# 基于文字局部结构相似度量的开放集文字识别方法

刘 畅 [1] 杨 春 [1] 殷绪成 [1]

**摘 要** 开放集文字识别 (Open-set text recognition, OSTR) 是一项新任务, 旨在解决开放环境下文字识别应用中的语言模型偏差及新字符识别与拒识问题. 最近的 OSTR 方法通过将上下文信息与视觉信息分离来解决语言模型偏差问题. 然而, 这些方法往往忽视了字符视觉细节的重要性. 考虑到上下文信息的偏差, 局部细节信息在区分视觉上接近的字符时变得更加重要. 本文提出一种基于自适应字符部件表示的开放集文字识别框架, 构建基于文字局部结构相似度量的开放集文字识别方法, 通过对不同字符部件进行显式建模来改进对局部细节特征的建模能力. 与基于字根 (Radical) 的方法不同, 所提出的框架采用数据驱动的部件设计, 具有语言无关的特性和跨语言泛化识别的能力. 此外, 还提出一种局部性约束正则项来使模型训练更加稳定. 大量的对比实验表明, 本文方法在开放集、传统闭集文字识别任务上均具有良好的性能.

**关键词** 开放集文字识别, 开放集学习, 泛用零样本学习, 组成学习

**引用格式** 刘畅, 杨春, 殷绪成. 基于文字局部结构相似度量的开放集文字识别方法. 自动化学报, 2024, **50**(10): 1977−1987

**DOI** 10.16383/j.aas.c230545 **CSTR** 32138.14.j.aas.c230545

## Open-set Text Recognition via Part-based Similarity

LIU Chang[1] YANG Chun[1] YIN Xu-Cheng[1]

**Abstract** Open-set text recognition (OSTR) is an emerging task that aims to address language bias and novel characters in open-world text recognition applications. Recent OSTR methods have achieved some success by decoupling the potentially biased context information with visual information. However, they tend to overlook the increasing importance of visual details. Given the biases in contextual information, detailed visual information became much more important in differentiating visually close characters. This work proposes an adaptive part-representation-based open-set text recognition framework and an open-set text recognition method via part-based similarly to improve the visual details modeling by explicitly modeling different character parts. Unlike radical-based methods, the proposed framework adopts a data-driven parting scheme, hence is language agnostic. A localization constraint is further proposed to address the instability caused by the parting scheme. The full framework steadily outperforms its baseline and yields reasonable performance on the close-set benchmarks.

**Key words** Open-set text recognition (OSTR), open-set learning, generalized zero shot learning, composition learning

**Citation** Liu Chang, Yang Chun, Yin Xu-Cheng. Open-set text recognition via part-based similarity. *Acta Automatica Sinica*, 2024, **50**(10): 1977−1987

文字识别是一个实际应用较为广泛的研究领域. 传统的文字识别任务只考虑训练集中见过的语言与字符[1], 不能很好地建模经常出现新字符的场景, 如演变迅速的网络图像文字识别应用环境和已知信息有限的古籍识别等任务. 对于字符集扩充的需求, 一些方法使用增量训练的方式来进行适应, 如 MRN[2] 提出通过补充额外训练专家模型并更新路由网络的方式进行适应, 其复杂度随专家数量的增长而提高. 麻斯亮等[3] 提出了一种基于训练不确定性的方式对训练集中的新字形进行发现并进行增量适应. 但这两种方法都仍限于识别已知类别, 故需要不断依赖用户反馈进行标注和训练. 对于未知类别, 小样本文字识别方法[4−6] 提出在提供辅助信息或示例样本后识别新字符的能力的需求与方案. 然而, 这两类方法均不能做到主动发现数据中的新字符[7]. 为了解决上述方法和任务建模上的局限性, Liu 等[8] 将这类需要发现新字符和新语言并进行增量识别的场景抽象为任务.

目前开放集文字识别 (Open-set text recognition, OSTR) 方法[8−10] 将字符视为一个整体进行特征提取并与其对应的类别中心计算相似度, 这种建

收稿日期 2023-09-04 录用日期 2024-04-19
Manuscript received September 4, 2023; accepted April 19, 2024
新一代人工智能国家科技重大专项 (2020AAA0109701), 国家杰出青年科学基金 (62125601), 国家自然科学基金 (62076024) 资助
Supported by National Science and Technology Major Project (2020AAA0109701), National Science Fund for Distinguished Young Scholars (62125601), and National Natural Science Foundation of China (62076024)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 北京科技大学计算机与通信工程学院 北京 100083
1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083

模方式对字符的细节特征表示不够理想, 容易造成图 1 所示的形近字的混淆[11] (本文中识别结果由图像、图像中文字的真值、模型预测结果构成. 真值行中白色代表训练中见过的字符, 黄色代表新字符. 结果行中, 绿色和红色分别表示识别正确和错误的结果. 紫色块表示模型拒识对应位置字符, 下同). 虽然在封闭集文字识别中, 形近字可以通过建模并利用上下文信息, 一定程度上缓解这一问题, 但是开放环境下, 训练集提供的上下文信息可能有较大偏差[9, 12], 从而使形近字问题的影响更为突出.

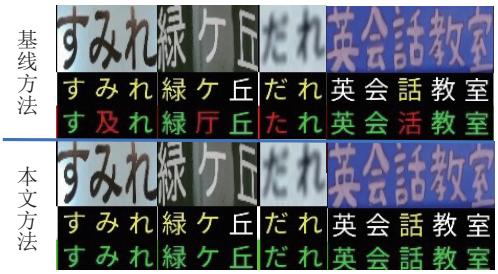



图 1 基于整字符识别方法的形近字混淆

Fig. 1 The confusion among close characters of the whole-character-based method

为了从视觉信息角度解决形近字混淆问题, Yu 等[11] 和 Zhang 等[13] 提出使用部件组成信息作为正则项来提升模型对细节信息的建模能力. 然而这些方法需要对于标签的结构性知识, 并缺乏在不同语言之间泛化的能力. 另一些方法则使用 Stroklet[14] 或局部图像[15] 等传统浅层部件特征进行字符表示. 这些方法虽然无需依赖领域知识, 但由于其过程的复杂性, 难以端到端地集成到基于深度学习的文字识别框架中. Hamming OCR[16] 提供了一种语言无关的部件编码方式, 然而这种方法会带来难度较大的实现和训练上的复杂度.

为了解决上述两类方法适用范围与复杂度上的局限性, 本工作提出一种基于自适应字符部件表示的开放集文字识别框架, 来改善模型对细节结构的建模. 该框架通过字符与标准模板的各个“部件”之间的相似程度进行分类或拒识操作. 与基于知识的部件构造方法[17−20] 不同的是, 本工作中部件由自适应的端到端训练得到. 这一特性使得该模型不再需要对特定语言所有字符的结构知识. 同时和 Hamming OCR[16] 相比, 该方法使用连续的特征向量作为部件表示, 一定程度上降低了模型的训练复杂度.

此外, 我们发现不加约束时部分部件的注意力图 (Attention map) 会关注无关区域, 导致模型性能波动较大. 我们认为这一现象是由于训练集涉及的字符数量过少, 导致部件表示过拟合到局部最优解. 具体来说, 模型训练涉及的 3 791 个字符类别不足以构成对标签空间足够密集的采样, 导致仅通过数据驱动学习得到的“部件”表示缺少泛化能力. 为了解决这一问题, 我们提出了一个局部性约束损失来约束部件注意力的局部性, 使其解空间更接近于结构知识意义上的部件, 如笔画和字根.

