研究与开发

特征增强和双线性特征向量融合的移动端工业货箱文本检测

胡海洋[1,2]，厉泽品[1,2]，李忠金[1,2]
（1. 杭州电子科技大学计算机学院，浙江 杭州 310018;
2. 浙江省脑机协同智能重点实验室，浙江 杭州 310018）

摘　要：在实际工业环境下，光线昏暗、文本不规整、设备有限等因素，使得文本检测成为一项具有挑战性的任务。针对此问题，设计了一种基于双线性操作的特征向量融合模块，并联合特征增强与半卷积组成轻量级文本检测网络 RGFFD（ResNet18+GhostModule+特征金字塔增强模块（feature pyramid enhancement module，FPEM）+ 特征融合模块（feature fusion module，FFM）+可微分二值化（differenttiable binarization，DB））。其中，Ghost 模块内嵌特征增强模块，提升特征提取能力，双线性特征向量融合模块融合多尺度信息，添加自适应阈值分割算法提高 DB 模块分割能力。在实际工厂环境下，采用嵌入式设备 UP2 board 对货箱编号进行文本检测，RGFFD 检测速度达到 6.5 f/s。同时在公共数据集 ICDAR2015、Total-text 上检测速度分别达到 39.6 f/s 和 49.6 f/s，在自定义数据集上准确率达到 88.9%，检测速度为 30.7 f/s。

关键词：文本检测；半卷积；特征向量融合；特征增强；特征融合
中图分类号：TN929.5
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022139

Feature enhancement and bilinear feature vector fusion for text detection of mobile industrial containers

HU Haiyang[1,2], LI Zepin[1,2], LI Zhongjin[1,2]
1. School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
2. Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou 310018, China

Abstract: In the real factory environment, due to factors such as dim light, irregular text, and limited equipment, text detection becomes a challenging task. Aiming at this problem, a feature vector fusion module based on bilinear operation was designed and combined with feature enhancement and semi-convolution to form a lightweight text detection network RGFFD (ResNet18 + Ghost Module + FPEM(feature pyramid enhancement module)) + FFM(feature

收稿日期：2022−01−03；**修回日期：**2022−06−10
通信作者：李忠金，lizhongjin@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.61572162，No.61802095）；浙江省重点研发计划项目（No.2018C01012）；浙江省自然科学基金资助项目（No.LQ17F020003）
Foundation Items: The National Natural Science Foundation of China (No.61572162, No.61802095), The Zhejiang Provincial Key Science and Technology Project (No.2018C01012), The Zhejiang Provincial National Science Foundation of China (No.LQ17F020003)

fusion module) + DB (differentiable binarization)). Among them, the Ghost module was embedded with a feature enhancement module to improve the feature extraction capability, the bilinear feature vector fusion module fused multi-scale information, and an adaptive threshold segmentation algorithm was added to improve the segmentation capability of the DB module. In the real industrial environment, the RGFFD detection speed reached 6.5 f/s, when using the embedded device UP2 board for text detection of container numbers. At the same time, the detection speed on the public datasets ICDAR2015 and Total-text reached 39.6 f/s and 49.6 f/s, respectively. The accuracy rate on the custom dataset reached 88.9%, and the detection speed was 30.7 f/s.

Key words: text detection, semi-convolution, feature vector fusion, feature enhancement, feature fusion

0 引言

现实场景中文字承载的高级语义信息能够帮助人们更好地理解周围世界，场景文本检测作为场景文本读取的关键组成部分，一直是计算机视觉领域的热门研究方向，例如工业自动化、自动驾驶和盲人辅助等。

早期的文字检测技术使用传统的模式识别方法，其主要分为两种：一是以连通区域分析为核心技术的文字检测方法[1-3]，二是 Minetto 等[4]提出的以滑动窗口为核心技术的文字检测方法。传统的模式识别方法一般包含 4 个步骤：字符候选区域生成；候选区域滤除；文本行构造；文本行验证。然而烦琐的检测步骤导致文字检测的实时性差，同时准确率得不到保证。

随着计算机视觉和模式识别领域的发展，卷积神经网络[5]开始崭露头角，逐步成为主流的目标特征提取网络。因此先从训练数据中提取有效的文本特征并建立模型，然后将模型运用于实际环境，并通过文本检测算法完成文本检测任务的深度学习方式逐渐成为主流。

目前，基于深度学习的文本检测方法可以分为两类：一种是基于目标检测方法的回归检测算法，目标检测框架由 SSD[6]、FasterRCNN[7]、ResNet[8]等进行针对文字特性的改进得到，这类方法的主要特点是通过回归水平矩形框（anchor）、旋转矩形框以及四边形等形状获得文字检测结果；另一种是基于文本分割方法进行文本检测，此类方法主要借鉴语义分割的思路，将文本像素分到不同的实例中，并通过一些后处理方法获得文本像素级别的定位结果，可以精确定位任意形状的文字，该类方法主要有 Liao 等[9]提出的可微分二值化（differentiable binarization，DB）后处理算法等。

与传统文字检测方法相比，深度学习的文字检测[10]已经简化了很多步骤，但是网络的加深带来了更大的计算量。ResNet50 具有大约 25.6 MB 大小的参数，以及需要 4.1×10^9 FLOPS（floating point operations per second，每秒浮点运算次数）的计算量处理一张 224 dpi × 224 dpi 的图像。因此，深度神经网络设计的最新趋势是探索可移植、高效、轻量的网络架构，并为移动设备提供可接受的性能。Han 等[11]采用裁剪的方法，对不重要的权值进行裁剪，以此提升网络性能。Howard 等[12]利用深度卷积和逐点卷积相结合构建了 MobileNet 轻量网络架构，在与 VGG16 精度相同的情况下，参数量和计算量减少了 2 个数量级。ShuffleNet[13]改进了通道的 shuffle 操作，增强了轻量网络的性能。

