Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 2 期

DOI: 10.12086/oee.2021.200270

一种车辆识别代号检测和识别的弱监督学习方法

曹　志[1,2]，尚丽丹[1,2]，尹　东[1,2*]

[1] 中国科学技术大学信息科学技术学院，安徽 合肥 230027；
[2] 中国科学技术大学中国科学院电磁空间信息重点实验室，安徽 合肥 230027

摘要：车辆识别代号对于车辆年检具有重要的意义。由于缺乏字符级标注，无法对车辆识别代号进行单字符风格校验。针对该问题，设计了一种单字符检测和识别框架，并对此框架提出了一种无须字符级标注的弱监督学习方法。首先，对 VGG16-BN 各个层次的特征信息进行融合，获得具有单字符位置信息与语义信息的融合特征图；其次，设计了一个字符检测分支和字符识别分支的网络结构，用于提取融合特征图中的单字符位置和语义信息；最后，利用文本长度和单字符类别信息，对所提框架在无字符级标注的车辆识别代号数据集上进行弱监督训练。实验结果表明，本文方法在车辆识别代号测试集上得到的检测 Hmean 数值达到 0.964，单字符检测和识别准确率达到 95.7%，具有很强的实用性。

关键词：卷积神经网络；弱监督学习；自然场景文本检测；自然场景文本识别；车辆识别代号

中图分类号：TP391.4；TP181　　**文献标志码：**A

曹志，尚丽丹，尹东. 一种车辆识别代号检测和识别的弱监督学习方法[J]. 光电工程，2021，**48**(2): 200270
Cao Z, Shang L D, Yin D. A weakly supervised learning method for vehicle identification code detection and recognition[J]. *Opto-Electron Eng*, 2021, **48**(2): 200270

A weakly supervised learning method for vehicle identification code detection and recognition

Cao Zhi[1,2], Shang Lidan[1,2], Yin Dong[1,2*]

[1]School of Information Science Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
[2]Key Laboratory of Electromagnetic Space Information of Chinese Academy of Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China

Abstract: The vehicle identification code (VIN) is of great significance to the annual vehicle inspection. However, due to the lack of character-level annotations, it is impossible to perform the single-character style check on the VIN. To solve this problem, a single-character detection and recognition framework for VIN is designed and a weakly supervised learning algorithm without character-level annotation is proposed for this framework. Firstly, the feature information of each level of VGG16-BN is fused to obtain a fusion feature map with single-character position information and semantic information. Secondly, a network structure for both the character detection branch and the character recognition branch is designed to extract the position and semantic information of a single character in the

收稿日期：2020-07-18； 收到修改稿日期：2020-10-23
基金项目：安徽省重点研究与开发计划项目(1804a09020049)
作者简介：曹志(1996-)，男，硕士研究生，主要从事计算机视觉方面的研究。E-mail：caozhihf@126.com
通信作者：尹东(1965-)，男，硕士，副教授，主要从事计算机视觉方面的研究。E-mail：yindong@ustc.edu.cn

fusion feature map. Finally, using the text length and single-character category information, the proposed framework is weakly supervised on the vehicle identification code data set without character-level annotations. On the VIN test set, experimental results show that the proposed method realizes the Hmean score of 0.964 and a single-character detection and recognition accuracy rate of 95.7%, showing high practicability.

Keywords: convolutional neural network; weakly supervised learning; scene text detection; scene text recognition; vehicle identification number (VIN)

1 引 言

车辆识别代号(Vehicle identification number, VIN)由 17 位字母和数字组合而成，是汽车上一组独一无二的号码，在车辆年检中对于核实车辆的唯一身份有重要的作用。VIN 的人工审核包含两个部分：审核图片中 VIN 是否与实际 VIN 匹配；审核图片中 VIN 字符风格(字体类型)是否与 VIN 拓印风格一致。

随着深度学习技术的发展，利用计算机自动审核已经成为趋势。VIN 的自动审核可以借助通用光学字符识别(optical character recognition，OCR)技术，通用 OCR[1-3]技术从包含文本的不特定场景中检测并识别出文本，分为自然场景文本检测[4-10]与自然场景文本识别[11-14]。

自然场景文本检测的发展主要经历了三个阶段：检测出水平方向的文本、检测出任意角度的文本和检测出弯曲的文本。CTPN[6]方法通过对 Faster RCNN[15]多个候选框合并，每次只检测文本框的一个小部分，最终可以实现对水平文本框的检测。RRPN[7]方案通过让 Faster RCNN 中 RPN 部分多预测一个角度参数，从而实现对倾斜文本框的检测。EAST[8]通过更改 SSD[16]检测算法，直接进行像素级文本预测，可以实现对倾斜文本的检测。TextSnake[9]则是利用分割网络预测文本二值区域、文本中心区域、文本中 15 个圆的半径等共 5 个分割图，实现对弯曲文本的检测。CRAFT[10]是通过预测单个字符和字符间关系的概率图，实现对弯曲文本的检测。但这些方法都是针对文本整体进行检测，无法实现对于单个字符的检测，也就无法满足 VIN 字符风格的校验需求。

自然场景文本识别主要有两种方式：基于 RNN 结构的对整体文本进行识别和基于分割方法的对每个字符进行的识别。CRNN[13]首先利用 CNN[17]和 BLSTM[18]提取特征，然后利用 BLSTM 和 CTC[19]部件获得字符图像上下文的关系，从而提升文本识别准确率，但是需要大量的训练数据，而且当识别弯曲文本时由于大量背景噪声的引入，会造成识别率的下降。Liao 等[14]提出了从二维空间对每个字符进行分类，这种方式由于是对每个字符进行识别，所以具有更少的搜索空间，更容易训练，但是这种方法在训练阶段需要字符级的标注，所以目前只能在人工合成数据集上进行训练。

现有的框架由于缺乏字符级标注，无法检测出单个字符，也就无法实现对字符风格的校验。本文设计了一种端到端的针对单字符的文本检测识别框架，并基于此提出了一种可以在 VIN 数据集上无须字符级标注的弱监督训练方法，并改进了字符分支所用的 loss 函数，向其中引入未知类别，可以最大程度上学习到其他字符特征。