实验表明, 我们提出的基于自适应字符部件表示的开放集文字识别框架在开放集文字识别任务[8] 上取得了较好的性能. 该框架也具有一定程度的封闭集文字识别能力, 可以在生产环境中替代一些常见的封闭集文字识别方法[21−23].

本文的主要贡献如下:

1) 提出了一种基于自适应字符部件表示的开放集文字识别框架, 能够通过建模语言无关部件的方式缓解形近字混淆的问题.

2) 针对基于自适应字符部件表示的开放集文字识别框架存在的模型性能不稳定的问题, 提出了一个局部性约束正则项, 通过压缩部件表示的解空间方式对这个问题进行了有效缓解.

3) 本文提出的方法在开放集和封闭集上均有较好的性能表现[1].

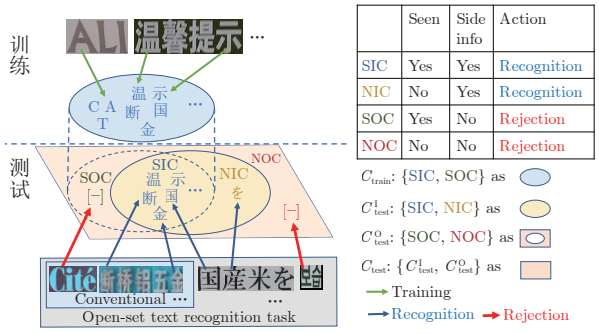
## 1 相关工作

### 1.1 开放集文字识别

随着文字识别方法的发展, 其应用场景也趋于复杂化. 一些开放环境, 如古籍识别、街景文字识别或网络图像文字识别等场景下, 会较频繁地出现训练集中没有覆盖到的字符. 传统封闭集方法[23] 往往会将这些新字符错误识别为已知字符, 导致识别错误和无法及时发现这些新字符. 同时, 传统方法需要使用大量数据进行重新训练才能识别这些新字符, 导致较长的响应时间. Liu 等[8] 提出开放集文字识别任务来建模这一用例, 该任务旨在提供主动发现新字符的能力, 并在不进行重新训练的前提下对新字符提供一定程度的识别能力, 同时还要保证对已知字符较好的识别能力.

任务层面上, 开放集文字识别任务 (见图 2) 将字符集按是否出现在训练集 (Seen/Novel, S/N) 的词条中和是否在测试时提供辅助信息 (In-set/Out-of-set, I/O) 划分为: 已知−集内 (Seen in-set characters, SIC)、已知集外 (Seen out-of-set characters, SOC)、新−集内 (Novel in-set characters, NIC) 和新集外 (Novel out-of-set characters, NOC) 四类. 模型需要对含有不提供辅助信息字符的词条

[1] 代码, 模型, 文档见: https://github.com/lancercat/OAPR




| | Seen | Side info | Action |
|---|---|---|---|
| SIC | Yes | Yes | Recognition |
| NIC | No | Yes | Recognition |
| SOC | Yes | No | Rejection |
| NOC | No | No | Rejection |

$C_{\text{train}}$: {SIC, SOC} as
$C^{\text{I}}_{\text{test}}$: {SIC, NIC} as
$C^{\text{O}}_{\text{test}}$: {SOC, NOC} as
$C_{\text{test}}$: $\{C^{\text{I}}_{\text{test}}, C^{\text{O}}_{\text{test}}\}$ as

图 2　开放集文字识别任务示意图[24]，经许可转载自文献 [24], ©《中国图象图形学报》编辑出版委员会, 2023

Fig. 2　An illustration of the open-set text task[24], reproduced with permission from reference [24], ©Editorial and Publishing Board of Journal of Image and Graphics, 2023

做拒识, 方便管理员进行检视. 并识别仅含有提供辅助信息字符的词条.

拒识性能通过行级别的召回率 (Recall, RE)、精准率 (Precision, PR) 和 F 指标 (F-measure, FM) 进行计算:

$$\begin{cases} RE = \dfrac{\sum_{i=1}^{N} \text{Rej}\left(Pred_{[i]}\right) \text{Rej}\left(Gt_{[i]}\right)}{\sum_{i=1}^{N} \text{Rej}\left(Gt_{[i]}\right)} \\ PR = \dfrac{\sum_{i=1}^{N} \text{Rej}\left(Pred_{[i]}\right) \text{Rej}\left(Gt_{[i]}\right)}{\sum_{i=1}^{N} \text{Rej}\left(Pred_{[i]}\right)} \end{cases} \tag{1}$$

式中, $N$ 为测试集样本数量, $Pred_{[i]}$ 为测试集中的第 $i$ 个样本的预测标签, $Gt_{[i]}$ 为测试集中的第 $i$ 个样本的标签真值, Rej 为判断输入序列中是否含有未知字符的示性函数.

$RE$ 与 $PR$ 两个指标通过 $FM$ 进行综合度量,

$$FM = \frac{2RE \cdot PR}{RE + PR} \tag{2}$$

模型的识别性能则主要通过行级别准确率 (Line accuracy, LA) 进行计算,

$$LA = \frac{1}{N} \sum_{i=1}^{N} \text{Same}\left(Gt_{[i]},\ Pred_{[i]}\right) \tag{3}$$

式中, 函数 Same 为判定两个输入序列是否完全一致的示性函数.

模型性能可以辅以字符级别的准确率 (Character accuracy, CA) 进行参考:

$$CA = 1 - \frac{\sum_{i=1}^{N} \text{NED}\left(Pred_{[i]},\ Gt_{[i]}\right)}{\sum_{i=1}^{N} \text{Len}\left(Gt_{[i]}\right)} \tag{4}$$

式中, Len 为返回输入序列长度的函数, 函数 NED 返回两个输入序列间编辑距离.

方法层面上, OSOCR 方法[8] 提供了一个实现开放集文字识别的基本框架. 一些研究者试图通过切断上下文信息对字符特征的影响来解决上下文偏差对性能造成的影响[9]. 然而这些方法忽视的一个问题是, 在上下文信息缺失, 甚至有偏的情形下, 形近字之间的混淆问题变得更加突出. 本文旨在通过显式建模字符部件, 并计算基于部件的相似度来增强对细节信息的建模, 从而改善开放环境下模型的性能.

### 1.2　部件相似度相关方法综述

基于目标部件组成信息的方法在不同领域均有广泛应用, 通常可以按照部件是否根据先验知识进行显式划分分为两类.

一类方法通过显式地利用组成信息的先验知识来改善模型的性能[25] 或直接对各个部件 (属性) 进行预测来进行识别[26]. 这类方法在文本领域中较为常见, 其主要利用的先验知识包括字根[17, 27]、输入法编码 (五笔、郑码)[28] 和笔画序列[19] 等.

在直接利用部件相似度进行分类的方法中, 分类的主要实现方式分为基于规则的匹配[17, 27] 和基于编码的匹配[18, 29]. 在基于规则的匹配方法中, 模型首先对图片中的每个字符生成一个结构信息预测序列[19, 30], 并将这个预测序列与所有字符标签对应的真值序列进行比较, 取相似度最高的匹配结果对应的标签作为该字符的预测结果. 基于编码的匹配方法则将结构信息通过编码器编码为特征向量, 并与从待分类样本中提取的特征进行相似度计算得到置信度[18]. 显式建模部件信息虽然比较直观且有效, 但是存在对领域知识依赖性强和难以泛化到其他语言的问题.