工厂中的货箱运输环境如图 1 所示，其中，开发板、显示器、摄像机部署在叉车上，叉车行驶的平均速度为 3 m/s 左右，摄像机拍摄货箱编号。只有当图片的检测帧率为 12 f/s（frames per second，每秒传输帧数）以上时，显示器才可以清楚地显示每张图片的检测结果，而处于移动端的文本检测，则需要达到更高的检测帧率才能满足要求，因此需要搭建轻量网络架构，而与轻量

网络 MobileNet、ShuffleNet 文本检测方法相比，工厂环境下的文本检测有其复杂性和特殊性：它所处的运输环境背景混乱、光线变化频繁、文本不规整等。因此在工厂环境下轻量网络文本检测方法无法在保证实时性的同时，达到较高的准确率。

图 1　工厂中的货箱运输环境

针对在工厂货箱运输场景中存在的问题，本文提出一种基于轻量级网络的货箱编号检测方法。文本检测模型如图 2 所示，首先，使用 ResNet18 作为基础网络架构，用改进的 Ghost 模块替换基础残差模块，其中，Ghost 模块嵌入文献[14]中提出的轻量级特征增强技术 Squeeze- and-Excitation，对部分卷积后的特征进行重标定，提高重要特征的权重。其次，采用双分支结构，第一分支使用文献[15]中提出的特征金字塔增强模块（feature pyramid enhancement module，FPEM）提取图像高级和低级信息，第二分支利用本文提出的双线性特征融合向量模块融合不同尺度的特征向量，增强尺度多变的文本特征表达能力。而后特征融合模块（feature fusion module，FFM）级联所有特征向量。最后，采用 DB 语义分割算法获得最终结果，其中，修改损失函数为文献[16]中提出的 DiceLoss 和 MaskLoss。同时在推理阶段采用自适应阈值分割算法替换固定阈值，更能适应工厂环境的光线变化。

为了能够训练新型轻量网络框架并评估它的优势，本文创建了一个复杂工厂环境下的货箱文字数据集，数据集中包含了不同种类的货箱，不同视角下、不同形状的文字。实验表明，本文提出的新型轻量级网络框架 RGFFD（ResNet18+GhostModule+特征金字塔增强模块（feature pyramid enhancement module，FPEM）+ 特征融合模块（feature fusion module，FFM）+可微分二值化（differenttiable binarization，DB ））在实时性和精确度方面都优于其他的网络框架。本文的主要贡献为以下 3 点。

- 提出了新型的轻量网络架构解决实际工业场景中移动设备的文字检测，并达到了可观的精确度。

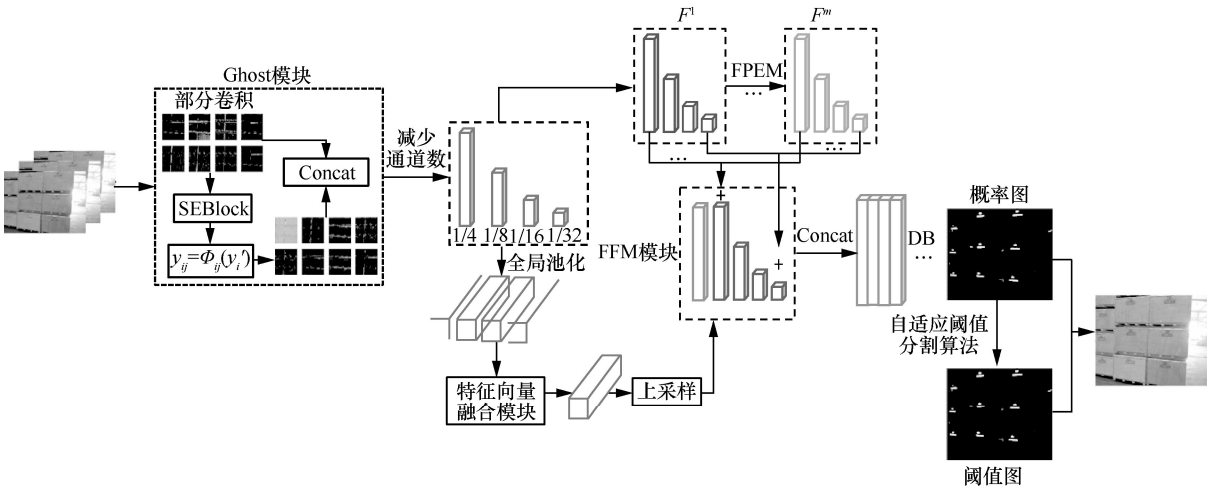


图 2　文本检测模型

- 制作了一个货箱文字数据集对模型进行训练和评估，并最终实现在线部署。
- 在自定义数据集上，本文的模型在识别精度和泛化能力上都超过了主流的文字检测方法。

1 相关工作

近年来，设计轻量高效的神经网络架构一直是热门的研究领域。VGGNets[17]模型表明，增加网络的深度可以显著提高网络的学习特征的能力。同时 Batch Normalization 操作通过调整输入每一层的分布，提升了深层网络学习过程的稳定性，产生了更平滑的优化曲面。ResNet 通过使用残差模型构建更深层次和更强大的网络。而网络的加深带来了更大的计算量，Szegedy 等[18]提出采用逐点卷积(1×1)减少参数计算量，并在同一层级利用不同大小的卷积核提取图像不同维度的信息，在保证模型质量的前提下，减少参数量。InceptionV2[19]则通过恰当地分解卷积与积极地正则化尽可能地利用有效的运算。虽然网络的计算成本得到了降低，但是仍无法在一些嵌入式的设备中运行。MobileNet 的提出使网络的计算成本出现了大幅度的下降，其主要思想是对每个通道单独利用卷积核进行卷积操作，然后利用逐点卷积融合特征，有效替代传统的卷积层，然而精确度却无法得到保障。MobileNetV2[20]对此进行了扩展，引入线性瓶颈和反向残差结构，主要思想是在深度卷积之前增加逐点卷积操作，使得特征提取能够在高维运行。Howard 等[21]提出 MobileNetV3，其添加了轻量级注意力机制，将 swish 替换为 h-swish，以更少的计算量获得更好的性能。ShuffleNet 提出逐点组卷积，有效地降低了因为逐点卷积而形成的通道之间的约束，同时采用通道混洗方法提高了通道组之间的信息流通，提高了信息的表示能力。SqueezeNet 广泛使用 1×1 卷积，采用 Squeeze-and-Excitation 模块减少参数数量，提升网络的特征提取能力。Lin 等[22]提出了特征金字塔网络（feature pyramid network，FPN），通过提取并融合上下文信息，使小物体的检测更准确，但由于网络计算复杂，参数量大无法满足实时性要求。Wang 等[15]提出了 FPEM，采用可分解卷积，降低网络计算量，并通过级联的方式完成高低级特征的提取，同时利用 FFM 特征融合模块融合不同层次的特征，在保证精度的同时参数量仅为 FPN 的 1/5。

