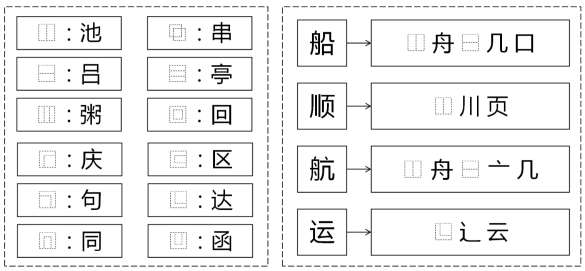
$$L_r=\sum_{I_i,w_i\in X}|w_i-P_i^w| \tag{30}$$

其中，X 为训练集，I_i 和 w_i 分别为第 i 个图像以及该图像真实字符个数标签。

因此，总体训练损失为：

$$L=\lambda_1 L_r+\lambda_2 L_c+L_w \tag{31}$$

其中，λ_1 和 λ_2 是平衡多任务的权衡系数。



(a)12 种基本中文字符结构　(b)中文字符转化为部首序列

图 4　中文字符与部首关系

Fig. 4　Relationship between Chinese characters and radicals

3　实验

3.1　数据集

本文收集了一个中国船名数据集(Chinese Ship Licence Dataset，CSLD)，其来源于真实采集的图像，通过安装摄像头，实地采集真实场景的船舶图像，并采用专门训练的船名检测网络，检测船舶图像的船名区域并截取出来。对于图像标签，采用预训练网络识别和人工矫正的方法对真实采集到的船舶照片的船名进行了标注。经过人工过滤和筛选，本数据集一共采集到 218086 张图像，将其按 8∶1∶1 的比例划分为训练集、验证集和测试集。

对于船名识别而言，识别结果的准确率与抓拍船名图像的质量密切相关。本文构建的数据集中，包括从真实场景中捕获的各种不同质量的图像，如图像清晰度低、船名腐蚀、光照过暗、文本部分遮挡等。为了探究不同网络模型在不同环境下的识别结果，本文进一步将数据划分为 SCSLD 数据集(Simple Chinese Ship Licence Dataset)和 DCSLD 数据集(Difficult Chinese Ship Licence Dataset)。SCSLD 数据集包含 193284 张图像，DCSLD 数据集包含 24802 张图像，同样按照 8∶1∶1 的比例划分为训练集、验证集和测试集。图 5 和图 6 分别给出了 SCSLD 和 DCSLD 数据集的部分图像。

图 5　SCSLD 数据集

Fig. 5　SCSLD dataset

图 6　DCSLD 数据集

Fig. 6　DCSLD dataset

3.2　实现细节

本文提出的方法是在 Ubuntu20.04 系统上利用 PyTorch 框架实现的，该系统配备英特尔 i7 10900K CPU、32GB 内存和 RTX 3080 GPU。对于所有数据集，图像在输入网络模型之前被调整到 32×256 大小。本文使用权重衰减为0.05 的 Adam 优化器进行训练，使用标准交叉熵作为损失函数，初始学习速率设置为 0.0001。同时，训练过程中使用了余弦学习速率调度器，一共训练 100 个 epoch。对于网络模型具体参数，其中外部 Transformer 采样尺度 $S=4$，内部 Transformer 采样尺度 $s=1,3$ 个阶段的嵌套 Transformer 个数、内部嵌入向量维度、外部嵌入向量维度，即 $[n_i,d_i,D_i]$ 分别为 [3,4,64]，[4,8,128]，[3,16,256]。

3.3　对比实验

将本文所提出的网络模型与其他经典的文字识别网络模型进行对比，包括 CRNN[5]，ASTER[6]，TransOCR[24]，SVTR[25]，VIPSR[26]，对比结果如表 1 所列。CRNN 采用 CNN 提取图形特征并使用 LSTM 对特征进行序列编码，最后使用 CTCloss[27] 解码出识别结果。ASTER 则在 CRNN 的基础上引入了空间变换器网络(STN)来处理文本的倾斜等不规则性。TransOCR 是基于 Transformer 的代表性方法之一，其采用 ResNet-34 作为编码器，自注意力模块作为解码器。SVTR 采用了 PVIT[21] 的整体架构，并加入了局部注意力机制和全局注意力机制，以实现字符多粒度特征的提取。VIPSR 在 SVTR 的基础上进行了改进，设计了一种新的排列方法，有效融合了局部和全局注意力模块的字符特征。