另一类方法[31−32] 则通过训练得到对于部件的表示. 虽然这些方法解决了对先验知识依赖的问题, 但是在文字识别领域应用较少. 一些早期非深度学习方法使用聚类得到 Strokelet, 作为部件来对字符进行表征[14], 或利用 Hough Forest 基于从字符图像上提取的局部图像进行集成分类[15]. 这些方法由于过程复杂, 较难用可导的连续函数进行表示或近似, 导致其难以端到端地集成到深度学习文字识别模型中. Li 等[16] 提出了一种可以端到端训练的字符部件表示方法, 但其实现与训练复杂度都比较高.

基于自适应字符部件表示的开放集文字识别框架可以进行端到端的训练, 同时, 相对整字符表示的基线模型有一定速度上的提升, 还能够缓解目前基于隐式部件的文字识别方法面临的复杂度问题.

## 2　基于自适应字符部件表示的开放集文字识别框架

为了解决现有文字识别方法对细节建模能力有限这一问题, 我们提出一种基于自适应字符部件表

示的开放集文字识别框架 (见图 3). 与现有的开放集文字识别框架[8–9] 不同, 该框架的特征图建模文字图像和原型的局部特征, 每个时序对应的字符特征由该字符的几个部件的特征拼接构成, 字符与原型之间的相似度由对应部件的相似度得到.

该框架由 4 个主要模块构成, 分别为共享的特征提取器模块, 行级部件注意力模块 (Part attention line module, PALM), 字符级部件注意力模块 (Part attention character module, PACM) 和部件相似度分类模块 (Part similarity recognition module, PSRM). 该框架的训练和测试流程与 Liu 等[8] 的框架相似. 训练时, 模型每一轮次采样一组样本图像, 并根据标签对字形进行采样[8]. 模型对采样结果进行前馈.

首先, 采样后的字形送入卷积骨干网络提取特征, 并使用 PACM 模块对该特征进行采样, 从而得到 Batch 字符集对应的原型 $P$. 然后, 对样本图像使用相同的卷积骨干网络提取特征, 使用 PALM 模块预测样本字符串的长度 $\hat{t}$, 同时将特征进行序列化, 从而得到各个时序的字符的部件特征表示 $F$. 最后, 通过 PSRM 模块对各时刻 $i$ 对应字符的部件表示 $F_{[i]}$ 和字符原型 $P$ 进行对比, 识别或拒识各时刻字符类别, 输出预测结果 $\hat{\boldsymbol{Y}}$.

模型训练主要通过监督长度预测 $\hat{t}$ 和识别结果 $\hat{\boldsymbol{Y}}$ 更新模型权重, 部件的局部性则通过局部性约束正则项 $L_{pac}$ 实现约束.

除测试集字符原型 $P$ 可以在测试前进行缓存外, 测试过程大体与训练前馈过程相同. 原型的添加或删除可以随需求变更, 调用骨干网络和 PACM 增量进行.

## 2.1 部件表示提取

### 2.1.1 行级部件注意力模块

行级部件注意力模块 (PALM, 见图 4) 将文本行级别的视觉特征图 $M$ 提取为字符视觉特征构成的序列 $F$, 其中每一个时序 $i$ 对应的特征 $F_{[i]}$ 由 $p$ 个部件特征构成. PALM 与其他开放集识别方法[8–9] 中的 L-CAM 类似, 不同之处是该模块对每个时序的字符提取复数个部件特征, 并对特征图进行了空间编码[33].

该模块对骨干网络不同层的图像特征断开梯度[9] 并进行空间位置编码, 送入特征金字塔网络 (Feature pyramid networks, FPN) 进行序列长度和各个字符部件位置的预测.

FPN 的各层中间特征首先经过平均池化 (Global average pooling, GAP) 得到全图特征, 再进行拼接后通过一个 MLP 进行序列长度 $\hat{t}$ 的预测, 该预测值用来在测试时截断预测序列.

该模块还通过对 FPN 的最后一层输出特征 $M_t$ 进行卷积, 得到各时序字符位置 $A_c \in \mathbf{R}^{t\times 1\times w\times h}$ 和各字符部件位置 $A_p \in \mathbf{R}^{t\times p\times w\times h}$, 并将二者相乘得到字符部件的注意力图 $A \in \mathbf{R}^{t\times p\times w\times h}$, 其中, $w$, $h$

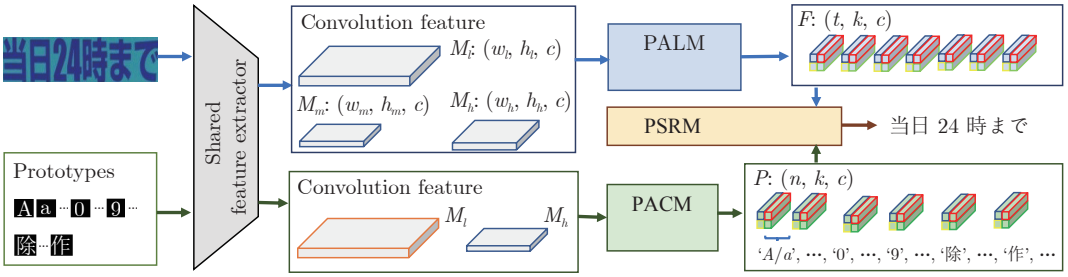



图 3 本文提出的基于自适应字符部件表示的开放集文字识别框架

Fig. 3 The proposed open-set text recognition framework with adaptive part representation

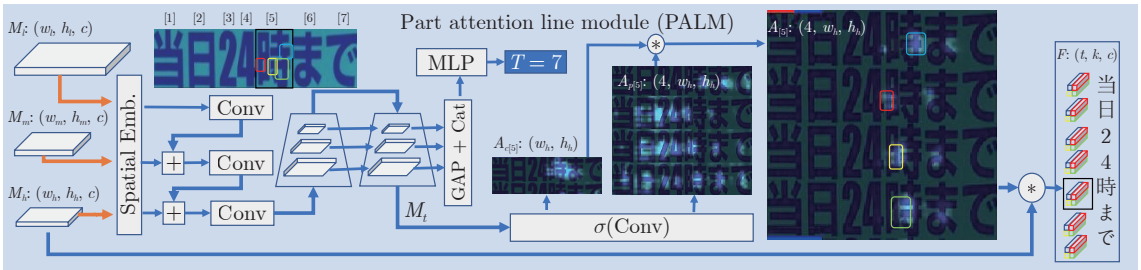



图 4 行级部件注意力模块

Fig. 4 The proposed part attention line module

分别为特征图的宽和高.

$$\begin{cases} A_c = \sigma\left(\mathrm{Conv}_c\left(M_t\right)\right) \\ A_p = \sigma\left(\mathrm{Conv}_p\left(M_t\right)\right) \\ A = A_c \odot A_p \end{cases} \tag{5}$$

式中, $\odot$ 为逐位置相乘, $\mathrm{Conv}_p$ 和 $\mathrm{Conv}_c$ 分别为 1×1 的卷积层, $\sigma$ 为 Sigmoid 激活函数, $p$ 为预设部件的数量, 本文中 $p$ 取 4, 即每个字符被自适应地划分为 4 个部件.