采用轻量网络架构提取图像文本特征，最后利用文本检测算法绘制文本框，文本检测方法大致可以分为两类：基于回归的方法和基于分割的方法。

基于回归的方法是直接回归文本边界框，准确定位文本。Liao 等[23]提出的 TextBoxes 基于 SSD 修改了 anchor 和卷积核的尺度，用于文本检测。Liao 等[24]提出的 TextBoxes++应用四边形回归来检测多方向文本。Tian 等[25]首次提出将文字区域分割成一系列小尺度的候选框，同时引入循环神经网络（recurrent neural network，RNN），增加了检测的精度，但只能检测水平方向的文字。Shi 等[26]采用角度概念，切割图片为段，使用 link 检测将属于同一文本的段进行连接，以处理长文本实例。Liao 等[27]通过使用旋转不变特征进行分类，使用旋转敏感特征进行回归，将分类和回归解耦，以便在多方向和长文本实例上取得更好的效果。Zhou 等[28]提出了全卷积操作，并直接生成预测文本框，利用局部感知非最大抑制（locality- aware non maximum suppression，LNMS）产生最终结果，实现端到端的文本检测。然而这些方法都是为四边形文本检测而设计的，无法识别任意形状的文本。

与基于回归方法不同，基于分割的方法通常结合像素级预测和后处理算法得到边界框，可以检测不规则形状的文本。Deng 等[29]提出 Pixel-Link 概念，对输入图像执行文本和非文本预测以及链接预测，然后通过后处理以获取文本框并过滤噪声，分隔不同的文本实例。Wang 等[30]通过分割具有不同规模内核的文本实例，并使用

渐进式尺度扩展算法获得最终文本框。然而 PSENet 因网络计算量大，网络检测文本的实时性大大降低。Wang 等[15]引入特征金字塔增强模块，并通过级联的方式提取多级信息，而后采用特征融合模块融合多尺度信息，最后通过像素聚合模块预测相似性向量聚合文本像素，在不降低精度的情况下，减少网络的计算量。DBNet 提出了一个可微分二值化模块预测收缩区域，并且收缩区域以恒定的膨胀率扩张，最终获得文本框。Tian 等[31]提出了学习形状感知嵌入（learning shape-aware embedding，LSAE）方法，将图像像素映射到特征空间中，对属于同一文本实例的像素进行聚类，很好地分割相邻的文本，并且可以检测长文本。

2　本文方法

2.1　特征增强的 Ghost 模块

深层卷积神经网络中通常由大量的卷积操作组成，这就需要大量的计算量，大多数方法采用逐点卷积处理跨通道的特征，然后采用深度卷积处理空间信息，以此减少网络的计算量。普通的卷积操作会产生大量的冗余信息，特征冗余如图 3 所示，其中有部分是相似的，因此不需要一个接一个地生成这些冗余的带有大量参数运算的特征映射，而是将相似的特征映射通过某种简单的线性操作进行获取，以此减少计算量。

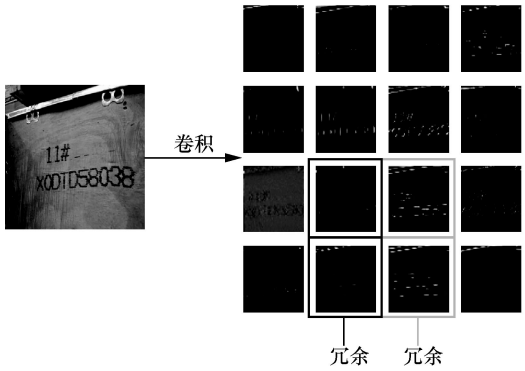


图 3　特征冗余

为了解决特征冗余的问题，避免不必要的卷积操作，本文采用改进的 Ghost 模块替换残差网络的基础残差 3×3 卷积模块。常规 Ghost 模块主要由两种关键技术组成，分别是部分卷积和 Chollet 等[32]提出的 DepthWise 卷积。相较于常规卷积，部分卷积只有部分特征图是利用卷积核生成的。Depthwise 操作的一个卷积核只负责特征映射图的一个通道，一个通道只与一个卷积核进行卷积操作。最后将线性变换的特征图与原先的特征图进行拼接操作，转换为普通卷积操作后通道数相同的特征图。

本文为了使网络能够在训练和测试阶段获得更加完整的文本图像特征，且只提高少许网络的复杂性，因此采用轻量级特征增强技术 Squeeze-and-Excitation，Squeeze-and-Excitation 模块如图 4 所示，通过显式地建模卷积特征通道之间的相互依赖性提高网络的性能。