2 本文方法

2.1 整体结构

本文设计的字符检测与识别框架如图 1 所示。其结构主要由三个部分组成：特征提取网络(Backbone)、字符检测分支(text detection branch)和字符识别分支(character branch)。

特征提取网络(Backbone)：用于提取图像中单字符位置和语义信息，VGG-16BN[20]作为常用的特征提取网络，需要对输入图片进行 5 次下采样以获得最终的特征图，但是会丢失单字符这种小目标。小目标是指占图像比例小于 10%或者尺寸小于 32×32 的目标，在下采样过程中会导致小目标特征信息的丢失，所以较难处理。为了解决这个问题，设计 Conv fuse 模块将高层次的语义特征通过 Upsample 的方式与浅层特征逐层融合，由于浅层特征对于小目标更加敏感，高层特征融合了语义信息，所以最终的融合特征图 F 包含了单字符小目标的位置和语义信息。

字符检测分支：即是对单字符的位置和文本行的位置进行编码，需要卷积核提取字符的前景和背景信息。图 2 展示了在不同数据集下卷积核的实际感受野，虽然理论上感受野是均匀分布的，但是实际上卷积核感受野是以二维高斯的形式分布的[21]。所以使用二维

高斯图对字符的位置进行编码，检测分支使用 4 个 3×3 的常规卷积核逐层从融合特征层 F 中提取字符的位置信息，最后使用两个 1×1 的卷积核分别解码出 regmap 和 affmap。

二维高斯图表示字符中心出现在该位置的可能性。如图 1 中 regmap 所示，颜色越红(深)，说明该位置为字符中心的可能性越高。通过 affmap 对属于同一个文本行的相邻字符的中心概率进行预测，如图 1 中颜色越红(深)，说明该位置越有可能是两个属于同一文本框相邻字符的中心。

字符识别分支：由于检测分支采用了回归的方式对字符的概率进行编码，在字符识别分支，则需要对字符的类别进行编码，这里使用像素类别分类的编码方式。由于常规卷积核无法适应不同字符的结构特征，因此使用可形变卷积核(DCNV2)[22]在融合特征层 F 上收集不同字符结构的空间信息，接着使用一个常规 3×3 卷积核汇总收集到的字符类别信息，最后利用一个常规的 3×3 卷积核将这些信息编码成 38 通道的概率图，经过 softmax 后得到 clsmap。Clsmap 每个像素有 38 个通道，每个通道表示该像素属于 38 个类别(26 个字母、10 个数字、1 个背景类、一个特殊字符)的概率。在推理时，使用 argmax 即可获得每个像素的最大可能的类别。图 1 中 clsmap 不同颜色表示网络预测该区域像素的类别，白色为背景类。

这里值得注意的是，DCNV2 与常规卷积核的区别，如图 3 所示，常规卷积核感受野如图 3(a)蓝色区域，过大会引入背景噪声，不利于对单个字符的识别；过小则无法捕捉字符的结构特征。而图 3(b)所示的

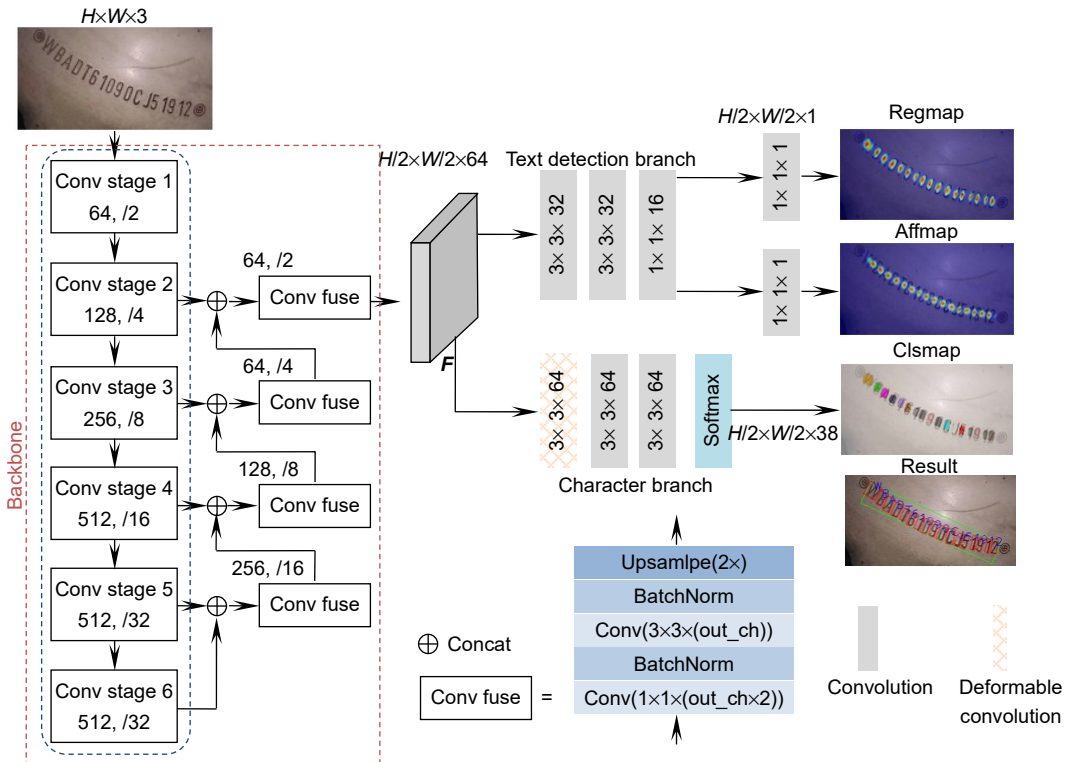


图 1 总体框架图

Fig. 1 Overall framework

图 2 实际有效感受野[21]

Fig. 2 Actually effective receptive field[21]