最后, 各个字符的部件特征表示 $F$ 根据 $A$ 对未断开梯度的 $M_h$ 采样得到:

$$F = \sum_{i=1}^{i=w}\sum_{j=1}^{j=h}\left(A \odot M_h\right)_{[i,\,j]} \tag{6}$$

### 2.1.2　局部性约束

实作中, 我们发现引入部件表示后, 同一参数训练得到的模型性能差异较大 (见消融实验, 第 3.2 节), 且模型性能对初始状态敏感. 同时我们观察到模型学习得到的部件存在注意力趋于分散的现象, 说明仅依赖数据驱动难以得到有效的部件表征. 这说明模型有概率在早期轮次陷入局部最优, 导致性能不稳定问题, 并表现为初值敏感. 为了解决这一问题, 我们提出了局部性约束正则项来约束采样点到采样区域重心的距离. 具体来说, 第 $t$ 时刻的第 $k$ 个部件的重心坐标 $g_{[t,\,k]}$ 定义为:

$$g_{[t,\,k]} = \sum_{i=1}^{i=w}\sum_{j=1}^{j=h} A_{[t,\,k,\,i,\,j]} \odot G_{[i,\,j]} \tag{7}$$

式中, $G \in [-1,\,1]^{w\times h\times 2}$ 为与注意力图等大的 2D 均匀坐标网格, 左上点坐标为 $(-1,\,-1)$, 右下点坐标为 $(1,\,1)$.

对于任一部件 $(t,\,k)$, 其采样点到其重心的平均距离, 可以通过对距离矩阵采样计算:

$$d_{[t,\,k]} = \sum_{i=1}^{i=w}\sum_{j=1}^{j=h}\left(A_{[t,\,k]} \odot \left\|G - g_{[t,\,k]}\right\|\right)_{[i,\,j]} \tag{8}$$

该局部性约束使得样本中所有部件的平均距离尽可能小, 即:

$$L_{md} = \frac{1}{t^* p}\sum_{t=1}^{t=t^*}\sum_{k=1}^{k=p} d_{[t,\,k]} \tag{9}$$

式中, $t^*$ 为文字标签长度的真值.

同时, 为了防止某个部件的注意力图被拉向全 0 这一显然解, 我们提出注意力图响应下界损失 $L_{mb}$ 对每个部件注意力图的最大值进行下界约束:

$$L_{mb} = \frac{1}{t^* p}\sum_{t=1}^{t=t^*}\sum_{k=1}^{k=p}\mathrm{ReLU}\left(\lambda_a - \max_{(i,\,j)} A_{[i,j]}\right)_{[t,\,k]} \tag{10}$$

式中, $\lambda_a$ 为设置的部件激活阈值, 本文中设置为 0.8. 则, 局部性约束 $L_{pac}$ 的形式为:

$$L_{pac} = L_{md} + L_{mb} \tag{11}$$

由于模型初始化时注意力图是随机的, 导致任一字符的注意力图的重心靠近图像中点, 这种情况下, 该约束会将不同时刻的注意力图向图像中点拉近, 影响模型的收敛速度. 为此, 我们设定该约束只在模型初步收敛后再生效, 即 $L_{pac}$ 在模型训练的前 10 k 个轮次不生效.

## 2.2　基于部件分类

该框架使用基于部件的原型构造和距离度量模块来实现分类.

对于基于部件的原型构造, 首先, 该框架使用骨干 (Backbone) 网络提取字形 (Glyph) 的局部视觉特征. 然后用字符级部件注意力模块 (PACM, 见图 5) 提取部件特征并组合成为原型. 此处, PACM 模块通过对截断梯度的低层次 (Low-level) 特征进行卷积得到各个部件的注意力图. 最后, 模型通过部件的注意力图采样高层次 (High-level) 特征图, 并对其模长进行正则化, 得到该字形对应的原型.

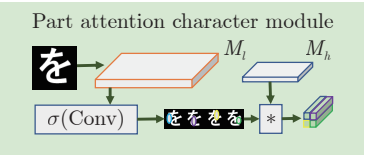



图 5　字符级部件注意力模块

Fig. 5　The proposed part attention character module

对于基于部件的距离度量, 部件相似度分类模块 (PSRM, 见图 6) 度量序列中字符 $t$ 的部件特征 $F_{[t]}$ 与原型 $P$ 之间的相似度, 并给出该字符的预测置信度 $\hat{\boldsymbol{Y}}_{[t]}$.

对于任一原型 $P_{[i]}$ 与字符部件特征 $F_{[t]}$, 部件相似度分类模块首先比较对应部件 $p$ 间的相似度 $z'_{[t,\,p,\,i]}$:

$$z'_{[t,\,p,\,i]} = \left|F_{[t,\,p]}\right|\cos\left(P_{[i,\,p]},\ F_{[t,\,p]}\right) \tag{12}$$

对于该相似度 $z'$ 加入一个失配阈值作为部件失配的置信度:

$$z_{[t,\,p]} = \left[z'_{[t,\,p]},\ s_{unk}\left|F_{[t,\,p]}\right|\right] \tag{13}$$

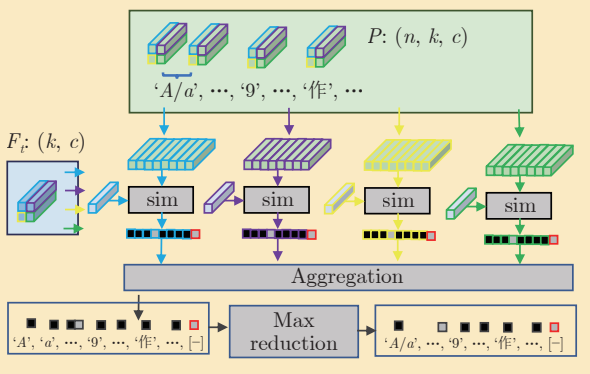



图 6 部件相似度分类模块

Fig. 6 The proposed part similarity recognition module

式中, $s_{unk}$ 为部件适配的置信度, 为可训练参数, 通过梯度下降进行更新.

各部件的置信度 $z_{[t,\,p]}$ 通过 Aggr 函数归并为字形相似度 $\tilde{S}$:

$$\tilde{S}_{[t]} = \mathrm{Aggr}\left(z_{[t,\,1]}, \cdots, z_{[t,\,p]}\right) \tag{14}$$

本文中 Aggr 实现为求均值. 与其他开放集文字识别方法[8−9] 类似, $\tilde{S}$ 经过最大归约 (Max-reduction) 得到字符置信度预测 $\hat{\boldsymbol{Y}}$:

$$\begin{cases} S^c_{[t,\,i]} = \max\limits_{\{j|\phi(j)=i\}} \left(\tilde{S}_{[t,\,j]}\right) \\ \hat{\boldsymbol{Y}}_{[t]} = \mathrm{softmax}\left(S^c_{[t]}\right) \end{cases} \tag{15}$$

式中, $\phi(j)$ 为映射函数, 返回原型 $j$ 对应的标签.

### 2.3 优化

与 OpenCCD[9] 类似, 该模型主要通过监督部件表示提取中的长度预测 $\hat{t}$ 和部件相似度分类模块中的标签预测 $\hat{\boldsymbol{Y}}$ 来进行训练:

$$L_{rec} = L_{ce} + L_t \tag{16}$$

其中 $L_{ce}(\hat{\boldsymbol{Y}},\, \boldsymbol{Y}^*)$ 为逐时序的字符分类交叉熵损失,

$$L_{ce}\left(\hat{\boldsymbol{Y}},\, \boldsymbol{Y}^*\right) = \frac{1}{t^*}\sum_{t=1}^{t=t^*} \mathrm{CrossEntropy}\left(\hat{\boldsymbol{Y}}_{[t]},\, \boldsymbol{Y}^*_{[t]}\right) \tag{17}$$

$L_t\left(\hat{t},\, t^*\right)$ 则为对序列长度预测的交叉熵损失,

$$L_t\left(\hat{t},\, t^*\right) = \mathrm{CrossEntropy}\left(\hat{t},\, t^*\right) \tag{18}$$

整个框架的损失定义为识别损失加部件正则项:

$$L = L_{rec} + \lambda_r L_{pac} \tag{19}$$

式中, $\lambda_r$ 为正则项的系数, 本文中经验性的设定为 0.1.