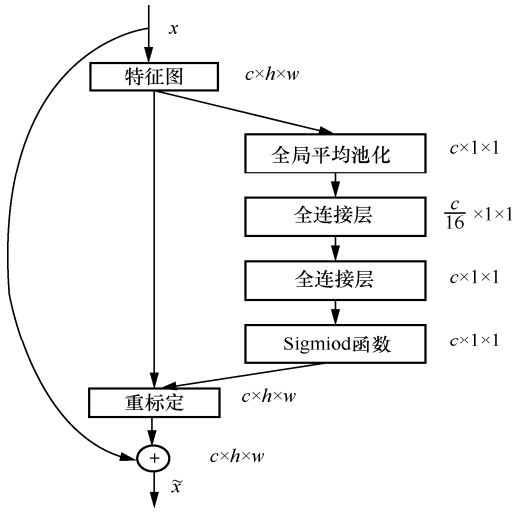
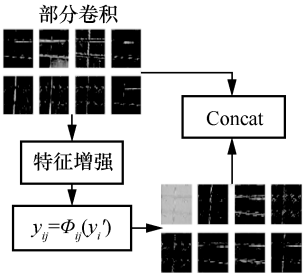


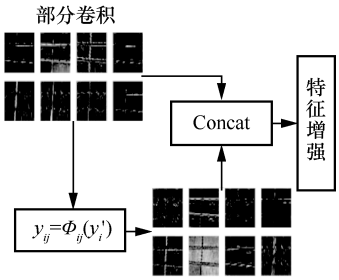
图 4　Squeeze-and-Excitation 模块

Ghost 模块改进方法示意图如图 5 所示，其中改进方法 1 将特征增强模块 Squeeze-and-Excitation 嵌入 Ghost 模块中，在部分卷积之后进行特征增强。改进方法 2 选择在部分卷积和 DepthWise 卷积后进行特征增强。相较于方法 2，方法 1 在进行特征增强时所需要的网络计算量更少，Ghost 模块部分卷积操作只产生通道数为 $N/2$ 的特征

图，因此只需要对一半的特征图进行特征增强。而在工厂环境下，叉车运行速度较快，需要网络检测图片的速率达 20 f/s 以上，才可以清晰地显示图片。因此本文选择改进方法 1，减少网络计算量，提升网络检测速率。



(a) 方法1



(b) 方法2

图 5　Ghost 模块改进方法示意图

2.2　双线性特征向量融合模块

工厂环境复杂，在不同视角存在大量尺度不同的文本，因此为了融合不同尺度的文本特征，增强尺度多变的文本特征表达能力，本文提出了双线性特征向量融合模块。

长短期记忆（long short-term memory，LSTM）[33]是特征向量融合模块的核心成分，LSTM 首先被应用在文本识别。特征融合模块细节如图 6 所示，本文特征向量融合模块仅由 4 个特征向量组合而成，因此本文舍弃了长期记忆，采用简单的线性操作融合以前的输入信息，将不同层次的特征向量依次输入特征向量融合模块中。其中，tanh 网络创建一个可以存储的向量 $\boldsymbol{C}_t$，sigmoid 网络层为此向量中的每个值输出一个 0～1 的数值 i_t，决定要存储哪些状态值，最后通过简单的线性操作进行融合。通过训练，可以使最后一个特征向量对应的输出存储了所有特征向量重要的信息。因为本文提出的双线性特征向量融合模块只需要经过简单的线性操作，就可以完成不同尺度特征向量的融合，因此在不影响实时性的同时，增加了网络检测的精确率。双线性特征向量融合模块公式化为：

$$i_t=\sigma\left(W_i\times\left[h_{t-1},\ x_t\right]+b_i\right) \tag{1}$$

$$C_t=\tanh\left(W_C\times\left[h_{t-1},\ x_t\right]+b_C\right) \tag{2}$$

$$h_t=\mathrm{Conv}\left(h_{t-1},\ x_t\right)+i_t\times C_t \tag{3}$$

其中，i_t 为 sigmoid 网络层的输出，C_t 为 tanh 网络层的输出，σ 为 sigmoid 网络层，h_{t-1} 为上一次的输出，x_t 为第 t 次的输入。W_i、W_C、b_i、b_C 为权重。

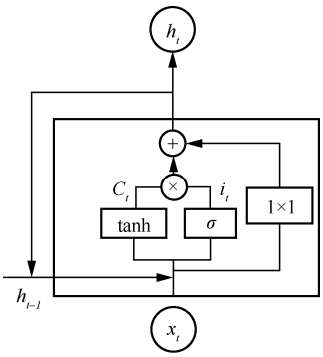


图 6　特征融合模块细节

2.3　特征金字塔和特征融合

FPN 采用特征金字塔模型，对高低层的语义信息进行融合，提高网络检测不同尺度的目标的精度，然而特征融合采用上采样、逐个位相加、向量拼接技术，大大增加了网络计算量，无法保证网络的实时性。FPEM 能够通过融合低级和高级信息增强不同尺度的特征。FPEM 模块细节如图 7 所示，FPEM 是可级联的模块，随着级联层数的增加，不同尺度的特征图会得到更充分的融合，特征图的感受野也随之增大。此外，因为 FPEM 是通过可分解卷积构建的，其计算开销非常小，仅为 FPN 的 1/5 左右。

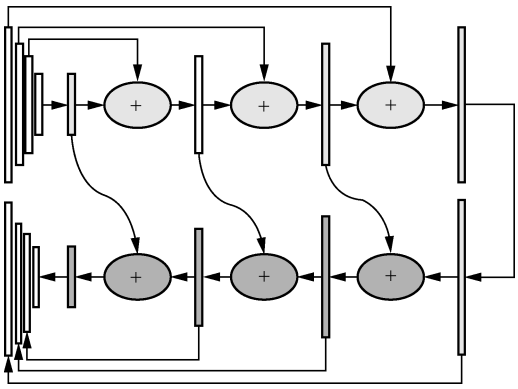


图 7　FPEM 模块细节

FFM 模块示意图如图 8 所示，特征融合模块 FFM 对 FPEM 级联产生的不同层次的特征 $F^1,F^2,\cdots,F^m$ 进行融合。为增强不同尺度文本的特征表达能力，本文对 FFM 进行改进，将特征融合后的向量进行上采样并与原模型相级联，获得通道数为 5×128，大小为原图 1/4 的最终特征图。