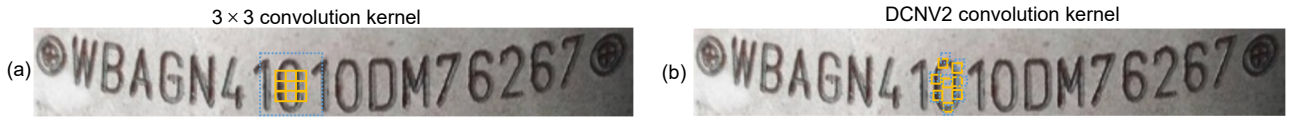


图 3　不同卷积核的对比

Fig. 3　Comparison of different convolution kernels

DCNV2 可以自适应学习到字符的结构特征，减少噪声的引入。

总体上而言，本文设计的 VIN 检测和识别框架 Backbone 提取融合单字符位置信息和类别信息的融合特征图；字符检测分支将位置信息解码成单字符出现的概率图；字符识别分支将语义信息解码成单字符类别分割图，最终可以检测并识别单个字符。通过对图片的同一位置提取语义信息和位置信息，可以获得更加鲁棒的语义特征，从而实现比单个分支更好的效果。

2.2　弱监督学习算法

为了解决缺乏字符级标签的问题，本文提出了一种弱监督学习方法，以实现对整体框架端到端的训练。弱监督训练分为两个步骤：一是利用已有的人工合成的字符级标注数据集 SynthText[23]对网络进行强监督训练；二是将 VIN 数据集与具有字符级标注的 SynthText、Icdar13[24]和 SCUT-FORU[25]数据集进行混合弱监督训练。

强监督训练: 用于赋予所提框架初始的单字符检测和识别能力，在人工合成数据集 SynthText 上进行预训练。对于具有字符级标注的 SynthText 的图片，需要根据字符级标签生成 regmap、affmap 和 clsmap，从而完成对检测和识别分支的训练。

对于检测分支，如图 4 所示，首先生成一个均值为 0 方差为 1 的二维正态高斯分布图，接着将高斯图映射到每个字符框区域内，即可获得 regmap。Affbox 由两个相邻的字符框获得：通过连接单字符框的对角线，可以获得上下两个三角形，两个相邻字符框共有四个三角形，依次连接三角形的中心，即可获得一个 affbox。按照生成 regmap 的方式将高斯图映射到每个 affbox，获得 affmap。

对于识别分支，由于相邻字符框重叠部分像素会有类别歧义，所以对字符框做缩小处理。将每个字符框保持中心不变，边长缩小为原来的 1/2 获得 class box，再对 class box 内的像素分配为对应的类别标签。

按照图 4 所示过程获得 regmap、affmap、clsmap 后，即可对网络进行强监督训练。

弱监督训练：用于提升网络对 VIN 单个字符的检测和识别能力，在不具备字符级标注的 VIN 数据集和具备字符级标注的 Icdar13、SCUT-FORU、SynthText 进行混合弱监督训练。

弱监督训练即是利用网络对 VIN 图片估计出字

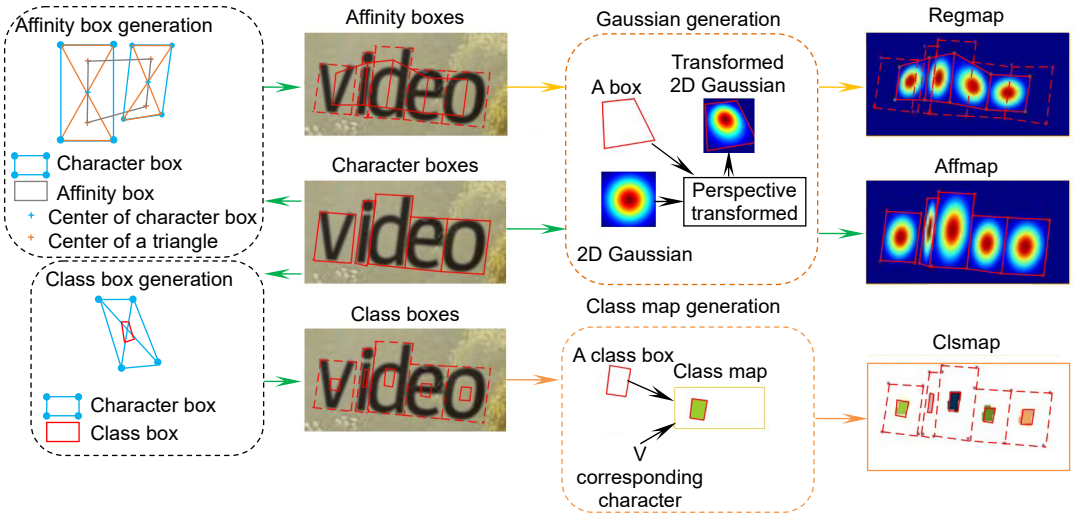


图 4　具有字符级标注的标签生成过程

Fig. 4　Label generation for images with character-level annotations

符级伪标签，并不断迭代，从而获得质量越来越高的伪标签，完成对网络的端到端训练。字符级伪标签的生成过程如图 5 绿色实线箭头所示，主要由以下 5 个步骤构成：

1) 利用文本框标注，得到 crop VIN。将 crop VIN 输入上个 epoch 训练得到的模型中，网络预测出 regmap 和 clsmap；

2) 利用分水岭算法[26]在 regmap 上获得每个字符框，利用 clsmap 获得每个字符框的类别。将字符框从左到右排列，获得对应字符框类别序列，字符框类别序列即为预测字符串 predict。

3) 利用字符串匹配算法，将预测出的 predict 与实际 VIN 即 gt 进行字符串匹配，获得匹配后的字符串 match。利用 match 删除多余字符框，对字符框类别进行修正。字符串匹配算法将在图 6 中详述。

4) 将 crop VIN 区域的字符框和对应类别映射回原图。计算本次伪标签置信度 confd，与上次生成伪标签的置信度进行对比，保留置信度较高的伪标签。confd 的计算规则如 2.3 节式(4)。

