DOI:10.16136/j.joel.2023.11.0625

基于字符注意力的自然场景文本识别

熊 炜[1,2,3*],孙 鹏[1],赵 迪[1],刘 粤[1]

(1.湖北工业大学 电气与电子工程学院,湖北 武汉 430068;2.襄阳湖北工业大学产业研究院,湖北 襄阳 441003;3.美国南卡罗来纳大学 计算机科学与工程系,南卡 哥伦比亚 29201)

摘 要:自然场景文本识别中采用固定大小的卷积核提取视觉特征,后仅进行字符分类的方法,其全局建模能力弱且忽视了文本语义建模的重要性,因此,本文提出一种基于字符注意力的自然场景文本识别方法。首先构建不同于卷积网络的多级 efficient Swin Transformer 提取特征,其可使不同窗口的特征进行信息交互;其次设计了字符注意力模块(character attention module,CAM),使网络专注于字符区域的特征,以提取识别度更高的视觉特征;并设计语义推理模块(semantic reasoning module,SRM),根据字符的上下文信息对文本序列进行建模,获得语义特征来纠正不易区分或模糊的字符;最后融合视觉和语义特征,分类得到字符识别结果。实验结果表明,在规则文本数据集 IC13 上识别准确率达到了 95.2%,在不规则的弯曲文本数据集 CUTE 上达到了 85.8%,通过消融及对比实验证明了本文提出的方法可行。

关键词:Swin Transformer;字符注意力;语义推理;特征融合

中图分类号:TP391.1 **文献标识码**:A **文章编号**:1005-0086(2023)11-1158-10

Natural scene text recognition based on character attention

XIONG Wei[1,2,3*], SUN Peng[1], ZHAO Di[1], LIU Yue[1]

(1. School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, Hubei 430068, China; 2. Xiangyang Industrial Research Institute, Hubei University of Technology, Xiangyang, Hubei 441003, China; 3. Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201, USA)

Abstract: In natural scene text recognition, a fixed size convolution kernel is used to extract visual features, and then character classification is performed. The global modeling ability of this method is weak and it ignores the importance of text semantic modeling. Therefore, this paper proposes a natural scene text recognition method based on character attention. Firstly, a multi-level efficient Swin Transformer network is constructed to extract features, which is different from the convolutional network. This network can make the features of different windows interact with each other. Secondly, the character attention module (CAM) is designed to make the network focus on the features of the character region, so as to extract the visual features with higher recognition ability. Then, the semantic reasoning module (SRM) is designed to model the text sequence according to the context information of characters. And the module can obtain semantic features to correct the indistinguishable or fuzzy characters. At last, visual and semantic features are fused to get the results of character recognition. The experimental results show that the recognition accuracy in this paper reaches 95.2% on the regular text data set IC13 and 85.8% on the irregular curved text data set CUTE. The feasibility of the proposed method is proved by ablative and comparative experiments.

Key words: Swin Transformer; character attention; semantic reasoning; feature fusion

* **E-mail**:xw@mail.hbut.edu.cn

收稿日期:2022-09-06 **修订日期**:2022-12-26

基金项目:国家自然科学基金(61571182,61601177)、湖北省自然科学基金(2019CFB530)、湖北省科技厅重大专项(2019ZYYD020)和襄阳湖北工业大学产业研究院科研项目(XYYJ2022C05)和国家留学基金(201808420418)资助项目

0 引　言

随着生活越来越智能化，人类可以轻易地获取文本图像，并且希望快速而简单地得到其中蕴含的高层语义信息。比如，在自动驾驶领域需要提取交通标识中的限速信息、街道信息、信号灯时间；在票证识别领域需要提取身份编号、票据金额等关键信息；在图像分类中需要根据图像中的信息进行分类，用户根据关键字搜索得到相应的图片等等，因此文本识别变得尤为重要。对于场景单一、高质量的文本图像，采用传统的光学字符识别方式非常有效，其往往需要手工设计特征，再进行单字符的分割与识别。但在自然场景中，文本字体、颜色、大小、方向不一，且不具有规则，甚至图像清晰度低，传统方法已不能达到较高的识别率。

目前对于自然场景文本识别主要采用深度学习的方法，大体可划分为规则文本的识别以及不规则的文本识别。识别规则文本的方法主要有基于 CTC(connectionist temporal classification)、序列到序列的方法[1-4]。2017 年，文献[1]提出了 CRNN(convolution recurrent neural network)方法：其通过卷积层提取特征，对字符分类，后经双向长短期记忆(bi-directional long short term memory，BLSTM)网络转录识别。同年，文献[3]提出了 CNN(convolution neural network)结合 CTC 的方式，使用层叠的 CNN 代替 RNN，可并行计算并降低了复杂度。对于不规则文本的识别方法主要有：1) 校正方法，即通过校正模块将不规则文本转化为规则文本；2) 分割方法，即先获取文本图像中每个字符的位置信息，分割出单个字符视觉特征，接着对特征进行分类识别出字符，最后将所有的字符连接起来得到识别结果。2016 年，文献[5]提出了使用空间变换网络(spatial transformer networks，STN)将图像中不规则文本变换成水平方的文本，然后以规则文本的方式进行处理。2019 年文献[6]以及 2021 年文献[7]提出的方法中，也使用薄板样条插值(thin plate spline，TPS)将图像中不规则文本变换成水平方的文本。2021 年，文献[8]则提出了分割的识别方法 Mask TextSpotter，其通过 Faster R-CNN 做候选框的回归得到文本掩码，实现文本实例的分割。基于校正的方法往往引起字符的扭曲变形，导致字符识别错误；而基于分割的方法需要进行字符级的标注，并且性能容易受到字符分割结果的影响，分割错误会导致字符识别错误，甚至无法识别。如今，越来越多的研究员采用基于注意力的方法识别不规则文本避免校正和分割，2019 年，文献[9]提出 2D Attention 的方法通过多层 Transformer 提取二维空间上的特征序列；2020 年，文献[10]提出的 DAN(decoupled attention network)方法使用了一种解耦的注意力网络；2021 年，文献[11]提出的 ABINet(autonomous，bidirectional and iterative network)方法以及文献[12]提出的 CTC-Attention 方法使用了 Transformer 作为注意力。