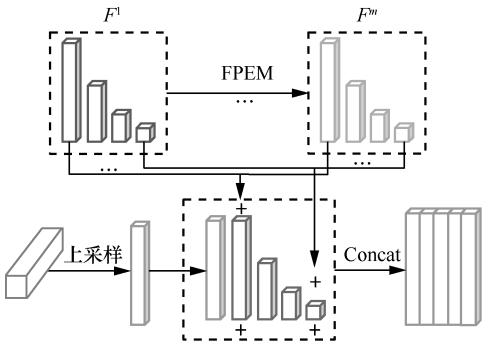


图 8　FFM 模块示意图

2.4　自适应阈值后处理 DB 算法

DBNet 采用可微分二值化处理，使阈值在训练期间能随着网络一起优化，同时基于阈值图和概率图获取近似二值图。DBNet 提供的可微的二值化计算式为：

$$B_{i,j}=\frac{1}{1+\mathrm{e}^{-k\left(P_{i,j}-T_{i,j}\right)}} \tag{4}$$

其中，$P_{i,j}$ 表示该区域有文字的概率，如果没有文字区域，$P_{i,j}$ 为 0；$T_{i,j}$ 是由网络学习到的阈值图；k 表示放大系数。

总的损失函数 L 可以表示为概率图的损失与二值图的损失与阈值图的损失的加权和：

$$L=L_s+\alpha\times L_b+\beta\times L_t \tag{5}$$

其中，L_S 是概率图的损失值，L_b 是二值图的损失值，L_t 是阈值图的损失。其中，α 和 β 分别被设置为 5 和 10。其中 L_S 和 L_b 分别使用 DiceLoss 损失函数来进行训练，DiceLoss 常用于医学图像分割，其目的就是解决前景比例太小的问题。其计算式为：

$$\text{loss}=1-\frac{2\left|\text{pred}\cap gt\right|}{\left|\text{pred}\right|+\left|gt\right|} \tag{6}$$

DiceLoss 损失函数特点就是收敛速度快，且收敛速度优于交叉熵等分类损失函数。L_t 采用 Mask L1 Loss 损失函数来进行训练，通过掩模进行 L_t 损失函数的计算。其计算式为：

$$L_t=\frac{\sum_{i=1}^{n}\left|y_i-y_i^p\right|\times\text{mask}}{\left|\text{mask}\right|} \tag{7}$$

其中，L_t 损失为平均绝对差值，而后增加了 mask 掩模，对 mask 指定的区域进行 L_t 损失函数的计算，n 为膨胀后的图像区域在 mask 掩模下需要进行计算的数量总和。绝对值 $\left|y_i-y_i^p\right|$ 为预测文本框与标签的距离差值。

DB 算法检测不规整文本如图 9 所示，本文采用自适应阈值替换固定阈值获取阈值图。最终文本框的获取有 3 个步骤。

步骤 1　采用网络输出概率图，其值范围为[0,1]，DBNet 设定常数阈值（0.2）获取阈值图，本文不设定固定的阈值获取阈值图，而是通过设置九宫格掩模对概率图进行自适应阈值计算获得阈值图，自适应阈值计算式为：

$$t=\frac{\sum_{i=-1}^{1}\sum_{j=-1}^{1}p_{x+i,y+j}}{9} \tag{8}$$

$$f'_{x,y}=\begin{cases}255 & f_{x,y}\geqslant t\\ 0, & \text{其他}\end{cases} \tag{9}$$

其中，x、y 为图像中的坐标，$f_{x,y}$ 是坐标为(x,y)的像素值，t 为自适应阈值算法求得的阈值，$p_{x+i,y+j}$ 为原图坐标为$(x+i,y+j)$的像素值，$f_{x,y}$

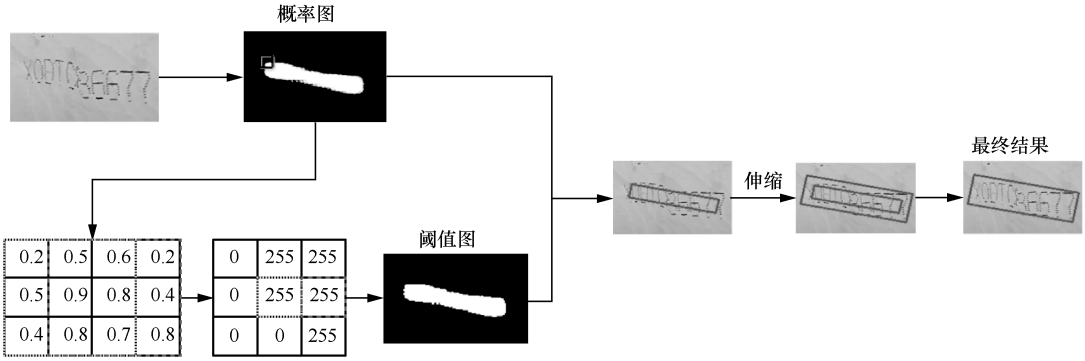


图 9　DB 算法检测不规整文本

是坐标为(x, y)的阈值。

步骤 2 从概率图和阈值图中获取连接区域。

步骤 3 扩大文本区域，进行膨胀，扩大的倍数为$D' = (A' \times r') / L'$，A'为多边形的面积，L'为多边形的面积，r'被设定为 1.8。