为了保证实验的公平性，所有网络的输入都调整为 32×256 大小，且所有模型均没有使用数据增强和预训练等其他策略。所有网络分别单独在 CSLD，SCSLD，DCSLD 数据集上进行训练。本文所收集的船名数据集涵盖了长江、珠江及沿海流域近 100 多个水上卡口点位的船舶数据，其中部分船名图像存在重复，导致验证集和测试集中小部分数据在训练集中重复出现。为了进一步验证模型的泛化性，对于本文提出的模型，本文分别在 CSLD，SCSLD，DCSLD 这 3 个数据集

中的训练集上抽样 40%和 60%的数据，并在相同的测试集上验证最终效果。

对于识别结果的评估，本文选择了在文本识别中两个主流的度量标准，即准确性（ACC）和标准化编辑距离（NED），计算式如式(32)、式(33)所示：

$$ACC=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(P_i,\hat{P}_i) \tag{32}$$

$$NED=1-\frac{1}{N}\sum_{i=1}^{N}\frac{D(P_i,\hat{P}_i)}{m(l(P_i),l(\hat{P}_i))} \tag{33}$$

其中，$\mathbb{I}(\cdot)$为两个字符串是否相等的运算函数，$E(\cdot)$为两个字符串之间编辑距离的运算函数，$l(\cdot)$为字符串长度运算函数，$m(\cdot)$为最大值运算函数，N 为样本数量。

实验结果如表 1 所列，与现有的先进方法相比，本文所提出的方法在船名数据集上取得了最优的性能。在 DCSLD 数据集上，本文提出的方法的优势最为明显，其原因是嵌套 Transformer 能够全面感知多粒度的字符成分特征，能够有效应对复杂环境状况。与其他网络模型相比，本文提出的网络模型在 ACC 和 NED 度量标准中，总体上都展现出了领先优势。从在 40%和 60%训练集上进行训练的实验结果可以看出，两者精度在 DCSLD 数据集上均出现了较大幅度的下降，进一步表明了构建困难数据集对模型整体训练的重要性，能够有效提高模型的鲁棒性和泛化能力。图 7 和图 8 分别给出了 DCSLD 数据集和 SCSLD 数据集上多个网络识别结果及其错误案例。

表 1 现有方法的测试实验结果

Table 1 Results of test experiments for existing methods

	CSLD		SCSLD		DCSLD		Params	FPS (s^{-1})
	ACC	NED	ACC	NED	ACC	NED		
CRNN	87.27	0.972	89.84	0.978	50.20	0.881	13.4×10^6	602
ASTER	91.34	0.978	93.37	0.984	62.10	0.897	28.2×10^6	180
TransOCR	92.30	0.980	94.23	0.986	64.42	0.903	84.9×10^6	212
SVTR	90.98	0.978	93.19	0.983	59.17	0.891	7.03×10^6	**645**
VIPTR	89.78	0.975	91.76	0.981	61.16	0.900	$\mathbf{6.02\times10^6}$	344
Ours(40%)	89.05	0.972	91.47	0.979	54..07	0.873		
Ours(60%)	91.04	0.978	93.19	0.984	60.20	0.895	8.04×10^6	369
Ours	**92.68**	**0.983**	**94.50**	**0.987**	**66.34**	**0.917**		

测试图像							
真实标签	**骏浩568**	**浙长兴货358**	**皖瑞海0688**	**江苏推870**	**苏财荣机808**	**金沙机8999**	**润杨集99999**
CRNN	驰塘568	浙长兴货356	皖瑞_0688	江苏推8_0	苏财荣_808	金沙机8999	润杨集集99999
ASTER	强浩568	浙长兴货358	航瑞海0688	江苏_890	苏财荣航808	金沙机8999	润杨集99999
TransOCR	银塘568	浙长兴货358	航瑞河0688	江苏推870	苏财荣_808	金沙机8999	润杨集99999
SVTR	银浩568	浙长兴货356	航瑞海0688	江苏推810	苏财荣航808	金沙机8999	润杨集99999
VIPTR	骏浩568	浙长兴货358	皖瑞海0688	江苏推810	苏财荣机808	金沙机9999	润杨集99999
Ours	骏浩568	浙长兴货358	皖瑞海0688	江苏推810	苏财荣机808	金沙机8999	润杨集99999