5) 根据映射回原图的 psedudo-label，对于检测分支，按照图 4 所示方式生成对应 pseudo-regmap、pseudo-affmap 。Confidence-map 用于评估 pseudo-regmap 和 pseudo-affmap 像素点的置信度，其中灰色部分表示置信度小于 1，白色部分置信度为 1。对于识别标签，若匹配字符串 match 不包含未知类别 #，直接按照图 4 所示方式生成 clsmap；若匹配字符串 match 包含未知类别#，则首先将 VIN 区域内像素类别设为未知，对应 pseudo-clsmap 黑色部分，接着按照图 4 方式生成 clsmap。未知类别的引入让网络在无法正确估计所有伪标签的情况下，可以尽可能学习到其他准确估计的字符特征。

图 5 中红色实线展示了利用上个 epoch 生成的伪标签对当前模型进行弱监督训练的过程，绿色实线箭头展示了字符级伪标签的生成过程。

图 6 中，predict 表示根据网络输出的 regmap 和 clsmap 推理出的字符串；gt 表示实际的 VIN 码；@表示字符框被识别为背景类；“#”表示进行字符串匹配后，预测框的类别与实际类别不匹配，此时将其设置为未知类别；“$”表示进行字符串匹配后，预测的框多余，需要被删除。将 predict 相对 gt 进行左右移位，图 6 中黑色“\”表示预 predict 与 gt 对应位类别一致，取匹配个数最多的位移为最终位移。

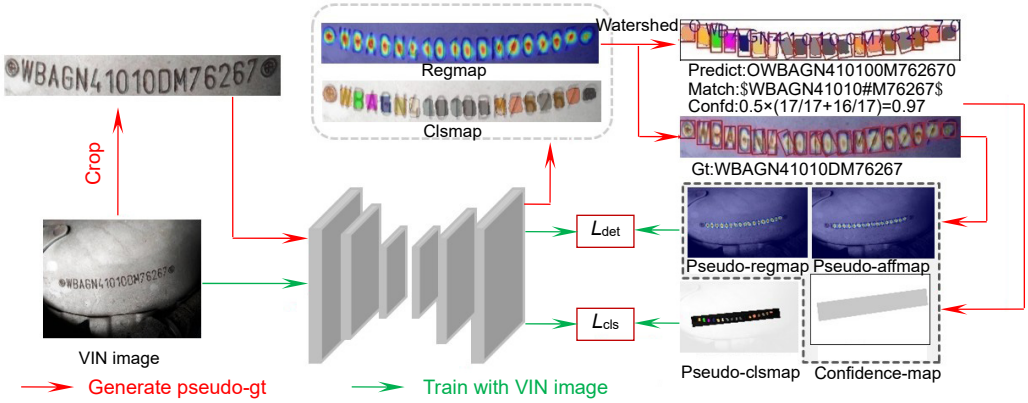


图 5 VIN 伪标签的生成过程

Fig. 5 Pseudo-gt generation for VIN

-3	-2	-1	0																1	2	3	correct
predic			O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2				
gt			2	H	H	Y	D	2	8	3	5	B	H	2	0	0	1	1	2					
O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2						0
	O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2					2
		O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2				13
			O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2			2
				O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2		0
					O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2	0
						O	2	H	E	@	D	2	G	1	5	B	H	2	0	0	1	1	2	0
		$	2	H	#	#	D	2	#	#	5	B	H	2	0	0	1	1	2					match

图 6 字符串匹配算法

Fig. 6 String matching algorithm

由于分水岭算法有欠分割或者过分割的问题，这会导致检测到的框过小或者过大，所以不能将对应位置的字符框类别进行强行类别分配，使用“未知类别#”来替代不匹配的字符类别。“未知类别#”引入作用是让识别分支尽可能学习到可以成功预测的字符的特征。如图 7 所示，当预测字符串中存在未知类别，则仅对正确匹配的字符框生成伪标签，字符区域其余类别值为黑色未知。图 7 中预测图蓝色框为正确匹配，红色框为未知类别，在伪标签图中，正确匹配的区域和背景加入训练。未知类别的引入可以在无法完全预测正确的情况下，对其余字母特征进行学习。

由于估计的框可能有误差，所以在实际进行训练时，还需要将 VIN 训练集与具有字符级标注的数据集 SynthText，Icdar13 和 SCUT-FORU 按照一定比例混合进行训练，目的是让网络学习到更通用的特征，从而对 VIN 估计出更加精确的字符框和字符类别。

总而言之，本文所提弱监督学习框架为：首先在具有字符级标注的人工合成数据集进行强监督训练，获得初始的权重；接着利用初始权重预测出 VIN 训练集的字符级标签，并且利用字符串匹配算法对标签进行匹配，获得估计的伪标签；最后将估计出字符级标注的 VIN 数据集与其他具有字符级标注数据集混合训练，并不断迭代伪标签，获得质量越来越高的伪标签，从而完成对 VIN 数据集的端到端训练，全程无需人工介入。

2.3 损失函数

1) 总体损失函数

本文总体损失函数如下:

$$L = L_{\text{det}} + \alpha L_{\text{cls}} \quad , \tag{1}$$

式中：L_{det} 为检测损失函数，α 为多任务系数(设为 1)和 L_{cls} 为识别损失函数。

2) 检测分支损失函数

由于是对单个字符的概率回归，所以使用均方误差损失函数，同时因为进行弱监督训练，故需对每个像素点的损失进行加权。其损失函数如下：

$$L_{\text{det}} = \sum_{p} S_{\text{conf}}(p) \cdot \left(\left\| S_{\text{r}}(p) - S_{\text{r}}^{*}(p) \right\|_2^2 + \left\| S_{\text{a}}(p) - S_{\text{a}}^{*}(p) \right\|_2^2 \right), \tag{2}$$

式中：p 表示单个像素点；$S_{\text{conf}}(p)$ 表示像素 p 的置信度，当使用具有字符级标注的数据集进行训练时，其值为 1；$S_{\text{r}}(p)$ 和 $S_{\text{a}}(p)$ 为实际 regmap 和 affmap 像素 p 的数值；$S_{\text{r}}^{*}(p)$ 和 $S_{\text{a}}^{*}(p)$ 表示网络预测的 regmap 和 affmap 像素 p 的数值。