3　实验

3.1　数据集

利用本文的网络 RGFFD，在 ICDAR2015、Total-text 以及本文采集并设计的数据集上进行实验和评估，具体如下。

- ICDAR2015 数据集：主要面向自然场景的文本检测，有 1 000 张训练图片和 500 张测试图片，其中图片的像素大小为 1 280 dpi × 720 dpi。
- 自定义数据集：针对复杂工厂环境下货箱编码检测的问题而采集的数据集，包含了在不同形状的木板箱子上从不同角度采集的 3 000 多张的文本图片，其中文本样式各异，并且倾斜角度不同，文字背景信息复杂。文本框区域用矩形的 4 个坐标点来进行记录。其中图片的像素大小为 1 280 dpi×720 dpi。
- Total-text 数据集：包含各种形状文本的数据集，包括水平、多方向和曲线文本实例，由 1 255 张训练图像和 300 张测试图像组成，文本实例以单词级标注。

3.2　实验环境和模型的训练

在本实验中，实验数据训练在 64 位的 Windows 系统上，内存为 16 GB，CPU 是 8 核，显卡为 RTX2060。所有的实验都是通过 PyTorch 深度学习框架完成的，模型的循环次数（epoch）为 500，学习率被设置为 0.001，每次训练选取样本数量为 8。

实验检测结果在开发板 UP2 board 中运行，开发板具体配置为 Windows10 系统，内存为 8 GB，容量大小为 64 GB，处理器为 N4200。

3.3　网络参数设置

本实验的反向传播选用的是 Adam 优化算法，Adam 优化器是一个寻找全局最优点的优化算法，算法引入了二次梯度校正。

3.4　评价指标

由于常规的视频需要达到 12 f/s，显示器才可以清楚地显示每张图片，而处于移动端的文本检测，叉车平均行驶速度为 3 m/s，则需要更高的帧率才能满足要求。经实验可知，当帧率达到 20 f/s 或以上时，图片可以流畅地显示，因此本文算法的性能由精确率、召回率、帧率（检测速度）（f/s）以及检测速度与 20 的差值这 4 个指标来衡量。

4　结果比较

4.1　模型对比实验

本文方法与近年出现的其他方法在多个数据集上进行了对比，在公共数据集中算法的实验结果由算法作者提供，自定义数据集中算法的实验结果

由本文作者实现。本文方法在 ICDAR2015 数据集中与不同文本检测方法的比较见表 1。由实验结果可知，TextFuseNet 在 ICDAR2015 数据集中的检测精确率和召回率已与 SOTA 相近，但是网络复杂，计算参数量大，导致网络的检测速度较慢，无法满足 20 f/s 的要求。PAN（pixel aggregation network）作为 PSENet 的改进版，将精确率和检测速度做到了很好的平衡，但是本文提出的 RGFFD 网络同样采用 PAN 的特征金字塔和特征融合模块并进行改进，在精确率上超过 PAN 0.4%，同时在召回率上与 PAN 接近，且利用改进的 Ghost 模块替换常规卷积，在速度方面更是超过 PAN 2.1 f/s，与其他网络相比本文方法检测速度达到最快。

表 1　本文方法在 ICDAR2015 数据集成中与不同文本检测方法的比较

方法	精确率	召回率	检测速度/（f·s^{-1}）	检测速度−20/（f·s^{-1}）
PAN[15]	85.5%	81.9%	37.5	17.5
EAST[28]	83.6%	73.5%	13.2	−6.8
SegLink[26]	73.1%	76.8%	13.9	−6.1
TextSnake[34]	84.9%	80.4%	1.1	−18.9
TextFuseNet[35]	**91.3%**	**88.9%**	8.3	−11.7
PSENet[30]	86.9%	84.5%	1.6	−18.4
CTPN[25]	74.2%	51.6%	7.1	−12.9
本文方法	85.9%	81.4%	**39.6**	**19.6**

注：加粗字体为每列最优的结果。

本文方法在 Total-text 数据集中与不同文本检测方法的比较见表 2，TextFuseNet 在精确率和召回率上已和 SOTA 相近，但是检测速度不满足要求，无法达到 20 f/s，而 PAN 在精确率和召回率上与 TextFuseNet 相近，且速度更是达到 39.9 f/s。本文提出的 RGFFD 网络精确率与 PAN 相近，且检测速度高于 PAN 5.7 f/s。与其他网络相比，本文方法检测速度达到最快。

表 2　本文方法在 Total-text 数据集中与不同文本检测方法的比较

方法	精确率	召回率	检测速度/（f·s^{-1}）	检测速度−20/（f·s^{-1}）
PAN[15]	87.3%	81.5%	39.9	19.9
EAST[28]	50.1%	36.2%	19.8	−0.2
SegLink[26]	30.3%	23.8%	9.1	−10.9
TextSnake[34]	82.7%	74.5%	4.7	−15.3
ATRR[36]	80.9%	76.2%	25.2	5.2
TextFuseNet[35]	**87.5%**	**83.2%**	7.1	−12.9
PSENet[30]	84.8%	79.7%	3.9	−16.1
本文方法	86.9%	78.1%	**45.6**	**25.6**

注：加粗字体为每列最优的结果。

本文方法在自定义数据集中与不同文本检测方法的比较见表 3，TextFuseNet 的精确率和召回率都显示出自己的优势，但是由于网络复杂、计算量大，网络的检测速度未能达到 20 f/s。本文的方法 RGFFD，精确率和召回率都与 TextFuseNet 相近，且检测速度达到 20 f/s 的要求，并超出 TextFuseNet 23.2 f/s。PAN 虽然检测速度也达到 20 f/s 的要求，但是精确率却没有 RGFFD 高，因此本文的方法 RGFFD 性能更优异。

表 3　本文方法在自定义数据集中与不同文本检测方法的比较

方法	精确率	召回率	检测速度/（f·s^{-1}）	检测速度−20/（f·s^{-1}）
PAN[15]	86.7%	80.6%	25.4	5.4
EAST[28]	79.2%	61.3%	9.5	−10.5
SegLink[26]	70.4%	56.8%	4.2	−15.8
TextSnake[34]	81.8%	75.2%	0.6	−19.4
PSENet[30]	87.5%	79.4%	2.1	−17.9
CTPN[25]	65.3%	49.5%	2.5	−17.5
TextFuseNet[35]	**89.5%**	**81.4%**	7.5	−12.5
本文方法	88.9%	80.5%	**30.7**	**10.7**