在上述的识别方法中，规则的文本识别方法对于不规则文本识别正确率低，而通过校正或者分割的方法实现对不规则文本的识别，往往代价较大，并且忽视了文本序列中字符与字符之间的内在联系。针对以上问题，本文提出了一种基于注意力机制的场景文本识别方法，是一种无需校正或分割的方法，通过搭建多级 efficient Swin Transformer 网络提取丰富的全局特征，后经注意力机制得到视觉特征；其次通过语义推理模块(semantic reasoning module，SRM)并结合字符上下文信息，对文本序列进行建模得到语义特征；最后使用特征融合单元融合不同模态的视觉特征和语义特征，并分类得到最终的文本识别结果，有效地避免了分割字符的缺陷，同时考虑到了字符与字符之间的内在联系。

1　文本识别网络结构

本文提出的文本识别网络结构如图1所示，主要有3个部分：视觉模型(visual module)，语义模型

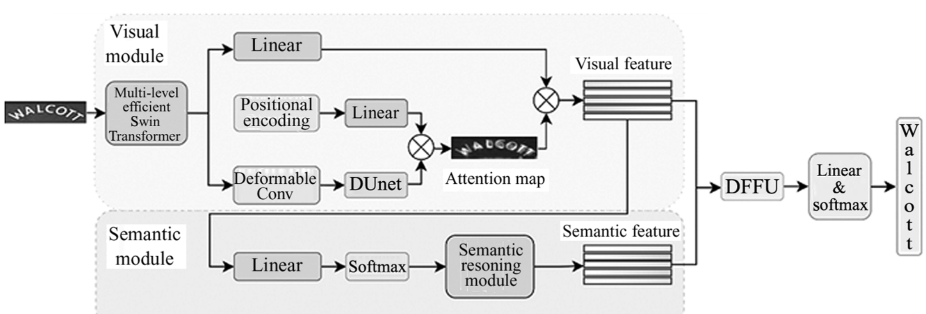

图 1　基于字符注意力的文本识别结构图

Fig. 1　The structure diagram of text recognition based on character attention

(semantic module)以及双特征融合单元(double features fusion unit,DFFU)。其中视觉模型由多级 efficient Swin Transformer 和字符注意力模块(character attention module,CAM)组成。

1.1　视觉模型

1.1.1　多级 efficient Swin Transformer

单纯的文本识别,输入图像是裁剪过后的图像,背景信息单一,且图中的每个字符来自于一个单词,因此全局信息对于单个字符的识别有着相关性;此外,普通 3×3 卷积或者空洞卷积(dilated convolution,DConv)的感受野都是固定大小的,无法实现远距离跨区域的信息交互。同时考虑到输入图像尺寸较小,多尺度的特征提取、融合操作反而容易丢失重要的细节信息,因此对于全局特征 F 的提取,本文选择了具有远程信息交互的 Swin Transformer[13] 以及具有通道信息交互的高效通道注意力机制(efficient channel attention, ECA)[14],设计了多级 efficient Swin Transformer 模块,如图 2 所示,Patch partition 层将输入图片划分为 $H/4\times W/4$ 个小块,每个小块由 4×4 个像素组成,Linear embedding 层将输入图像的通道维度变成预先设置好的值即 512,Swin Transformer block 提取输入图像空间上的细粒度特征;为了挖掘通道上的重要信息,在 Swin Transformer block 之后以残差的方式嵌入 ECA,使模型重点关注某些通道特征。

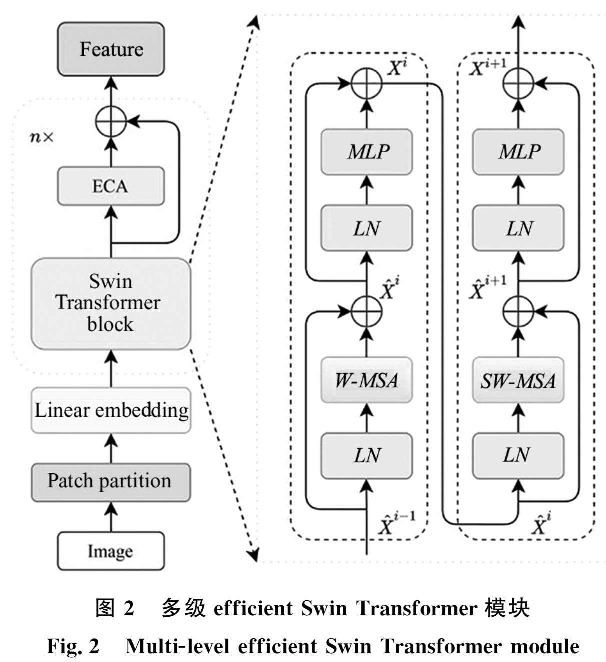

图 2　多级 efficient Swin Transformer 模块

Fig. 2　Multi-level efficient Swin Transformer module

Swin Transformer 是在 Transformer 的基础上进行改进的,它对输入图像进行窗口划分,通过移动窗口的方式,实现不同窗口区域间的远程信息交互,因此它不仅可以对全局信息进行建模,相比于 Transformer 具有更强特征提取能力。Swin Transformer 由层归一化(layer normalization,LN)、基于窗口的多头自注意力(window-based multi-head self-attention,$W\text{-}MSA$)、多层感知机(multi-layer perceptron,MLP)以及基于移动窗口多头自注意力(shifted window-based multi-head self-attention,$SW\text{-}MSA$)组成,结构如图 2 右侧所示。在 $W\text{-}MSA$、MLP 以及 $SW\text{-}MSA$ 模块前都添加了 LN 模块,组成 4 个小单元,每个小单元通过残差的方式连接。对于输入特征数据 $X\in\boldsymbol{R}^{C\times H\times W}$,其中 C、H 和 W 分别表示为特征图的通道数、高度和宽度,从第 i 层到第 $i+1$ 层的计算式如下[12]:

$$\hat{X}^i = W\text{-}MSA[LN(\hat{X}^{i-1})] + X^{i-1}, \tag{1}$$

$$X^i = MLP[LN(\hat{X}^i)] + \hat{X}^i, \tag{2}$$

$$\hat{X}^{i+1} = SW\text{-}MSA[LN(X^i)] + X^i, \tag{3}$$

$$X^{i+1} = MLP[LN(\hat{X}^{i+1})] + \hat{X}^{i+1}, \tag{4}$$

式中,i 表示层号,$\hat{X}^i$、X^i 分别表示第 i 层 $W\text{-}MSA$、MLP 输出的特征,$\hat{X}^{i+1}$、X^{i+1} 则分别表示第 $i+1$ 层 $SW\text{-}MSA$、MLP 输出的特征。

1.1.2　CAM

通过多级 efficient Swin Transformer 提取到的全局特征 $\boldsymbol{F}$,具有与字符无关的背景信息,因此有必要让网络模型将注意力聚焦在字符区域。此外,对于字符识别需要从全局特征 $\boldsymbol{F}$ 映射得到字符特征序列,再进行分类从而识别出字符,如图 3 所示,L 为预测的文本序列最大长度。

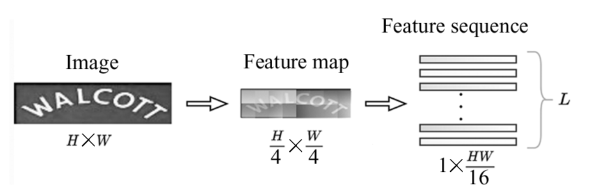

图 3　特征图映射示意图

Fig. 3　Schematic diagram of feature mapping

通常的做法是采用自注意力机制[15],它是机器翻译使用的一种基于查询的注意力机制,能很好地解决文本序列的预测问题,因此本文在自注意力机制的基础上针对文本识别进行了改进,设计了 CAM 让网络模型关注的焦点置于字符区域,从而提升模型的识别准确率。如图 4 所示,CAM 由 $\boldsymbol{V}$、$\boldsymbol{Q}$、$\boldsymbol{K}$ 3 个分支组成,$\boldsymbol{V}$ 由全局特征 $\boldsymbol{F}$ 线性变换得到。考虑到文本特征本身就是位置相关的,通过位置编码得到

$\boldsymbol{Q}$,可以让模型利用特征图中的位置信息,并且在不同文本中字符与字符的间距不同,位置编码后经过线性变换可增加模型的泛化能力。考虑到字符形状是不规则的,而可变形卷积(deformable convolution)[16]的感受野是个多边形,以残差的方式嵌入使得模型一定程度上可以根据字符的形状过滤不相关的背景特征;此外,由于输入图像中字符所占区域较大,所以在 Unet[12-17] 网络中使用空洞率为 2 的 DConv 替代普通卷积以扩大感受野,在编码解码部分使用 1×1 卷积,形成 DUnet(dilated Unet)网络,如图 4 右侧所示。使用语义分割中的 Unet 网络可以较好地处理字符区域与非字符区域之间的分界,因此,$\boldsymbol{K}$ 由全局特征 $\boldsymbol{F}$ 依次经过一个可变形卷积、DUnet 网络得到。

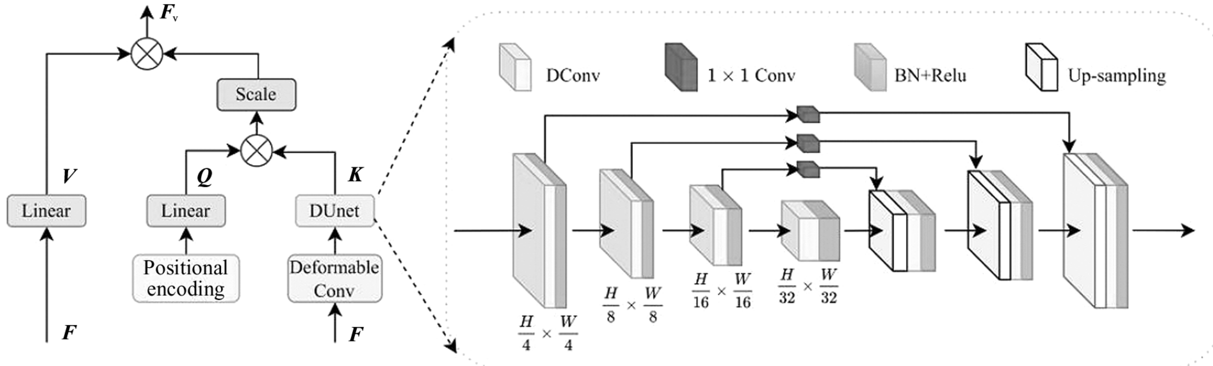

图 4 字符注意力模块结构图

Fig. 4 The structure diagram of character attention module

假设 $\hat{p}_k \in \{(-1,-1),(-1,0),(-1,1),\cdots,(1,1)\}$ 表示 3×3 可变形卷积核,$K=9$,$x(\hat{p})$、$y(\hat{p})$ 分别表示输入特征 $\boldsymbol{F}$ 在位置 $\hat{p}$ 的值、卷积映射后的值,则有[16]:

$$\mathrm{DConv}[x(\hat{p})]=y(\hat{p})=\sum_{k=1}^{K} w_k \cdot x(\hat{p}+\hat{p}_k+\Delta\hat{p}_k)\cdot\Delta m_k, \tag{5}$$