弱监督学习过程需要对检测分支生成的标签进行置信度评估，其规则如下：

$$S_{\text{conf}}(w) = \frac{1}{2}\left(\frac{l^{\text{p}}(w)}{l(w)} + \frac{l^{\text{c}}(w)}{l(w)} \right) \quad , \tag{3}$$

式中：w 表示 VIN 图片中被标注的区域，该区域像素的置信度正比于预测出的字符串长度和预测正确的类别个数；$l(w)$ 表示 VIN 的实际长度；$l^{\text{p}}(w)$ 表示预测出的字符的个数，预测出的字符个数与实际个数越接近，置信度越高；$l^{\text{c}}(w)$ 是预测正确的字符的个数，预测正确的个数越接近 $l(w)$，置信度越高。

获得 $S_{\text{conf}}(w)$ 后，通过式 (4) 可以获得 confidence-map，其作用是对 regmap 和 affmap 产生的 loss 进行加权，估计的置信度越高，权重越高。当置信度小于 0.5 时，可以认为预测出的字符位置严重偏离事实，如果将这些标签加入训练，会降低预测的准确性，对于这种情况将 w 等距离划分成 $l(w)$ 份，并将置信度设为 0.5，对于 w 以外的区域，其置信度为 1。对于识别分支，当置信度小于 0.5 时，则将 VIN 区域所有像素设为未知。

$$S_{\text{conf}}(p) = \begin{cases} \min(0.5, S_{\text{conf}}(w)) & p \in w \\ 1 & \text{otherwise} \end{cases} \quad 。 \tag{4}$$

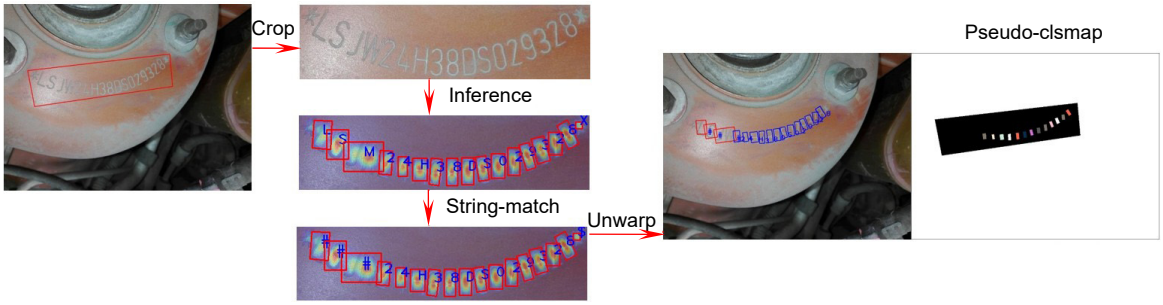


图 7 字符识别分支伪标签生成过程

Fig. 7 Generation process of character recognition branch pseudo label

3) 识别分支损失函数

引入未知类别，从而可以对识别分支进行弱监督训练。未知类别的像素即是不参与损失函数的计算，对应图 5 中 pseudo-clsmap 的黑色像素，其计算规则如式(5)所示：

$$L_{\text{cls}}=-\frac{4}{H\times W-4\cdot N_{\text{uk}}}\cdot\sum_{i=0}^{H/2}\sum_{j=0}^{W/2}W_{i,j}\left(\sum_{c=0}^{C-1}(Y_{i,j}=c)\ln\left(\frac{e^{X_{i,j,c}}}{\sum_{k=0}^{k=C-1}e^{X_{i,j,k}}}\right)\right)\quad,\tag{5}$$

式中：H 和 W 对应输入图片的高和宽，图 5 中 pseudo-clsmap 中的高和宽为 $H/2$ 和 $W/2$，通道数为 38；C 表示每个像素类别的个数，共有 38 类；$Y_{i,j}$ 对应 pseudo-clsmap 中第 i 行第 j 列像素的实际类别；$X_{i,j,c}$ 表示网络实际输出的 clsmap 第 i 行第 j 列第 c 个通道输出的数值；N_{uk} 对应未知类别像素的个数(因为未知类别的像素的损失被忽略，所以需要减去未知类别的像素的个数)；$W_{i,j}$ 用于平衡正负类像素样本不均衡并对未知类别的像素赋权为 0，负类为背景类，占了大多数像素，$W_{i,j}$ 的计算规则：

$$W_{i,j}=\begin{cases}\dfrac{N_{\text{neg}}}{N-N_{\text{neg}}} & Y_{i,j}\in[1,37]\\ 1 & Y_{i,j}=0\\ 0 & Y_{i,j}=\text{unknown}\end{cases},\tag{6}$$

式中：$N=H/2\times W/2-N_{\text{uk}}$，对应图 5 中 pseudo-clsmap 中除去未知类别的像素总数；$Y_{i,j}=0$，对应背景类，为 pseudo-clsmap 白色像素；N_{neg} 对应背景类像素的个数，VIN 区域以外的部分均为背景类。当 $0.5\le S_{\text{conf}}(w)<1$ 时，w 区域中除了预测正确的字符区域为正类，其余分配为未知类别；当 $S_{\text{conf}}(w)<0.5$ 时，将 w 区域内所有像素设为未知类；当 $S_{\text{conf}}(w)=1$ 时则按照生成的标签进行训练，说明此时网络已经完全能够检测并识别每个字符了。

2.4 推理过程

推理过程由四个步骤组成，如图 8 所示：1) 将 regmap 和 affmap 相加得到 VIN 区域二值图(binary map)，寻找二值图中的矩形轮廓，区域面积最大的矩形为检测到的 VIN 文本框；2) 在 regmap 和 clsmap 上裁剪出对应的区域，获得 crop regmap 和 crop clsmap；3) 在 crop regamp 使用分水岭算法分割出每个字符的框，并统计每个框内像素种类的个数，取像素数量最多的类别为字符框的类别；4) 将检测到的 VIN 区域框和单个字符框映射回原图，即可获得检测到的 VIN 区域框和单个字符框。