注：加粗字体为每列最优的结果。

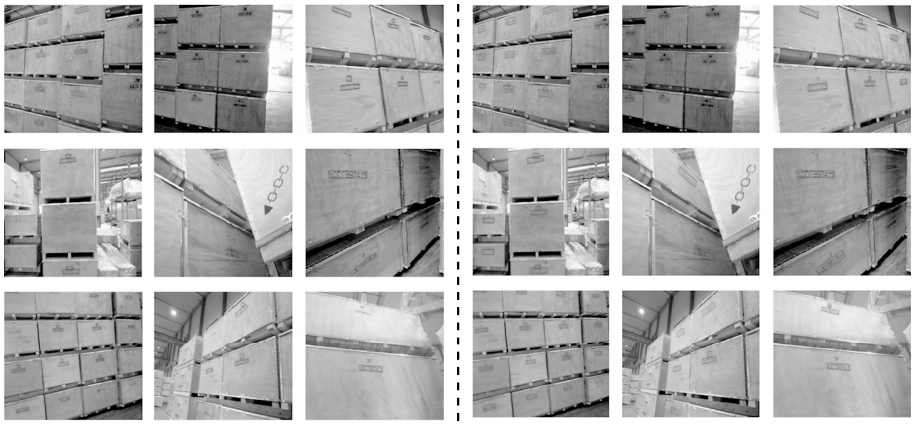
针对复杂工厂环境，将训练好的网络参数移植入开发板 UP2 board 进行检测，开发板中不同方法的效果比较见表 4。实验结果表明本文的方法检测速度最快。

表 4　开发板中不同方法效果比较

方法	检测速度/（f·s^{-1}）
PAN[15]	4.3
MobileNetV3[21]+DB[9]	5.1
本文	**6.5**

注：加粗字体为每列最优的结果。

开发板检测的文本效果如图 10 所示，图 10（a）为 MobileNetV3+DB 网络检测的结果，图 10（b）



(a) MobileNetV3+DB　　(b) 本文RGFFD

图 10　开发板检测的文本效果

为 RGFFD 的检测结果。可以看出本文的检测方法能够精确检测远处尺度较小的货箱文本，且针对不同角度，RGFFD 也能精确检测。因此，RGFFD 能够在工厂环境下高效地完成文本检测任务。

4.2　消融研究

为了验证本文不同模块在检测过程中发挥作用，针对不同模块设计了消融实验，所有消融实验均在自定义数据集上进行。

针对 Ghost 模块，设计了 3 组实验：第 1 组实验不对网络嵌入任何模块，第 2 组实验嵌入的 Ghost 模块采用改进方法 1（如图 5 所示），第 3 组实验嵌入的 Ghost 模块采用改进方法 2（如图 5 所示）。第 1 组实验、第 2 组实验与第 3 组实验形成对比，分别探究删除、不同改进方法对结果的影响，如果第 2 组实验结果优于第 1 组实验、第 3 组实验，则说明本文改进的 Ghost 模块在文本检测流程中发挥了不可或缺的作用。

本文实验的运行环境以及参数设置均与原实验相同。不同嵌入模块的实验结果对比见表 5。由实验结果可知：嵌入 Ghost 模块之后，网络的检测速度得到了提升，这说明 Ghost 模块为轻量级。而比较第 2 组实验、第 3 组实验，可以发现改进方法 1 相较于改进方法 2，计算的参数量更少，网络的检测速度高出 0.8 f/s，且精确率两者相接近，召回率高出 0.8%，进一步证明了方法 1 的有效性。

表 5　不同嵌入模块的实验结果对比

方法	精确率	召回率	检测速度/(f·s^{-1})
无	87.5%	79.2%	25.6
嵌入改进方法 1	**88.9%**	**80.5%**	**30.7**
嵌入改进方法 2	88.1%	78.7%	29.9

注：加粗字体为每列最优的结果。

此外，本文设计了消融实验验证特征向量融合模块的有效性。特征向量融合模块效果比较见表 6。由实验结果可知，嵌入特征向量融合模块虽然降低了检测速度，但是精确率和召回率得到了提升，相较于原来精确率提升了 1.1%，召回率提升了 0.3%。

表 6　特征向量融合模块效果比较

方法	精确率	召回率	检测速度/(f·s^{-1})
无	87.8%	80.2%	**32.3**
特征向量融合模块	**88.9%**	**80.5%**	30.7

注：加粗字体为每列最优的结果。

在推理阶段，本文设计了消融实验验证自适应阈值的 DB 算法的有效性。在自定义数据集中不同方法的检测结果如图 11 所示，实验结果表明采用自适应阈值分割算法获得的平均联合交叉（intersection over union，IOU）值要较固定阈值（0.2）高出 4%，而文本检测获得的 IOU 值越高，

越能判定文本框的存在，因此真阳率得到了 5.2% 的提升。因为工厂环境复杂、文本较小、光线变化频繁等因素，直接使用全局阈值分割算法容易引入噪声等不确定因素影响阈值，而自适应阈值分割算法，只获取像素点周围九宫格内的像素进行阈值计算，更好地限制了不利因素，且适应文本较小的情况。

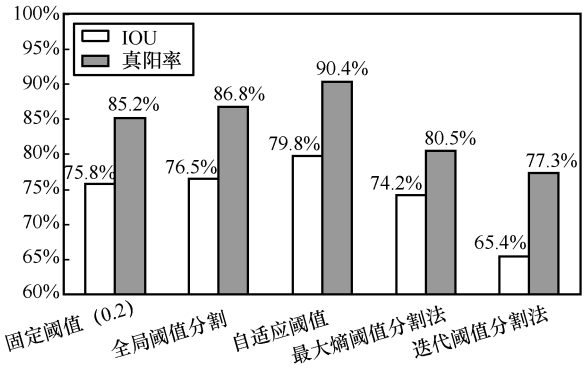


图 11　在自定义数据集中不同方法的检测结果

推理阶段检测结果如图 12 所示，展示了采用自适应阈值分割算法与固定阈值方法检测文本的实验图，图 12（a）为固定阈值（0.2）的检测结果，图 12（b）为自适应阈值分割算法的检测结果。可以发现在光照频繁变化的情况下，自适应阈值分割算法的判断能力更好，固定阈值算法会发生很多误判，因此自适应阈值分割算法在文本检测流程中发挥着不可或缺的作用。