式中,$\Delta\hat{p}_k$、Δm_k 分别是第 k 个位置的可学习的偏移量和调制标量,且 $\Delta m_k \in [0,1]$。

位置编码矩阵 P 的定义为[15]:

$$\boldsymbol{P}(pos,d_i)=\begin{cases}\sin(pos/10\,000^{2k/d}) & \text{if} \quad d_i=2k\\ \cos(pos/10\,000^{2k/d}) & \text{if} \quad d_i=2k+1\end{cases}, \tag{6}$$

式中,pos 为字符在文本序列中的位置,d_i 为字符向量的位置,$2k$ 表示偶数,$2k+1$ 表示奇数。

骨干网络提取到的全局特征为 $\boldsymbol{F}\in \boldsymbol{R}^{C\times H\times W}$,则可得到 $\boldsymbol{V}$、$\boldsymbol{Q}$、$\boldsymbol{K}$:

$$\boldsymbol{V}(\boldsymbol{F})=\boldsymbol{F}\boldsymbol{W}_{\mathrm{v}}, \tag{7}$$

$$\boldsymbol{Q}=\boldsymbol{P}\boldsymbol{W}_{\mathrm{p}}, \tag{8}$$

$$\boldsymbol{K}(\boldsymbol{F})=\mathrm{DUnet}[\mathrm{DConv}(\boldsymbol{F})+\boldsymbol{F}], \tag{9}$$

式中,$\boldsymbol{W}_{\mathrm{v}}$、$\boldsymbol{W}_{\mathrm{p}}$ 是学习到的线性变换矩阵,DUnet($\cdot$)为 DUnet 网络,DConv($\cdot$)为可变形卷积。

若 $\boldsymbol{Q}$、$\boldsymbol{K}$ 相乘的结果过大,经过 softmax 函数后,训练时网络梯度将非常小,不利于反向传播,因此对其进行了尺度化操作,最终视觉特征 $\boldsymbol{F}_{\mathrm{v}}$ 由公式(10)计算得到:

$$\boldsymbol{F}_{\mathrm{v}}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_l}}\right)\boldsymbol{V}, \tag{10}$$

式中,d_l 为字符向量的长度。

1.2 SRM

通常人类对于句中不认识的单词,往往会考虑上下文信息进行猜测;对于缺省字符的单词往往会根据其前后的字符进行猜测。例如“instance”,当缺省字符“a”时,即“inst_nce”,人类可通过上文信息“inst”以及下文信息“nce”很容易地猜测出缺省的字符“a”,最后预测出单词“instance”。

基于人类对文本预测的模式,本文加入了 SRM[12] 对文本序列中的字符相关性进行建模,其结构如图 5 所示。

根据字符预测顺序设置 $L\times L$ 大小的字符掩码矩阵 $\boldsymbol{M}_{\mathrm{mask}}$:

$$\boldsymbol{M}_{\mathrm{mask}}[i,j]=\begin{cases}0, \text{if } i\neq j\\ -\infty, \text{if } i=j\end{cases}, \quad i,j\in[1,L]。 \tag{11}$$

视觉特征 $\boldsymbol{F}_{\mathrm{v}}\in \mathscr{R}^{C\times H\times W}$,经过 Linear 和 softmax 操作求得字符的概率特征 $\boldsymbol{F}_{\mathrm{p}}$,将 $\boldsymbol{F}_{\mathrm{p}}$ 和 $\boldsymbol{M}_{\mathrm{mask}}$ 作为多头自注意力的输入,得到特征 $\boldsymbol{F}_{\mathrm{m}}$,后经过 Add&normalize 层、Feed forward 层、Linear&softmax

层得到语义特征 $\boldsymbol{F}_s$，其中 $\boldsymbol{F}_p$、$\boldsymbol{F}_m$ 的计算如下：

$$\boldsymbol{F}_p = \mathrm{softmax}\,(\boldsymbol{F}_v\boldsymbol{W}_1), \tag{12}$$

$$\boldsymbol{F}_m(\boldsymbol{F}_p,\boldsymbol{M}_{mask}) = \mathrm{softmax}\left[\frac{\boldsymbol{Q}(\boldsymbol{F}_p\boldsymbol{W}_2)^{\mathrm{T}}}{\sqrt{d}} + \boldsymbol{M}_{mask}\right]\boldsymbol{F}_p\boldsymbol{W}_3, \tag{13}$$

式中，$\boldsymbol{W}_n \in \boldsymbol{R}^{C\times H\times W}$ $(n\in\{1,2,3\})$为训练得到的变换矩阵，式(13)中d为多头自注意力的维度。当模型预测第 i 个字符时，i 与 j 相等，此时的掩码矩阵 $\boldsymbol{M}_{mask}$ 在(i,j)处为负无穷大，在式(13)中经过 softmax 函数，最终计算出的权重系数为 0，即忽略该字符本身信息，预测结果完全由其他字符的信息得到，避免了该字符信息对预测结果的干扰，提升了模型的预测能力；依次预测每个位置的字符，即可得到具有语义信息的字符特征 $\boldsymbol{F}_s$。