3 实 验

3.1 实验环境和参数设置

1) 实验环境

本文实验平台为：CPU 志强 E5、GPU/TitanXP、Ubuntu16.04、Python 3.6、Pytorch 1.0。

2) VIN 数据集

本文采用的 VIN 数据集在车检流水线上采集，标注 VIN 区域的四个顶点。VIN 数据集包含 2120 张训练图片和 1234 张测试图片，其分辨率从 480×360 到 2048×1536 不等，每张图片仅包含一个待识别的 VIN，由于每个 VIN 包含 17 个字符，所以整个数据集实际包含 36040 个训练字符和 20978 个测试字符。图 9 展示了 VIN 数据集部分图片，其标注框如红色实线框所

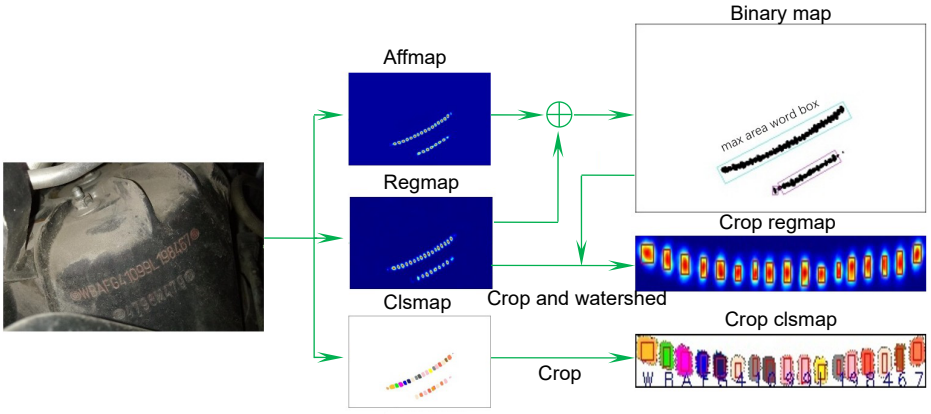


图 8 推理过程

Fig. 8 Reasoning process

图 9 VIN 数据集部分图示

Fig. 9 Illustration of VIN dataset

示。

3) 训练过程及参数设置

训练分为强监督训练阶段和真实数据混合弱监督训练阶段。强监督训练阶段在 SynthText 上完成，SynthText 数据集是大约包含 80 千张具有字符级标注的人工合成数据集。由于数据量巨大，此阶段不对数据集进行增广操作。在训练时，保持图片的长宽比不变，将图片放缩到 640×640，batch size 设为 8，在 SynthTex 上训练 1 个 epoch，学习率设为 1E-4，其余参数为优化器默认参数，优化器采用 Adam[27]。

4) 评价指标

使用检测精度和识别精度来对性能进行衡量。

对于检测性能，采用 IoU>0.7 的 Hmean(f-measure) 作为性能评价指标，其规则：

$$H_{\text{mean}}=2\times\frac{P_{\text{precision}}\times R_{\text{recall}}}{P_{\text{precision}}+R_{\text{recall}}}\ , \tag{7}$$

式中：$P_{\text{precision}}$ 为准确率(precision)，表示模型预测出的正样本中真样本的比例；R_{recall} 为召回率(recall)，表示模型预测出的正样本占所有正样本的比例。

对于单个字符的识别精度(Accuracy，用 A_{accuracy} 表示)，由于 VIN 并没有字符级标注，所以对于单字符的识别精度，其规则：

$$A_{\text{accuracy}}=\frac{T_{\text{correct}}}{T_{\text{all}}}\ , \tag{8}$$

式中：T_{correct} 为预测出的 VIN 经过字符串匹配算法正确匹配的字符的总数(correctNum)，T_{all} 是指待识别 VIN 字符的总数(allNum)。

3.2 实验与分析

为了说明所提方法的先进性和有效性，本文做了以下四个实验：实验一将本文所提算法的性能与当前主流算法对比；实验二对不同的参数进行消融，比较检测和识别精度；实验三对比了不同字符识别分支的准确率；实验四展示了本文所提弱监督算法迭代训练的结果。

实验一：测试图片的尺寸为 1200×1200，实验结果如表 1 所示。

本实验选取了最新的并且已经开源的文本检测框架 EAST[8]、TextSnake[9]、CRAFT[10]和文本识别框架 CRNN[15]与本文所提的算法进行对比。表 1 中，对于文本检测能力，与 EAST、TextSnake 和 CRAFT 分别比较 VIN 整体检测的 Recall、Precision 和 Hmean；对于识别能力，与 CRNN 比较 VIN 字符的 Accuracy。

对于文本检测框架，利用开源模型提供的在 SynthText 上的预训练权重，在 VIN 训练集上进行微调，获得最终的模型。测试时，为了公平对比，只保留最大面积的预测框。实验表明，相比于倾斜文本检测框架 EAST，本文算法检测 Hmean 值有 0.125 的提升；相比于弯曲文本检测算法 CRAFT 和 TextSnake，Hmean 值分别有 0.203 和 0.05 的提升。对于识别框架 CRNN，VIN 识别精度有 16.8%的提升。实验结果表明，本文所提出的框架无论是在 VIN 的检测还是字符的识别精度上都超过了主流的方法，而且可以实现对单

表 1 与其他算法进行对比

Table 1 Comparison of different algorithms

Methods	Recall	Precision	Hmean	Accuracy/%	Speed/(f/s)
EAST	0.832	0.845	0.839	——	17.3
TextSnake	0.957	0.960	0.959	——	18.2
CRAFT	0.761	0.761	0.761	——	8.4
CRNN	——	——	——	78.9	30.2
Ours	0.964	0.964	0.964	95.7	8.1