5　结束语

本文提出的轻量级文本检测网络 RGFFD，采用改进的 Ghost 模块大幅度降低计算量的同时嵌入特征增强模块（SEBlock）提升 Ghost 模块提取特征能力，而后连接特征金字塔（FPEM）融合图像高低层语义信息，同时使用双线性特征向量融合模块，增强尺度多变的文本特征表达能力，而后采用特征融合模块（FFM）融合各特征向量。实验结果表明，在工厂环境下，RGFFD 网络运行在嵌入式设备 UP2 板中，检测速度最快且精度高，因此本文的网络性能更优异。而采用 Dice Loss 和 Mask L1 Loss 作为损失函数的 DB 算法，收敛效果更好。在推理阶段，采用自适应阈值分割算法来获取阈值的方法比直接采用固定阈值的方法，更加能够适应各种环境变化，图片检测文本框的效果也更加精准。

参考文献：

[1] HUANG W L, LIN Z, YANG J C, et al. Text localization in natural images using stroke feature transform and text covariance descriptors[C]//Proceedings of 2013 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 1241-1248.

（a）固定阈值（0.2）　　（b）自适应阈值分割算法

图 12　推理阶段检测结果

[2] NEUMANN L, MATAS J. Real-time lexicon-free scene text localization and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1872-1885.

[3] MATAS J, CHUM O, URBAN M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004, 22(10): 761-767.

[4] MINETTO R, THOME N, CORD M, et al. T-HOG: an effective gradient-based descriptor for single line text regions[J]. Pattern Recognition, 2013, 46(3): 1078-1090.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[M]//ComputerVision–ECCV2016. Cham: Springer International Publishing, 2016: 21-37.

[7] ZHONG Z Y, SUN L, HUO Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches[J]. International Journal on Document Analysis and Recognition (IJDAR), 2019, 22(3): 315-327.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[9] LIAO M, WAN Z, YAO C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Piscataway: IEEE Press, 2020: 11474-11481.

[10] 苏赋, 吕沁, 罗仁泽. 基于深度学习的图像分类研究综述[J]. 电信科学, 2019,35(11): 58-74.
SU F, LV Q, LUO R Z. Review of image classification based on deep learning[J]. Telecommunications Science, 2019, 35(11): 58-74.

[11] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 1577-1586.

[12] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[13] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of 2018 IEEE/CVF Conference on ComputerVision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on ComputerVision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[15] WANG W H, XIE E Z, SONG X G, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 8439-8448.

[16] MILLETARI F, NAVAB N, AHMADI S A. V-net: fully convolutional neural networks for volumetric medical image segmentation[C]//Proceedings of 2016 Fourth International Conference on 3D Vision (3DV). Piscataway: IEEE Press, 2016: 565-571.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[18] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

[19] IOFFE S, SZEGEDY C. Normalization: accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167.

[20] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of 2018 IEEE/CVF Conference on ComputerV ision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4510-4520.

[21] HOWARD A, SANDLER M, CHENB, et al. Searching for MobileNetV3[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 1314-1324.

[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on ComputerVision and Pattern Recognition. Piscataway: IEEE Press, 2017: 936-944.

[23] LIAO M, SHI B, BAI X, et al. Textboxes: a fast text detector with a single deep neural network[C]//Thirty-first AAAI Conference on Artificial Intelligence. Piscataway: IEEE Press, 2017.

[24] LIAO M H, SHI B G, BAI X. TextBoxes++: a single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(8): 3676-3690.

[25] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]//Proceedings European Conference on Computer Vision. Heidelberg: Springer, 2016: 56-72.

[26] SHI B G, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3482-3490.

[27] LIAO M H, ZHU Z, SHI B G, et al. Rotation-sensitive regression for oriented scene text detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5909-5918.

[28] ZHOU X Y, YAO C, WEN H, et al. EAST: an efficient and

accurate scene text detector[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2642-2651.

[29] DENG D, LIU H, LI X, et al. Pixellink: detecting scene text via instance segmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Piscataway: IEEE Press, 2018.

[30] WANG W H, XIE E Z, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 9328- 9337.

[31] TIAN Z T, SHU M, LYU P Y, et al. Learning shape-aware embedding for scene text detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4229-4238.

[32] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1800-1807.

[33] SHI X J, CHEN Z, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]//Advances in Neural Information Processing Systems.[S.l.:s.n.], 2015: 802-810.

[34] LONG S, RUAN J, ZHANG W, et al. Textsnake: a flexible representation for detecting text of arbitrary shapes[C]//Proceedings of the European Conference on Computer Vision (ECCV).Piscataway: IEEE Press, 2018: 20-36.

[35] YE J, CHEN Z, LIU J H, et al. TextFuseNet: scene text detection with richerfused features[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on ArtificialIntelligence Organization, 2020: 516-522.

[36] WANG X B, JIANG YY, LUO Z B, et al. Arbitrary shape scene text detection with adaptive text region representation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 6442-6451.

[作者简介]

胡海洋（1977− ），男，杭州电子科技大学教授，主要研究方向为机器视觉、智能制造。

厉泽品（1997− ），男，杭州电子科技大学硕士生，主要研究方向为计算机视觉、文本检测识别。

李忠金（1988− ），男，杭州电子科技大学讲师，主要研究方向为计算机视觉、移动边缘计算。