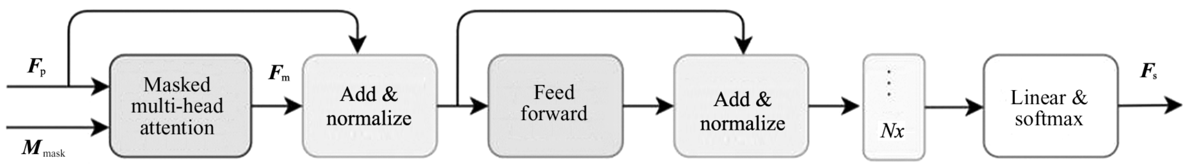

图 5　语义推理模块

Fig. 5　Semantic reasoning module

1.3　DFFU

视觉特征 $\boldsymbol{F}_v$ 以及语义特征 $\boldsymbol{F}_s$ 都是描述字符的特征，但两者却是不同的模态，视觉特征 $\boldsymbol{F}_v$ 是根据原始图像提取到的特征，语义特征 $\boldsymbol{F}_s$ 则是通过视觉特征 $\boldsymbol{F}_v$ 转化为字符概率推理预测得到的特征。两者都对文本识别结果有影响，为此设计了基于门控神经网络[18]的 DFFU，用来平衡两种不同模态的特征数据组合。其结构如图 6 所示，对视觉特征 $\boldsymbol{F}_v$、语义特征 $\boldsymbol{F}_s$ 进行拼接操作，sigmod 函数计算各特征的贡献比重。

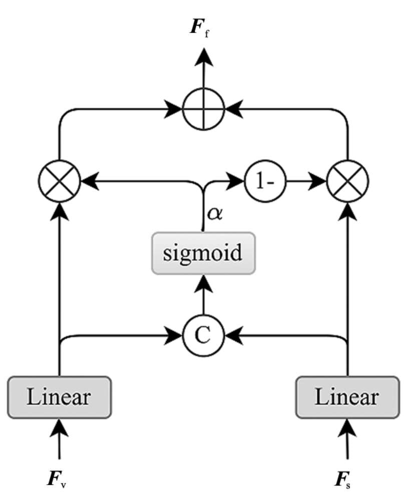

图 6　双特征融合单元

Fig. 6　Double feature fusion unit

融合后的特征 $\boldsymbol{F}_f$ 的表达式为：

$$\boldsymbol{F}_f = \boldsymbol{F}_v\boldsymbol{W}_4\alpha + \boldsymbol{F}_s\boldsymbol{W}_5(1-\alpha), \tag{14}$$

$$\alpha = \sigma\{[\boldsymbol{F}_v\boldsymbol{W}_4, \boldsymbol{F}_s\boldsymbol{W}_5]\}, \tag{15}$$

式中，$\boldsymbol{W}_4$、$\boldsymbol{W}_5$ 是训练时需要学习的参数，权重系数 α 由式(14)计算得到，$[\cdot,\cdot]$ 表示拼接操作，σ 为 sigmoid 函数。

1.4　损失函数

本文在训练模型时使用的损失函数 L 由 3 部分构成：L_v、L_s、L_f，三者按一定的权重求和作为整个模型的损失函数。

损失函数 L 的计算式为：

$$L = \lambda_v L_v + \lambda_s L_s + \lambda_f L_f, \tag{16}$$

式中，L_v、L_s、L_f 分别表示在视觉模型、语义模型、特征融合过程的损失。

L_v、L_s、L_f 均采用交叉熵损失函数表示，如式(17)所示：

$$L_{loss} = -\frac{1}{L}\sum_{i=1}^{L}\log p(g_i\,|\,I), \tag{17}$$

式中，$p(g_i\,|\,I)$ 表示将输入图像 I 中第 i 个字符预测为真实标签中字符 g_i 的概率。

2　实验与分析

2.1　实验数据集

本文选择 MJSynth(MJ)、SynthText(ST)两个数据集来训练模型，以及采用 6 个规则文本数据集和不规则文本数据集来验证模型性能。MJ 数据集有 890 万张训练图片，而 ST 有 700 万张训练图片。规则文本数据集的文本识别相对简单，有 IIIT5k、SVT、IC13 数据集，IIIT5k 共有 3 000 张经过裁剪后规则的测试图像，SVT 是谷歌从街景图像中裁剪而来的，共有 647 张图像；IC13 裁剪于 ICDAR2013 数据集，共有 1 095 张图像。而不规则文本数据集往往背景复杂、分辨率低、噪声干扰多，因此识别难度大，

数据集有 IC15、SVTP、CUTE。IC15 使用谷歌眼镜拍摄，图像拍摄位置、对焦随意，共有 1 811 张图像；SVTP 是谷歌从街景图像中裁剪而来，许多图像都发生了畸变，有 639 张图像；CUTE 有 288 张测试图像，是一个弯曲文本的图像数据集。

2.2　实验细节

本文实验所使用的软件平台为 Ubuntu18.04 以及 Pytorch 深度学习框架，使用的显卡为含有 8 G 内存的 NVIDIA RTX 3070。通过旋转、仿射、透视数据增强技术对图像进行增强，并将输入图像直接缩放到 32 × 128 大小。选择 Adam 算法作为网络模型的优化器，将初始动量设为 0.9，训练时参数 batch size 的值为 50，模型的初始学习率为 0.000 1，损失函数中的权重系数 λ_v、λ_s、λ_f 均设置为 1。