## 3 实验

### 3.1 实现细节

该框架的实现基于 Liu 等[8] 的代码库完成. 为了适应更小的文字, 模型的输入图片大小设置为 48 × 200 像素, Padding 方式被改为左上对齐以适应空间位置编码. 字符图像尺寸保持原设置为 32 像素, 部件数量设置为 4.

模型性能评测方式参考 Liu 等[8] 分为消融实验、开放集性能测试和封闭集性能测试. 具体来说, 消融实验和开放集性能测试模型在中英文训练集上进行训练, 并在日文数据集上进行测试. 训练用的数据集包括 ART[34]、LSVT[35]、CTW[36]、RCTW[37] 以及 MLT[38] 的中文和拉丁文部分. 训练集中标签含有英文, 中文一类简体字和数字以外字符的样本被丢弃, 以防止标签泄露. 模型在该混合数据集上训练 200 k 个轮次, Batchsize 设为 48. 测试用的数据集为 MLT 中日文子集, 未作额外处理[8].

对于封闭集识别任务, 我们按照 Liu 等[8] 的方法, 在 MJ[39] 和 ST[40] 数据集上进行训练, 并在 IIIT5K[41] CUTE[42]、SVT[43]、ICDAR2003 (IC03)[44] 和 ICDAR-2013 (IC13)[45] 等数据集上进行测试. 由于训练集规模的明显增大, 模型在该集合上训练 800 k 个轮次, Batchsize 设为 48.

### 3.2 消融实验

本文在泛化零样本学习 (Generalized zero-shot learning, GZSL) 划分上进行消融实验来验证基于自适应字符部件表示的开放集文字识别框架和局部性约束的有效性. 为了减小随机因素对实验结果的影响, 我们对每个模型以相同参数进行 3 次独立训练, 并计算 3 次模型的行级别准确率的均值 (Avg LA), 以及最好性能与最坏性能差值 (Gap LA). 定量结果如表 1 所示. 具体模型测试集性能曲线见图 7. 图 7 显示了消融实验中的各模型分别进行三次独立运行 (训练至 400 k 轮) 的模型性能. 其中: 折线对应各模型三次独立运行的平均性能, 阴影区域对应各模型的性能范围. 从定性实验的角度, 我们在图 8 中给出了基线模型和基于自适应字符部

表 1 消融实验

Table 1 Ablative studies

| | 自适应字符部件表示 | 局部性约束 | Avg LA ↑ | Gap LA ↓ |
|---|---|---|---|---|
| Ours | ✓ | ✓ | 39.61 | 4.91 |
| 仅自适应字符部件表示 | ✓ | | 38.91 | 6.54 |
| 字符整体特征 | | | 34.04 | 2.27 |

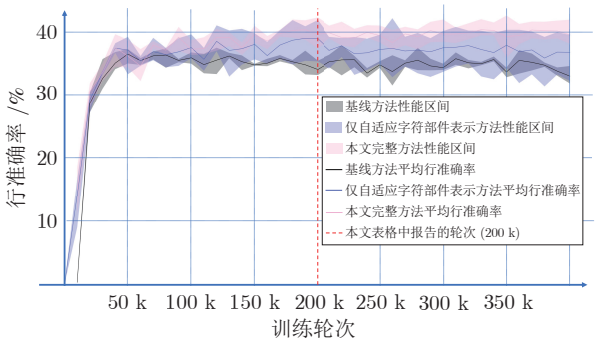



图 7　消融实验详细结果图

Fig. 7　Details of each individual run in the ablative studies

件表示的开放集文字识别框架识别结果的样本对照, 表现出对于形近字的混淆有一定改善效果.

从性能均值来说, 我们提出的基于部件的框架能够较为明显地提升模型性能. 由表 1 可知, 基于自适应字符部件表示的开放集文字识别框架可以在开放环境识别任务上产生 5% 左右的提升, 而局部性约束可以产生 0.7% 左右的额外提升. 此处, 不加约束的部件框架虽然大多数情况下能带来较好的识别性能, 但由于标签类别数量有限, 其通过数据驱动学习得到的部件表示容易陷入局部最优, 表现为模型性能受初值影响较大. 同时, 各部件对应的注意力区域表现出较差的局部性, 极端情况下可能收敛为散布在复数字符之间的区域. 这种性质导致仅依赖数据驱动时, 最差情况下性能与基线差别不大甚至略低.

从图 7 和表 1 可以看出, 我们提出的局部性约束正则项对性能上界有一定的提升, 但显著提升了性能下界, 说明该正则项有效地缓解了引入部件表示带来的过拟合问题. 然而, 同时应当注意的是, 虽然整体方法下界在大多数轮次性能都优于基线的上界, 该正则项并没有完全解决初值敏感问题, 表现为性能上下界差仍较基线有较大差异. 下一步, 我们考虑引入其他语言的字符集并进行增广[10] 来对标签空间进行更稠密的采样, 缓解部件表示学习中的过拟合问题.

此外, 图 7 中曲线说明模型在我们报告的训练轮次即 200 个轮次上已经收敛, 且该消融实验结论对于收敛后的绝大多数轮次成立.

### 3.3　开放集文字识别性能

我们按照 Liu 等[9] 的方式来评测模型在开放环境下的性能, 即验证在中文数据集上训练的模型对日文数据集的迁移识别和拒识泛化能力. 我们报告模型在 GZSL、开放集识别 (Open-set recognition, OSR)、泛化开放集识别 (Generalized open-set recogniton, GOSR) 和 OSTR 四个划分下的性能. 其中, GZSL 划分通过将所有测试数据集中字符加入 $\boldsymbol{C}_{test}^{k}$ 来重点衡量模型对新旧字符的识别能力, 这种划分更加接近常规的零样本文字识别任务[27, 29]. OSR 划分主要测试模型的拒识能力, 该划分将所有训练集中没有覆盖的字符划入 $\boldsymbol{C}_{test}^{u}$, 即对所有新字符均不提供辅助信息并要求模型对其进行拒识, 接近于一般的开放集识别任务定义[46]. GOSR 划分综合了 GZSL 和 OSR 两个任务, 要求模型识别提供新旧字符的辅助信息的字符, 并保留对不提供辅助信息的字符的拒识能力, 用于衡量模型在加入一定量新字符后的性能. 实际上, 这两种能力都会随着字符集的增长而下降, 这也是目前封闭集识别方法在中文数据集上[11] 性能远低于英文数据集性能[23] 的一个重要原因, 这也成为一个需要进一步探索的方向. OSTR 划分在 GOSR 的基础上, 额外引入了对训练集中包含的字符的拒识需求, 对应实际应用中, 用户希望将不常出现的字符或语言临时移除, 提高模型推理速度的需求. 全部四个划分下的定量实验结果见表 2, GZSL 划分下的定性结果见图 9.