图 7 DCSLD 识别结果

Fig. 7 DCSLD recognition results

测试图像							
真实标签	**皖金舵678**	**苏云50**	**振航集007**	**奇盛运通7**	**泰顺机108**	**皖湾讹货3869**	**皖阜南货1222**
CRNN	皖金舵678	苏云50	振机集007	奇盛运通1	泰顺机_08	皖湾止货3869	皖阜南货1222
ASTER	皖金舵678	苏云50	振机集007	奇盛运通1	泰顺机_08	皖湾池货3869	皖阜南货1222
TransOCR	皖金舵678	苏云50	振航集007	奇盛运通_	泰顺机108	皖湾池货3869	皖阜南货1222
SVTR	皖金舵678	苏云50	振航集007	奇盛运通7	泰顺机108	皖湾讹货3869	皖阜南货1222
VIPTR	皖金舵678	苏云50	振航集007	奇盛运通7	泰顺机108	皖湾讹货3869	皖阜南货_222
Ours	皖金舵678	苏云50	振航集007	奇盛运通7	泰顺机108	皖湾讹货3869	皖阜南货1222

图 8 SCSLD 识别结果

Fig. 8 SCSLD recognition results

针对 DCSLD 数据集，可以将其分为 3 种复杂环境场景，如图 9 所示，包括由于抓拍角度固定导致的图像中文本倾斜问题、由于水上环境复杂以及船牌腐蚀等问题导致抓拍到的图像中文本模糊问题，以及由于轮胎和绳索等物品的不正确摆放导致抓拍的图像中文字符被部分遮挡的现象。

为进一步探究不同模型在不同环境场景中的鲁棒性，将 DCSLD 数据集中的验证集手动划分为 3 种不同场景，即图像文本倾斜、图像文本模糊和图像文字部分遮挡。所有模型均使用在 CSLD 数据集上训练的结果，模拟在实际应用场景下的数据分布，并在 3 种复杂环境场景下对比不同模型之间的效果。

(a)图像文本倾斜

(b)图像文本模糊

(c)图像文字部分遮挡

图 9 复杂环境场景

Fig. 9 Complex environmental scenarios

实验结果如表 2 所列。对于图像文本倾斜问题，使用结合卷积网络结构和循环网络结构的 CRNN 模型将原始输入转换为一维特征，其空间感知能力明显弱于其他使用 Transformer 注意力机制结构的网络，但 ASTER 模型与本文所提出的模型通过引入空间变化网络，不仅能够有效地矫正文本，还能减少背景信息的干扰，模型性能有较大的提升。对于图像文本模糊与文字部分遮挡问题，本文所提出的模型通过嵌套 Transformer 结构与部首预测头，在文字遮挡或文字结构模糊不清时，通过 Transformer 提取字符与部首不同粒度的特征，引入中文字符部首序列的先验知识，增强特征表示，提高最终识别精度。

表 2　针对不同复杂环境场景的测试实验结果

Table 2　Results of test experiments for different complex environmental scenarios

methods	图像文本倾斜		图像文本模糊		图像文字部分遮挡	
	ACC	NED	ACC	NED	ACC	NED
CRNN	75.67	0.942	49.92	0.879	51.68	0.883
ASTER	82.56	0.952	51.32	0.882	57.51	0.891
TransOCR	85.62	0.958	56.21	0.887	61.23	0.894
SVTR	83.44	0.955	54.98	0.884	61.75	0.894
VIPTR	83.54	0.957	53.72	0.880	60.16	0.892
Ours	**88.52**	**0.973**	**57.21**	**0.872**	**65.72**	**0.913**

3.4　消融实验

为了明确每个设计的独立贡献以及各个模块的重要性，本文通过依次去除字符数量预测头（WCPH）、字符部首预测头（CRPH）、内部 Transformer 模块（Inner Transformer）和空间变化网络（STN）来进行消融实验。实验在 3 个数据集（CSLD，DCSLD，SCSLD）上进行。