2.3　评估指标

本文在测试时将词准确率作为主要评价指标，部分实验使用字符准确率。词准确率将整个单词作为评价依据，其要求整个单词中的字符及个数都预测正确，并且预测序列顺序与单词字符顺序相一致，词准确率为预测准确的单词数与总单词数之比；而字符准确率是指预测准确的字符数与预测的总的字符数之比，公式表示为：

$$Accuracy = \frac{n}{N}, \tag{18}$$

式中，N 表示需要识别的单词或字符总数，n 表示模型正确识别出单词或字符的数量。

2.4　消融实验

2.4.1　多级 efficient Swin Transformer

本文实验训练数据以及验证数据来自不同的数据集，因此在 6 个不同的测试集中进行了验证。

为了探究 efficient Swin Transformer 层数 n 对模型性能的影响，设置了层数 n 为 4、6、8、10 的消融实验，实验结果如图 7 所示，图中，Avg 表示平均准确率。当 n 为 4 时，规则文本数据集的平均识别准确率为 84.6%，不规则文本的为 64.7%；当 n 增大到 8 时，相比于前者分别提升了 6.3%、13.6%；而增加到 10 时，模型的识别准确率却有所下降，说明了 efficient Swin Transformer 层数 n 太小或太大，模型对文本的识别性能都无法达到最好的结果，层数少提取的特征不够丰富，识别准确率低，层数过多则不同区域交互的特征信息过多而引起模型过拟合，泛化能力下降。

图 8 是骨干网络的消融实验，可以得知本文提出的多级 efficient Swin Transformer 性能优于 Resnet45，同时也表明骨干网络中嵌入 ECA 也可以提升模型性能，在不规则文本数据集上有 0.3% 的性能提升。

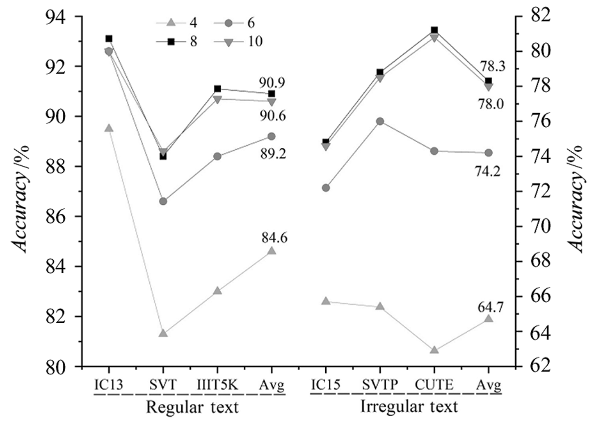

图 7　Efficient Swin Transform 层数 n 的消融实验

Fig. 7　Ablation experiment on the various numbers of efficient Swin Transformer

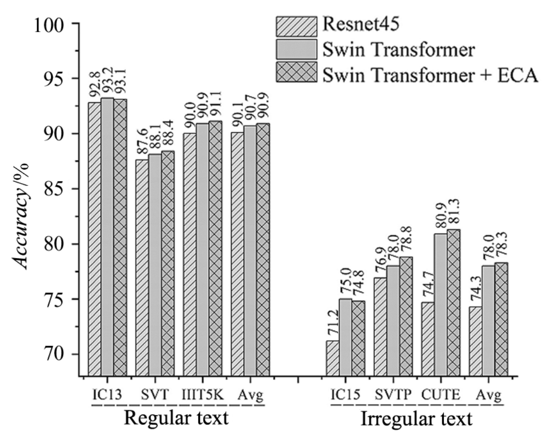

图 8　骨干网络的消融实验

Fig. 8　Ablation experiment of backbone network

2.4.2　DFFU

于特征融合而言，通常采用的方法有相加(add)、相乘(multiply)、拼接(concatenate)，因此本文对采用不同的特征融合方式进行了对比，实验结果如表 1 所示。可见本文的特征融合方式在 IC13、SVT、IC15、SVTP 数据上识别准确率相比于其他方法达到了最优，在规则文本数据集上比相加、相乘、拼接方法分别提升了 0.8%、0.1%、0.5%。说明了本文采用的特征融合方式比其他 3 种方法更有效。

2.4.3　CAM

为了验证本文对自注意力机制改进方式的有效性，针对自注意力 K、Q、V 的不同操作做了消融实验。表 2 为实验结果，表中，Linear 表示线性映射，PE(positional encoding)表示位置编码，Unet 表示普

通卷积的 UNet，DDUnet（deformable and dilated Unet）表示本文设计的含有可变形卷积、DConv 的 Unet。未对自注意力机制进行改进时，即实验 1，规则文本识别平均正确率为91.7％，不规则文本为79.6％。实验2在引入位置编码后，分别提升了1.4％、2.4％，表明了计算注意力权重时，嵌入位置信息对识别性能有提高。实验 2、3、4 表明引入 Unet 结构、可变形卷积和采用 DConv 增加感受野可以增大模型的识别准确率，尤其是不规则文本可以获得更好的性能。考虑到对 V 进行 DDUnet 操作与增大 efficient Swin Transformer 层数一样，增加了模型的深度，根据前面消融实验得知：层数的增加会降低模型性能，而 Q 是位置编码，因此没有对 V、Q 进行 DDU-net 操作的消融实验。

表 1　不同特征融合方式的对比

Tab. 1　Comparison of different feature fusion method

Fusion method	Regular text/％				Irregular text/％			
	IC13	SVT	IIIT5K	Avg	IC15	SVTP	CUTE	Avg
Add	93.1	88.4	91.1	90.9	74.8	78.8	81.3	78.3
Multiply	94.0	89.0	**91.8**	91.6	76.5	78.1	**81.9**	78.9
Concatenate	94.2	88.9	90.7	91.2	76.5	77.8	81.3	78.5
DFFU	**94.3**	**89.2**	91.7	**91.7**	**77.0**	**80.3**	81.6	**79.6**

表 2　字符注意力的消融实验

Tab. 2　Ablation experiment of character attention

ID	K	Q	V	Regular text/％				Irregular text/％			
				IC13	SVT	IIIT5K	Avg	IC15	SVTP	CUTE	Avg
1	Linear	Linear	Linear	94.3	89.2	91.7	91.7	77.0	80.3	81.6	79.6
2	Linear	Linear ＋PE	Linear	94.5	**91.9**	92.8	93.1	79.6	82.2	84.2	82.0
3	Unet	Linear ＋PE	Linear	94.6	91.7	93.1	93.1	80.0	82.5	85.4	82.6
4	**DDUnet**	**Linear ＋PE**	**Linear**	**95.2**	91.3	**93.9**	**93.5**	**80.3**	**85.4**	**85.8**	**83.8**