从表 2 结果来看, 模型的识别能力相较近年的方法[8] 有较明显的优势. 模型性能的提升来自两部分, 一部分来自基线 (表 1 中字符整体特征) 自身, 另一部分来自本文提出的方法 (见消融实验). 特别的, 本文的基线训练过程中移除了数据中没有的中文繁体字, 以确保比较的公平性. 在已有工作中, 这些字符会被当作负样本, 导致性能损失. 在不经过重新训练的前提下, 本文模型能够在没有见过的语

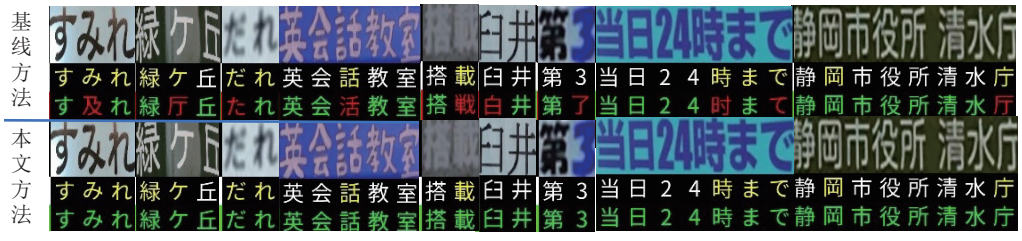



图 8　基线方法 (上侧) 与我们的模型 (下侧) 的识别结果对比

Fig. 8　More comparison between base method (top) and the proposed framework (bottom)

表 2 开放集文字识别性能

Table 2 Performance on open-set text recognition benchmarks

| 任务 | $C^k_{test}$ | $C^u_{test}$ | 方法 | 来源 | **LA** (%) | Recall (%) | Precision (%) | **F-measure** (%) |
|---|---|---|---|---|---|---|---|---|
| GZSL | Unique Kanji | ∅ | OSOCR-Large[8] | PR′ 2023 | 30.83 | — | — | — |
| | Shared Kanji | | OpenCCD[9] | CVPR′ 2022 | 36.57 | — | — | — |
| | Kana, Latin | | OpenCCD-Large[9] | CVPR′ 2022 | 41.31 | — | — | — |
| | | | **Ours** | — | **39.61** | — | — | — |
| | | | **Ours-Large** | — | 40.91 | — | — | — |
| OSR | Shared Kanji | Unique Kanji | OSOCR-Large[8] | PR′ 2023 | 74.35 | 11.27 | 98.28 | 20.23 |
| | Latin | Kana | OpenCCD-Large*[9] | CVPR′ 2022 | **84.76** | 30.63 | **98.90** | 46.78 |
| | | | **Ours** | — | 73.56 | 64.30 | 96.21 | **76.66** |
| | | | **Ours-Large** | — | 77.15 | 60.59 | 96.80 | 74.52 |
| GOSR | Shared Kanji | Kana | OSOCR-Large[8] | PR′ 2023 | 56.03 | 3.03 | 63.52 | 5.78 |
| | Unique Kanji | | OpenCCD-Large*[9] | CVPR′ 2022 | **68.29** | 3.47 | **86.11** | 6.68 |
| | Latin | | **Ours** | — | 65.07 | **54.12** | 82.52 | **64.65** |
| | | | **Ours-Large** | — | 67.40 | 47.64 | 82.99 | 60.53 |
| OSTR | Shared Kanji | Kana | OSOCR-Large[8] | PR′ 2023 | 58.57 | 24.46 | 93.78 | 38.80 |
| | Unique Kanji | Latin | OpenCCD-Large*[9] | CVPR′ 2022 | 69.82 | 35.95 | **97.03** | 52.47 |
| | | | **Ours** | — | 68.20 | **81.04** | 89.86 | **85.07** |
| | | | **Ours-Large** | — | **69.87** | 75.97 | 91.18 | 82.88 |

注: * 表示原论文中未报告的性能, 数据来自原作者代码仓库和释出的模型.

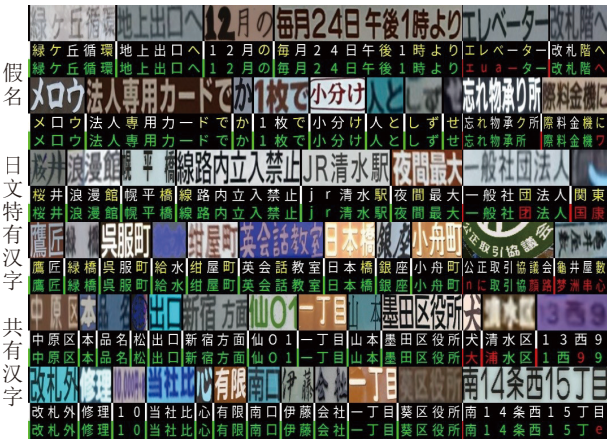



图 9 日文测试数据集上的识别结果 (GZSL 划分)

Fig. 9 Sample results from the Japanese testing data set (With GZSL split)

言上保持 40% 以上的整行识别正确率. 这意味着模型可以在数分钟内获得对新语言的一定程度的处理能力. 同时, 其拒识能力较其他方法[8] 有显著的优势, 并在所有测试中保持 45% 以上的召回率 (Recall), 这表明模型能够自动发现一半左右其不能处理的新字符, 及时通知管理员进行处理. 该模型同时能达到 82% 以上的准确率 (Precision), 意味着绝大多数来自模型的通知都对应数据中出现了缺失辅助信息的字符.

总的来说, 从以上指标看, 模型能够及时发现数据中的新字符, 并在新模型重新训练完成前, 或由于数据缺失无法重新训练时, 通过提供标准字形的方式, 快速获得一定程度上对这些数据的识别能力. 我们同时在韩文上进行了对该模型的定性实验, 结果见图 10. 该模型对韩文有一定识别能力, 但是形近字问题并没有被完全解决, 尤其是当这些细节特征在训练集上不体现区分度的时候, 如韩文中区分圆形部件和方形部件, 但中文通常不区分两者. 此时, 这些特征有时不会被网络建模. 对于这一问题, 我们计划在未来引入其他无关语言的字符集, 扩充标签空间来解决这一问题.

值得注意的是, 大模型的拒识性能相对比小模型较差, 主要体现在对新字符 Recall 偏低, 即新字

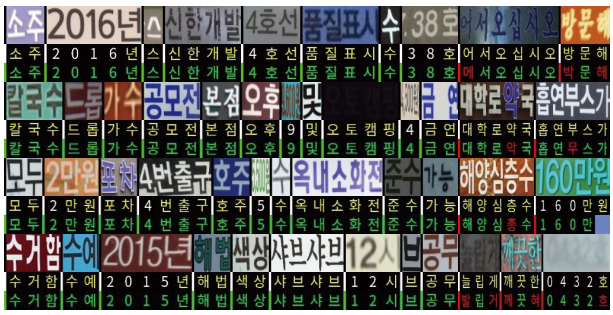


图 10 韩文数据集识别结果

Fig. 10 Sample recognition results from the Korean data set

表 3　封闭集文字识别基准测试性能及单批次推理速度

Table 3　Performance on close-set text recognition benchmarks and single batch inference speed