从表 3 的消融实验结果可知：

1）在 CSLD 和 DCSLD 数据集上，去除 WCPH 后，结果有轻微下降，但在 DCSLD 数据集上略有上升。这表明 WCPH 模块通过学习文本图像的全局特征，可以提供对目标字符数量的预测。在识别过程中，网络可以根据 WCPH 的预测结果进行相应的调整和约束，从而提高识别的准确性。但是在复杂环境下，由于模糊和遮挡等原因，模型对目标字符的数量预测效果反而不佳。

2）在 CSLD 数据集和 SCSLD 数据集上，去除 CRPH 模块和 WCPH 模块后，ACC 分别下降了 0.93% 和 0.75%；而在 DCSLD 数据集上，可以观察到精度下降明显，ACC 下降了 3.46%。因此，可以得出结论：使用多任务联合训练策略可以捕获更全面的信息，提升识别准确率，尤其在复杂环境情况下（如部分遮挡、图像模糊等），通过 CRPH 引入字符部首的先验知识，可以有效增强特征表示，实现更细粒度的监督。

3）本文提出的嵌套 Transformer 模块包含内部 Transformer 模块和外部 Transformer 模块，分别用于提取不同尺度特征。在去除内部 Transformer 模块、CRPH 模块和 WCPH 模块后可以明显观察到，在 3 个数据集上 ACC 和 NED 结果都有较大程度的下降，这验证了在中文文本识别任务中，文本的字形和结构特征对于正确的识别起着重要作用，内部 Transformer 模块的存在能够有效地提取这些细粒度特征，并为后续的文本识别任务提供更丰富和准确的特征表示。

4）在去除空间变化网络后，3 个数据集上的精度都呈现下降趋势。如图 10 所示，在图像输入到主干网络前，通过空间变化网络实现对文字图像进行校正、裁剪和形变处理，可以有效改善图像中文字倾斜问题，同时减少无关背景信息的干扰，提高识别结果的准确性。

表 3　消融实验结果

Table 3　Ablation experiment results

WCPH	CRPH	Inner Transformer	STN	CSLD		SCSLD		DCSLD	
				ACC	NED	ACC	NED	ACC	NED
√	√	√	√	**92.68**	**0.983**	**94.50**	**0.987**	66.34	**0.917**
	√	√	√	92.57↓	0.982↓	94.38↓	**0.987**	**66.39**↑	**0.917**
		√	√	91.75↓	0.980↓	93.75↓	0.985↓	62.88↓	0.904↓
			√	90.70↓	0.977↓	92.90↓	0.983↓	58.89↓	0.891↓
				89.05↓	0.972↓	92.47↓	0.982↓	56.22↓	0.852↓

图 10　空间变化网络处理结果

Fig. 10　Results of spatial transformer network processing

结束语　本文提出了一种基于嵌套 Transformer 的中文船名识别网络，该网络能够提取多粒度字符特征，这些特征不仅可以描述字符的局部模式，还能够在全局尺度上描述不同空间特征间的依赖关系。同时，本文利用多任务学习的中文部首先验知识来训练一个更通用和健壮的视觉编码器。首先使用 STN 网络对输入图像进行预处理；然后利用主干网络提取多粒度特性；最后结合中文部首的先验知识，进行多任务训练，包括字符数预测、字符识别和部首预测。这种训练方式不仅提高了网络模型在部首识别方面的性能，也增强了视觉编码器的泛化能力，使其能够在面对各种复杂中文场景时保持较强的鲁棒性。在船名数据集上的实验结果和详细的消融分析显示，本文方法优于现有的经典方法，且具有易于训练、低参数量和高帧率等特点。