图 9 为本文注意力机制消融实验的可视化图，(1)—(4)分别是表 4 中实验 1—4，由图中(1)可见，采用自注意力机制，模型预测每个字符其关注的焦点较为稀疏，可以解释为自注意力的 3 个输入来自于同一特征，不含有位置信息，对特征的关注度较为分散[14]。而采用位置编码，则是更多地关注于每个字符所在位置的特征，如图中(2)所示。图中(4)相比于(2)、(3)，模型关注的范围更大，基本可以覆盖整个字符，并且不会将注意力放到其他字符上。说明了本文改进后的自注意力扩大了模型对字符的位置注意范围，可以得到更准确、更完整的字符特征，利于字符的识别。

图 9　注意力消融实验可视化图

Fig. 9　Visualization of attention ablation experiment

图 10 是本文设计的字符注意力对来自不同数据集图像的注意力可视化，可以发现，无论是对背景复杂、字符尺度不一、密集文本、模糊文本、弯曲文本图像都可以准确地关注到字符所在的位置。因此，识别准确率、注意力可视化图都说明了改进后的注意力机制具有良好的效果。

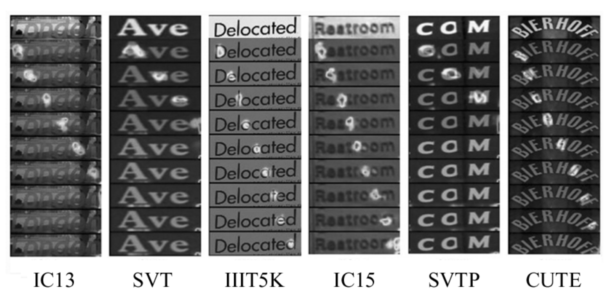

图 10　不同数据集的注意力可视化

Fig. 10　Visualization of attention in different datasets

2.4.4　SRM

最后为了探究 SRM 的有效性进行了消融实验，从字符识别的准确率和单词的准确率方面进行了比较，实验结果如表 3、表 4 所示，Baseline 表示本文的

多级 efficient Swin Transformer 和字符注意力。可见,模型加入 SRM 后,对不规则文本数据集的字符识别准确率提升了 1.0%,对单词的识别准确率提升了 3.1%,而模型本身对规则文本字符识别的准确率较高,无法发挥出 SRM 的推理优势,因此在规则文本数据集上提升较低。总体上看,SRM 的引入,提升了单个字符的识别准确率,从而提升了整个文本的识别准确率,尤其是提高了不规则文本的准确率。

表 3　语义建模对字符准确性的消融实验

Tab. 3　Ablation experiment of semantic modeling on character accuracy

Model	Regular text/%				Irregular text/%			
	IC13	SVT	IIIT5K	Avg	IC15	SVTP	CUTE	Avg
Baseline	**97.8**	**95.6**	96.4	96.6	89.4	89.9	89.1	89.5
Baseline + SRM	97.7	95.5	**96.8**	**96.7**	**90.9**	**91.2**	**89.3**	**90.5**

表 4　语义建模对词准确性的消融实验

Tab. 4　Ablation experiment of semantic modeling on word accuracy

Model	Regular text/%				Irregular text/%			
	IC13	SVT	IIIT5K	Avg	IC15	SVTP	CUTE	Avg
Baseline	94.2	90.3	93.4	92.6	77.4	81.6	83.0	80.7
Baseline + SRM	**95.2**	**91.3**	**93.9**	**93.5**	**80.3**	**85.4**	**85.8**	**83.8**

2.5　对比实验

除了消融实验,本文与将近些年其他的自然场景文本识别方法进行了对比,如表 5 所示。

由表中的数据可知,本文识别方法在规则文本数据集 SVT、不规则文本数据集 IC15、SVTP 上均优于其他方法,识别准确率在 SVT、IC15、SVTP 上分别达到了91.3%、80.3%、85.4%。在IC13数据集上虽然比 Mask TextSpotter 方法略低于 0.1%,对规则文本的识别优势并不明显,但在不规则文本数据上平均准确率比 Mask TextSpotter 高出了 1.4%。从平均准确率上看,本文提出的识别方法在规则和不规则文本数据集上,识别准确率相比于文中列出的大多数方法都要高,约为1%,可以说明本文方法具有不错的效果。

表 5　不同文本识别模型的对比

Tab. 5　Comparison of different text recognition models

Model	Regular text/%				Irregular text/%			
	IC13	SVT	IIIT5K	Avg	IC15	SVTP	CUTE	Avg
CRNN[1]	89.2	81.6	82.9	84.6	64.2	70.0	65.5	66.6
ASTER[6]	91.8	89.5	93.4	91.6	76.1	78.5	79.5	78.0
2D Attention[9]	92.7	90.1	94.0	92.3	76.3	82.3	86.8	81.8
Mask TextSpotter[8]	**95.3**	90.6	93.9	93.3	77.3	82.2	87.8	82.4
SEED[19]	92.8	89.6	93.8	92.1	80.0	81.4	83.6	81.7
Textscanner[20]	92.9	90.1	93.9	92.3	79.4	84.3	83.3	82.3
DAN[10]	93.9	89.2	94.3	92.5	74.5	80.0	84.4	79.6
RobustScanner[21]	94.5	88.1	**95.3**	92.6	77.1	79.5	**90.3**	82.3
SAR[22]	91.0	84.5	91.5	89.0	69.2	76.4	83.3	76.3
Ours	95.2	**91.3**	93.9	**93.5**	**80.3**	**85.4**	85.8	**83.8**