| 方法 | 来源 | IIIT5K | CUTE | SVT | IC03 | IC13 | GPU | TFlops | FPS |
|---|---|---|---|---|---|---|---|---|---|
| CA-FCN*[22] | AAAI′19 | 92.0 | 79.9 | 82.1 | — | 91.4 | Titan XP | 12.0 | 45.0 |
| Comb.Best[23] | ICCV′19 | 87.9 | 74.0 | 87.5 | 94.4 | 92.3 | Tesla P40 | 12.0 | 36.0 |
| PERN[47] | CVPR′21 | 92.1 | 81.3 | 92.0 | 94.9 | 94.7 | Tesla V100 | 14.0 | 44.0 |
| JVSR[48] | ICCV′21 | 95.2 | **89.7** | 92.2 | — | 95.5 | RTX 2080Ti | 13.6 | 38.0 |
| ABINet[49] | T-PAMI′23 | **96.2** | 89.2 | **93.5** | **97.4** | 95.7 | V100 | 14.0 | 29.4 |
| CRNN[21, 23] | T-PAMI′17 | 82.9 | 65.5 | 81.6 | 92.6 | 89.2 | Tesla P40 | 12.0 | **227.0** |
| Rosetta[23, 50] | KDD′18 | 84.3 | 69.2 | 84.7 | 92.9 | 89.0 | Tesla P40 | 12.0 | 212.0 |
| ViTSTR[51] | ICDAR′21 | 88.4 | 81.3 | 87.7 | 94.3 | 92.4 | RTX 2080Ti | 13.6 | 102.0 |
| GLaLT-Big-Aug[52] | TNNLS′23 | 90.4 | 77.1 | 90.0 | 95.2 | 95.3 | — | — | 62.1 |
| **Ours-Large** | — | 89.06 | 77.77 | 80.68 | 89.61 | 87.98 | Tesla P40 | 12.0 | 85.7 |

符有更大概率错误落入某一已知字符分界面[2], 我们推测该现象可能是由于大模型更多的参数引起的, 特征提取模块过拟合导致的.

### 3.4　封闭集文字识别性能

由于大多数实际应用场景以训练集中见过的字符为主, 这些文字样本的识别准确率是影响开放集文字识别方法实用性的重要因素. 这里我们参考 Liu 等[9] 的实验设置, 采用封闭集文字识别测试的协议事实标准[23], 对模型进行测试. 该组实验定量结果及与最近几年的轻量方法的对比见表 3, 表中方法按照 60 FPS 以下和 60 FPS 以上分为上下两个部分, 各部分按发表时间排序. 速度上看, 该模型较为轻量, 在 Tesla P40 显卡上单 batch 下可以达到 85.70 FPS 的推理速度, 且仅占用 1 034 MB 显存. 考虑到实际应用中性能差距往往可以通过增加训练数据予以弥补, 所以一定程度上的精度差距可以接受.

定性结果见图 11. 定性来看, 模型对相对规则的文字图片识别结果较好, 但对长文字和较为不规则的艺术字识别性能不够可靠. 定量上看, 该模型的性能与常用的方法[22−23, 50−51] 有一定可比性, 较新提出的封闭集方法存在一定合理差距. 我们的方法在最大的 IIIT5K 集上性能较好, 但由于上下文建模的缺失, 在上下文依赖较强的 IC03[44]、IC13[45] 数据集上有一定劣势. 这个问题可以通过引入一个可选的上下文模块[9, 53] 解决.

总的来说, 虽然该模型的封闭集识别性能相较 SOTA 有所欠缺, 但由于其对新字符的处理能力和较低的资源占用, 该方法仍然在开放场景, 如网络图片和多语言路牌识别场景下有一定实用性.

[2] 注意, 字符在特征空间的区域可能有交集.


图 11　封闭集上的识别结果展示

Fig. 11　Sample results from the close-set benchmark

## 4　结束语

本文提出了一种基于自适应字符部件表示的开放集文字识别框架, 有效地提升了开放集文字识别的性能. 模型整体框架能够缓解形近字的混淆问题, 而局部性约束有效地解决了该框架存在的训练不稳定的问题. 实验结果表明, 本文提出的方法在开放集、传统闭集文字识别任务上均具有良好的性能. 特别的, 本文模型在相近的语言上展示出了较好的开放集识别性能, 可有效应用于数据变化较快的网络图片文字识别任务和数据多样性较广的多语言场景识别任务. 未来可以针对跨语系语种 (语言文字形态差异大) 迁移能力有限等局限性, 开展进一步研究工作.

## References

1　Li Wen-Ying, Cao Bin, Cao Chun-Shui, Huang Yong-Zhen. A deep learning based method for bronze inscription recognition. *Acta Automatica Sinica*, 2018, **44**(11): 2023−2030
(李文英, 曹斌, 曹春水, 黄永祯. 一种基于深度学习的青铜器铭文识别方法. 自动化学报, 2018, **44**(11): 2023−2030)

2　Zheng T L, Chen Z N, Huang B C, Zhang W, Jiang Y G. MRN: Multiplexed routing network for incremental multilingual text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18598−18607

3　Ma Si-Liang, Xu Yong. Calligraphy character recognition method driven by stacked model. *Acta Automatica Sinica*, 2024,

**50**(5): 947−957
(麻斯亮, 许勇. 叠层模型驱动的书法文字识别方法研究. 自动化学报, 2024, **50**(5): 947−957)

4 Zhang Yi-Kang, Zhang Heng, Liu Yong-Ge, Liu Cheng-Lin. Oracle character recognition based on cross-modal deep metric learning. *Acta Automatica Sinica*, 2021, **47**(4): 791−800
(张颐康, 张恒, 刘永革, 刘成林. 基于跨模态深度度量学习的甲骨文字识别. 自动化学报, 2021, **47**(4): 791−800)

5 Zhang C H, Gupta A, Zisserman A. Adaptive text recognition through visual matching. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 51−67

6 Souibgui M A, Fornés A, Kessentini Y, Megyesi B. Few shots are all you need: A progressive learning approach for low resource handwritten text recognition. *Pattern Recognition Letters*, 2022, **160**: 43−49

7 Kordon F, Weichselbaumer N, Herz R, Mossman S, Potten E, Seuret M, et al. Classification of incunable glyphs and out-of-distribution detection with joint energy-based models. *International Journal on Document Analysis and Recognition*, 2023, **26**(3): 223−240

8 Liu C, Yang C, Qin H B, Zhu X B, Liu C L, Yin X C. Towards open-set text recognition via label-to-prototype learning. *Pattern Recognition*, 2023, **134**: Article No. 109109

9 Liu C, Yang C, Yin X C. Open-set text recognition via character-context decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 4513−4522

10 Liu C, Yang C, Yin X C. Open-set text recognition via shape-awareness visual reconstruction. In: Proceedings of the 17th International Conference on Document Analysis and Recognition. San José, USA: Springer, 2023. 89−105

11 Yu H Y, Chen J Y, Li B, Ma J, Guan M N, Xu X X, et al. Benchmarking Chinese text recognition: Datasets, baselines, and an empirical study. arXiv: 2112.15093, 2021.

12 Wan Z Y, Zhang J L, Zhang L, Luo J B, Yao C. On vocabulary reliance in scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11422−11431

13 Zhang J Y, Liu C, Yang C. SAN: Structure-aware network for complex and long-tailed Chinese text recognition. In: Proceedings of the 17th International Conference on Document Analysis and Recognition. San José, USA: Springer, 2023. 244−258

14 Yao C, Bai X, Shi B G, Liu W Y. Strokelets: A learned multi-scale representation for scene text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 4042−4049

15 Seok J H, Kim J H. Scene text recognition using a Hough forest implicit shape model and semi-Markov conditional random fields. *Pattern Recognition*, 2015, **48**(11): 3584−3599

16 Li B C, Tang X, Qi X B, Chen Y H, Xiao R. Hamming OCR: A locality sensitive hashing neural network for scene text recognition. arXiv: 2009.10874, 2020.