参 考 文 献

[1]　JIN L W，YIN J X，GAO X，et al. Study of Several directional

feature extraction methods with local elastic meshing technology for HCCR[C]// Proceedings of the 6th International Conference for Young Computer Scientist. Hong Kong: International Academic Publishers, World Publishing Corporation, 2001: 232-236.
[2] SU Y M, WANG J F. A novel stroke extraction method for Chinese characters using Gabor filters[J]. Pattern Recognition, 2003, 36(3): 635-647.
[3] CHANG F. Techniques for Solving the Large-Scale Classification Problem in Chinese Handwriting Recognition[M]. Berlin: Springer, 2008: 161-169.
[4] YU H, CHEN J, LI B, et al. Benchmarking Chinese Text Recognition: Datasets, Baselines, and an Empirical Study[J]. arXiv: 2112.15093, 2021.
[5] SHI B, 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.
[6] 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, 2019, 41(9): 2035-2048.
[7] LU N, YU W, QI X, et al. MASTER: Multi-aspect non-local network for scene text recognition[J]. Pattern Recognition, 2021, 117: 107980.
[8] 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. New York: IEEE Computer Society, 2021: 7098-7107.
[9] WANG W, ZHANG J, DU J, et al. DenseRAN for Offline Handwritten Chinese Character Recognition[C]// Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). New York: IEEE, 2018: 104-109.
[10] WANG T, XIE Z, LI Z, et al. Radical aggregation network for few-shot offline handwritten Chinese character recognition[J]. Pattern Recognition Letters, 2019, 125: 821-827.
[11] DENG X, HUANG Z, MA K, et al. RRecT: Chinese Text Recognition with Radical-Enhanced Recognition Transformer[C]// Proceedings of the International Conference on Artificial Neural Networks and Machine Learning - ICANN 2023. Berlin: Springer, 2023: 509-521.
[12] CAO Z, LU J, CUI S, et al. Zero-shot Handwritten Chinese Character Recognition with hierarchical decomposition embedding[J]. Pattern Recognition, 2020, 107: 107488.
[13] CHEN J, LI B, XUE X. Zero-shot Chinese character recognition with stroke-level decomposition[J]. arXiv: 2106.11613, 2021.
[14] LIU X, HU B, CHEN Q, et al. Stroke sequence-dependent deep convolutional neural network for online handwritten Chinese character recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(11): 4637-4648.
[15] YU H, WANG X, LI B, et al. Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 2023: 11909-11918.
[16] LIU B, ZHANG S, HONG Z, et al. A Horizontal Tilt Correction Method for Ship License Numbers Recognition[J]. Journal of Physics: Conference Series, 2018, 976(1): 012013.
[17] LIU D, CAO J, WANG T, et al. SLPR: A Deep Learning Based Chinese Ship License Plate Recognition Framework[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(12): 23831-23843.
[18] LIU B, WU S, ZHANG S, et al. Ship License Numbers Recognition Using Deep Neural Networks[J]. Journal of Physics: Conference Series, 2018, 1060(1): 012064.
[19] ZHANG W, SUN H, ZHOU J, et al. DCNN Based Real-Time Adaptive Ship License Plate Recognition (DRASLPR)[C]// Proceedings of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). New York: IEEE, 2018: 1829-1834.
[20] ZHOU C, LIU D, WANG T, et al. M3ANet: Multi-modal and multi-attention fusion network for ship license plate recognition [J]. IEEE Transactions on Multimedia, 2023, 26: 5976-5986.
[21] WANG W, XIE E, LI X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 548-558.
[22] HAN K, XIAO A, WU E, et al. Transformer in transformer [J]. Advances in neural information processing systems, 2021, 34: 15908-15919.
[23] DOSOVITSKIY A, BEYER L, KOLESBIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010.11929, 2020.
[24] CHEN J, LI B, XUE X. Scene Text Telescope: Text-Focused Scene Image Super-Resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2021: 12021-12030.
[25] DU Y, CHEN Z, JIA C, et al. SVTR: Scene Text Recognition with a Single Visual Model[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. New York: IEEE, 2022: 884-890.
[26] CHENG X, ZHOU W, LI X, et al. VIPTR: A Vision Permutable Extractor for Fast and Efficient Scene Text Recognition[J]. arXiv: 2401.10110, 2024.
[27] GRAVES A, FERNANDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// Proceedings of the 23rd International Conference on Machine Learning. New York: Association for Computing Machinery, 2006: 369-376.

WANG Teng, born in 2000, postgraduate. His main research interests include image processing and deep learning.

XIAN Yunting, born in 1982, Ph.D, lab master. His main research interests include artificial intelligence and image processing.

（责任编辑：何杨）