个字符的检测。

实验二：在 VIN 的测试集上，对比了四个因素对模型的影响：表 2 中，真实图片用于验证具有字符级标注数据集 Icdar13 与 SCUT-FORU 对于框架的影响；识别分支用于验证检测与识别任务是否能相互促进；DCNV2 用于验证在识别分支引入可形变卷积核对于检测和识别能力的影响；未知类别用于验证未知类别的引入对模型的影响。

表 2 中，方法 1 和 2 对比了在缺乏识别分支情况下真实图片对于检测能力的影响。由于 SynthText 为人工合成数据集，所以引入了真实场景特征后能在一定程度上提升模型泛化能力，从而提升检测能力；方法 6 和 7 则对比了加入识别分支后的真实图片的对比，检测和识别精度也有一定的提升，说明真实图片的特征能够提升模型的检测和识别能力，同时也表明识别分支的引入可以提升模型的鲁棒性。

方法 2 和 3 对比了识别分支对于模型的影响。方法 3 结果表明，字符识别分支能在一定程度上提升检测的精度，是因为识别分支能够促使主干网络提取包含语义信息的更加鲁棒的融合特征。

方法 3 和 5、方法 4 和 7 对比了 DCNV2 对于识别精度的影响，实验结果表明自适应感受野对于字符识别较为重要，字符识别能力的提升也能促进检测能力的提升。

方法 3 和 4、方法 5 和 7 对比了本文提出未知类别对于检测和识别能力的影响，当不具有未知类别时，网络仅能学习到完全正确估计的 VIN 图片的特征，但是当引入未知类别，网络还可以进一步学习到部分正确估计的 VIN 图片的特征，从而提升网络的精度。

实验三：对比了不同字符识别分支在 VIN 训练集上裁剪出的 VIN 区域的字符识别准确率，如表 3 所示。

表 3 中 3×3 表示尺寸为 3 的常规卷积核，dcn(3×3) 表示尺寸为 3 的可形变卷积核(DCNV2)。方法为首先在 SynthText 上预训练 1 个 epoch，保持除了字符识别分支以外的所有参数一致；接着利用预训练得到的模型权重，识别 VINSet 训练集图片裁剪出的 VIN 区域，对比不同结构字符识别准确率的大小。

在实验之初，考虑到更多的卷积核可以获得更大的感受野，从而更能提取到字符的内在特征，所以首先使用 4 个 3×3 的卷积核提取字符内在结构特征，然后使用 1×1 的卷积核提取通道的特征，发现在训练集上仅仅达到 63.1%的识别精度，并且发现 467QI 这五

表 2 不同模块对模型精度的影响

Table 2 Comparison of effects of different modules on model accuracy

方法	1	2	3	4	5	6	7
真实图片		✓	✓	✓	✓		✓
识别分支			✓	✓	✓	✓	✓
DCNV2					✓	✓	✓
未知类别				✓		✓	✓
Hmean	0.654	0.761	0.793	0.851	0.812	0.928	0.964
Accuracy/%	----	----	69.3	80.2	74.6	93.2	95.7

个字符的识别率接近 0。考虑到可能是最后 1×1 的卷积核没有办法获得空间上的信息，所以第二个结构只使用 3 个 3×3 的卷积核对字符识别，最后一个 3×3 卷积核用于融合空间和通道特征，发现识别精度可达 72.7%。但是发现某些字符识别精度较低。考虑到可能是感受野过大，引入了过多的背景噪声，所以采用 3 个 3×3 的卷积核继续试验，发现字符的识别精度各个类别都较高，但是对于字母 1JL 这样瘦长型的字符识别精度较低。为了解决这个问题，引入可形变卷积核(DCNV2)，对比了将 DCNV2 放在最后用于压缩通道和放在最前面用于提取初始的结构特征，发现放在最前面识别准确率较高。而且可形变卷积核的引入确实可以解决瘦长型字符识别精度较低的问题，所以在字符识别分支最后采用了 dcn(3×3)，3×3，3×3 这样的结构。

实验四：统计在 VIN 训练集上的字符识别准确率，如表 4 所示。

随着迭代训练的不断进行，字符的总体识别准确率在不断提高，这也说明了对单个字符的检测精度在不断提高。图 10 则展示了随着迭代训练的不断进行，一些初始时无法准确预测的 VIN 也逐渐可以准确预测。通过从总体的字符识别准确率以及迭代训练时的一些图例可以说明，所提的弱监督学习算法是有效的。

3.3 实验结果展示

从两个方面即推理出字符框及类别的最终结果图以及包含了网络三个分支输出可视化的实验图展示实验结果。这里所有的图片来自 VIN 测试集，测试尺寸 1200×1200。

图 11 表明通过弱监督学习后，本文所提出的 VIN 检测识别框架可以成功地检测并识别每个字符。图 12 展示了网络的输入输出及后处理结果。输入原始图片，网络输出 regmap、affmap、clsmap，通过后处理获得每个字符框的位置及对应类别。这里值得注意的是，利用最大面积可以筛选出 VIN 区域文本框。为了方便可视化，将网络输出结果叠加在原图上显示。

表 3 字符识别分支结构对比实验

Table 3 Comparative experiments on the branch structure of character recognition

字符识别分支结构	识别准确率/%
3×3,3×3,3×3,3×3,1×1	63.1
3×3,3×3,3×3,3×3	72.7
3×3,3×3,3×3	74.2
3×3,3×3,dcn(3×3)	76.8
Dcn(3×3),3×3,3×3	81.1

表 4 迭代训练结果

Table 4 Iterative training results

Epoch	识别正确字符数	准确率/%
0	29228	81.10
10	31067	86.20
20	32256	89.50
30	33554	93.10
40	35534	98.59