3　结　论

本文提出了一种无需校正、分割的自然场景文本识别方法,首先构建多级 efficient Swin Transformer,通过对输入图像进行窗口划分,使得不同窗口之间的特征可以实现空间上的信息交互,并且嵌入通道注意力突出某些重要通道的特征;接着设计了基于自注意力的 CAM 使得网络将关注的焦点置于特征图的文本区域,提取文本区域的视觉特征;然后使用 SRM 对文本序列建模,并得到语义特征;最后通过 DFFU 权衡视觉特征和语义特征贡献比重进行融合,经分类、转录得到最终的识别结果。通过在

6个不同的数据集上进行消融实验和对比实验,证明了本方法对规则和不规则文本识别的有效性。在将来研究中,将进一步探索新的文本位置注意力方法,使得网络提取更加完整的文本特征;此外,本文针对的是从左到右的一种文本行识别方法,而现实生活中还存在着从上到下的文本列,因此这也是将来需要考虑的问题。

参考文献:

[1] 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.

[2] BORISYUK F,GORDO A,SIVAKUMAR V. Rosetta: large scale system for text detection and recognition in images[C]//the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 19-23, 2018,London,United Kingdom. New York:ACM,2018:71-79.

[3] GAO Y,CHEN Y,WANG J,et al. Reading scene text with attention convolutional sequence modeling[J]. Neurocomputing,2017,339(C):161-170.

[4] LIU C Y,CHEN X X,LUO C J,et al. Deep learning methods for scene text detection and recognition[J]. Chinese Journal of Image Graphics,2021,26(6):1330-1367.
刘崇宇,陈晓雪,罗灿杰,等. 自然场景文本检测与识别的深度学习方法[J]. 中国图象图形学报,2021,26(6):1330-1367.

[5] SHI B,WANG X,LYU P,et al. Robust scene text recognition with automatic rectification[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016,Las Vegas,NV,USA. New York:IEEE,2016:4168-4176.

[6] 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.

[7] LI L R, ZHANG K,ZHANG Y L,et al. Natural scene text recognition algorithm based on multi-level feature selection[J]. Journal of Optoelectronics · Laser, 2022, 33(5):479-487.
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,2022,33(5):479-487.

[8] LIAO M H,LYU P Y,HE M H,et al. Mask textspotter: an end-to-end trainable neural network for spotting text with arbitrary shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,43(2):532-548.

[9] LYU P,YANG Z,LENG X,et al. 2D attentional irregular scene text recognizer[EB/OL]. (2019-06-13)[2022-08-26]. https://arxiv.org/abs/1906.05708.

[10] WANG T,ZHU Y,JIN L,et al. Decoupled attention network for text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence, February 7-12, 2020,New York,USA. Palo Alto,California,USA: AAAI Press,2020,34(7):12216-12224.

[11] MA Y Y,X B. Offline handwritten text recognition based on CTC Attention [J]. Progress in Laser and Optoelectronics,2021,58(12):1210007.
马洋洋,肖冰. 基于CTC-Attention脱机手写体文本识别[J]. 激光与光电子学进展,2021,58(12):1210007.

[12] FANG S,XIE H,WANG Y,et al. Read like humans: autonomous,bidirectional and iterative language modeling for scene text recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 19-25, 2021,Virtual. New York:IEEE,2021:7098-7107.

[13] LIU Z,LIN Y,CAO Y,et al. Swin transformer:hierarchical vision transformer using shifted windows[C]//IEEE/CVF International Conference on Computer Vision,October 10-17, 2021, Montreal, QC, Canda. New York: IEEE, 2021: 10012-10022.

[14] WANG Q L,WU B G,ZHU P F,et al. ECA-Net:efficient channel attention for deep convolutional neural networks [EB/OL]. (2019-10-08)[2022-08-26]. https://doi.org/10.48550/arXiv.1910.03151.

[15] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in Neural Information Processing Systems,December 4-9,2017,Long Beach,California. USA. Red Hook,NY,United States:Curran Associates Inc,2017:6000-6010.

[16] ZHU X,HU H,LIN S,et al. Deformable convnets v2:more deformable, better results[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition,June 15-20, 2019,Long Beach,CA,USA. New York:IEEE,2019:9308-9316.

[17] RONNEBERGER O，FISCHER P，BROX T. U-net：Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention，October 5-9，2015，Munich，Germany. Cham：Springer，2015：234-241.

[18] AREVALO J，SOLORIO T，MONTES-Y-GóMEZ M，et al. Gated multimodal units for information fusion[EB/OL]. (2017-02-07)[2022-08-26]. https://arxiv. org/abs/1702.01992.

[19] QIAO Z，ZHOU Y，YANG D，et al. SEED：semantics enhanced encoder-decoder framework for scene text recognition[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 13-19，2020，Seattle，WA，USA. New York：IEEE，2020：13528-13537.

[20] 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，February 7-12，2020，New York，USA. Palo Alto，California，USA：AAAI Press，2020，34(7)：12120-12127.

[21] YUE X，KUANG Z，LIN C，et al. RobustScanner：dynamically enhancing positional clues for robust text recognition[C]//European Conference on Computer Vision，August 2328，2020，online. Cham：Springer，2020：135-151.

[22] LI H，WANG P，SHEN C，et al. Show，attend and read：a simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence，January 27-February 1，2019，Honolulu，Hawaii，USA. Palo Alto，California，USA：AAAI Press，2019，33(1)：8610-8617.

作者简介：

熊　炜　(1976—)，男，博士，副教授，硕士生导师，主要从事数字图像处理和计算机视觉方面的研究.