17 Wang T, Xie Z, Li Z, Wang T, Xie Z, Li Z, et al. Radical aggregation network for few-shot offline handwritten Chinese character recognition. *Pattern Recognition Letters*, 2019, **125**: 821−827

18 Cao Z, Lu J, Cui S, Zhang C S. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition*, 2020, **107**: Article No. 107488

19 Chen J Y, Li B, Xue X Y. Zero-shot Chinese character recognition with stroke-level decomposition. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: IJCAI, 2021. 615−621

20 Zu X Y, Yu H Y, Li B, Xue X Y. Chinese character recognition with augmented character profile matching. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM, 2022. 6094−6102

21 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(11): 2298−2304

22 Liao M H, Zhang J, Wan Z Y, Xie F M, Liang J J, Lyu P Y, et al. Scene text recognition from two-dimensional perspective. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu, USA: AAAI Press, 2019. 8714−8721

23 Baek J, Kim G, Lee J, Park S, Han D, Yun S, et al. What is wrong with scene text recognition model comparisons? Dataset and model analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 4714−4722

24 Yang Chun, Liu Chang, Fang Zhi-Yu, Han Zheng, Liu Cheng-Lin, Yin Xu-Cheng. Open set text recognition technology. *Journal of Image and Graphics*, 2023, **28**(6): 1767−1791
(杨春, 刘畅, 方治屿, 韩铮, 刘成林, 殷绪成. 开放集文字识别技术. 中国图象图形学报, 2023, **28**(6): 1767−1791)

25 He J, Chen J N, Lin M X, Yu Q H, Yuille A. Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 11259−11268

26 Pourpanah F, Abdar M, Luo Y X, Zhou X L, Wang R, Lim C P, et al. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(4): 4051−4070

27 Zhang J S, Du J, Dai L R. Radical analysis network for learning hierarchies of Chinese characters. *Pattern Recognition*, 2020, **103**: Article No. 107305

28 He S, Schomaker L. Open set Chinese character recognition using multi-typed attributes. arXiv: 1808.08993, 2018.

29 Huang Y H, Jin L W, Peng D Z. Zero-shot Chinese text recognition via matching class embedding. In: Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer, 2021. 127−141

30 Wang W C, Zhang J S, Du J, Wang Z R, Zhu Y X. DenseRAN for offline handwritten Chinese character recognition. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). Niagara Falls, USA: IEEE, 2018. 104−109

31 Chen S, Zhao Q. Divide and conquer: Answering questions with object factorization and compositional reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 6736−6745

32 Geng Z G, Wang C Y, Wei Y X, Liu Z, Li H Q, Hu H. Human pose as compositional tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 660−671

33 Zhang H, Li F, Liu S L, Zhang L, Su H, Zhu J, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

34 Chng C K, Liu Y L, Sun Y P, Ng C C, Luo C J, Ni Z H, et al. ICDAR2019 robust reading challenge on arbitrary-shaped text-RRC-ArT. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia: IEEE, 2019. 1571−1576

35 Sun Y P, Ni Z H, Chng C K, Liu Y L, Luo C J, Ng C C, et al. ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR).

Sydney, Australia: IEEE, 2019. 1557−1562

36 Yuan T L, Zhu Z, Xu K, Li C J, Mu T J, Hu S M. A large Chinese text dataset in the wild. *Journal of Computer Science and Technology*, 2019, **34**(3): 509−521

37 Shi B G, Yao C, Liao M H, Yang M K, Xu P, Cui L Y, et al. ICDAR2017 competition on reading Chinese text in the wild (RCTW-17). In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan: IEEE, 2017. 1429−1434

38 Nayef N, Patel Y, Busta M, Chowdhury P N, Karatzas D, Khlif W, et al. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition-RRC-MLT-2019. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia: IEEE, 2019. 1582−1587

39 Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Synthetic data and artificial neural networks for natural scene text recognition. arXiv: 1406.2227, 2014.

40 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE Computer Society, 2016. 2315−2324

41 Mishra A, Alahari K, Jawahar C V. Scene text recognition using higher order language priors. In: Proceedings of the British Machine Vision Conference. Surrey, UK: BMVA Press, 2012. 1−11

42 Risnumawan A, Shivakumara P, Chan C S, Tan C L. A robust arbitrary text detection system for natural scene images. *Expert Systems With Applications*, 2014, **41**(18): 8027−8048

43 Wang K, Babenko B, Belongie S. End-to-end scene text recognition. In: Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE Computer Society, 2011. 1457−1464

44 Lucas S M, Panaretos A, Sosa L, Tang A, Wong S, Young R, et al. ICDAR 2003 robust reading competitions: Entries, results, and future directions. *International Journal of Document Analysis and Recognition*, 2005, **7**(2−3): 105−122

45 Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda L G I, Mestre S R, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, USA: IEEE Computer Society, 2013. 1484−1493

46 Geng C X, Huang S J, Chen S C. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(10): 3614−3631

47 Yan R J, Peng L R, Xiao S Y, Yao G. Primitive representation learning for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 284−293

48 Bhunia A K, Sain A, Kumar A, Ghose S, Chowdhury P N, Song Y Z. Joint visual semantic reasoning: Multi-stage decoder for text recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 14920−14929

49 Fang S C, Mao Z D, Xie H T, Wang Y X, Yan C G, Zhang Y D. ABINet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(6): 7123−7141

50 Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK: ACM, 2018. 71−79

51 Atienza R. Vision transformer for fast and efficient scene text recognition. In: Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer, 2021. 319−334

52 Zhang H, Luo G Y, Kang J, Huang S, Wang X, Wang F Y. GLaLT: Global-local attention-augmented light transformer for scene text recognition. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2023.3239696

53 Fang S C, Xie H T, Wang Y X, Mao Z D, Zhang Y D. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 7098−7107

**刘 畅** 吕勒奥理工大学博士后. 2024 年获得北京科技大学博士学位. 主要研究方向为小样本学习, 文本识别和文本检测.
E-mail: lasercat@gmx.us
(**LIU Chang** Postdoctoral at Luleå University of Technology. He received his Ph.D. degree from University of Science and Technology Beijing in 2024. His research interest covers few-shot learning, text recognition and text detection.)

**杨 春** 北京科技大学副教授. 2018 年获得北京科技大学博士学位. 主要研究方向为模式识别, 计算机视觉, 文档分析与识别. 本文通信作者.
E-mail: chunyang@ustb.edu.cn
(**YANG Chun** Associate professor at University of Science and Technology Beijing. He received his Ph.D. degree from University of Science and Technology Beijing in 2018. His research interest covers pattern recognition, computer vision, document analysis and recognition. Corresponding author of this paper.)

**殷绪成** 北京科技大学教授. 2006 年获得中国科学院自动化研究所博士学位. 主要研究方向为模式识别, 文字识别, 计算机视觉, 人工智能芯片, 工业智能与工业软件技术及应用.
E-mail: xuchengyin@ustb.edu.cn
(**YIN Xu-Cheng** Professor at University of Science and Technology Beijing. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2006. His research interest covers pattern recognition, text recognition, computer vision, AI chips, industrial intelligence and industrial software technology and applications.)