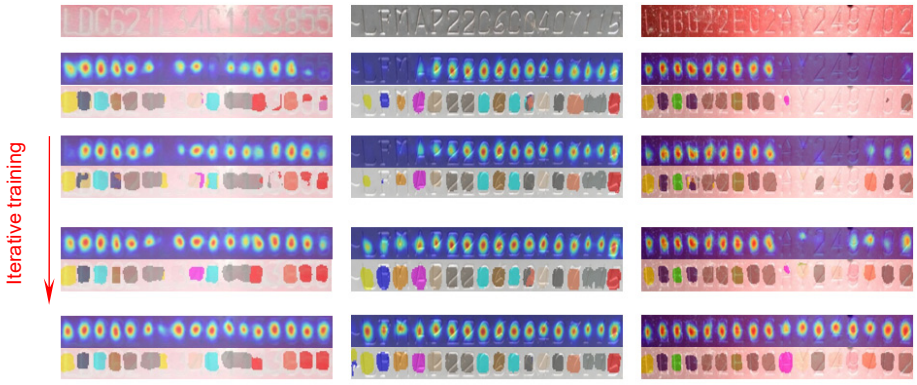


图 10 迭代训练图示

Fig. 10 Iterative training diagram

图 11 VIN 检测及识别结果

Fig. 11 VIN detection and recognition results

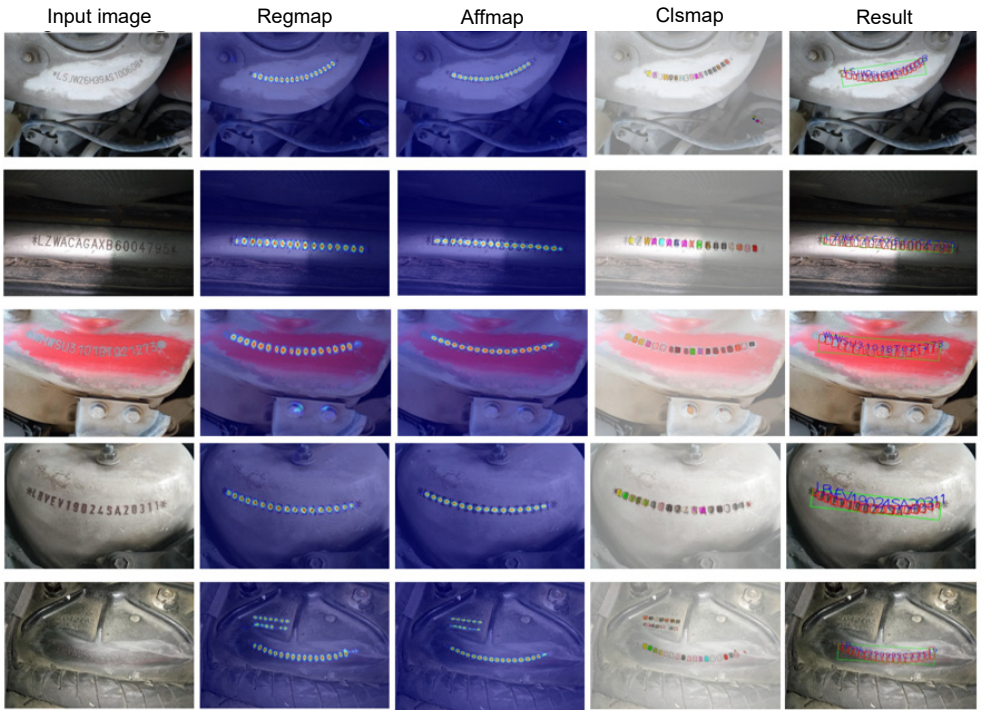


图 12 网络的输出及后处理结果

Fig. 12 Network output and post-processing results

4 结束语

针对车辆识别代号数据集缺乏字符级标注的问题，本文提出了一种可以端到端训练的 VIN 检测和识别框架，并且针对该框架提出了一种弱监督学习算法，使得检测和识别单个字符成为可能。实验结果表明，该网络不但能够进行弱监督学习、成功检测和识别单个字符，而且检测精度和识别精度能够达到较好的效

果。本文下一步的工作是实现 GPU 版本的后处理算法，以节省数据的迁移时间。

参考文献

[1] Subedi B, Yunusov J, Gaybulayev A, *et al*. Development of a low-cost industrial OCR system with an end-to-end deep learning technology[J]. *IEMEK J Embedded Syst Appl*, 2020, **15**(2): 51–60.

[2] Rashtehroudi A R, Shahbahrami A, Akoushideh A. Iranian license plate recognition using deep learning[C]//*Proceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP)*, 2020: 1–5.

[3] Naz S, Khan N H, Zahoor S, *et al*. Deep OCR for Arabic script-based language like Pastho[J]. *Expert Syst*, 2020, **37**(5): e12565.

[4] Liao M H, Wan Z Y, Yao C, *et al*. Real-time scene text detection with differentiable binarization[C]//*Proceedings of the AAAI*, 2020: 11474–11481.

[5] Liu Y L, Chen H, Shen C H, *et al*. ABCNet: real-time scene text spotting with adaptive Bezier-curve network[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 9809–9818.

[6] Tian Z, Huang W L, He T, *et al*. Detecting text in natural image with connectionist text proposal network[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 56–72.

[7] Ma J Q, Shao W Y, Ye H, *et al*. Arbitrary-oriented scene text detection via rotation proposals[J]. *IEEE Trans Multimed*, 2018, **20**(11): 3111–3122.

[8] Zhou X Y, Yao C, Wen H, *et al*. East: an efficient and accurate scene text detector[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 5551–5560.

[9] Long S B, Ruan J Q, Zhang W J, *et al*. Textsnake: a flexible representation for detecting text of arbitrary shapes[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 20–36.

[10] Baek Y, Lee B, Han D Y, *et al*. Character region awareness for text detection[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 9365–9374.

[11] Shi B G, Yang M K, Wang X G, *et al*. ASTER: an attentional scene text recognizer with flexible rectification[J]. *IEEE Trans Pattern Anal Mach Intell*, 2019, **41**(9): 2035–2048.

[12] Wang Q Q, Huang Y, Jia W J, *et al*. FACLSTM: ConvLSTM with focused attention for scene text recognition[J]. *Sci China Inf Sci*, 2020, **63**(2): 120103.

[13] 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 Trans Pattern Anal Mach Intell*, 2016, **39**(11): 2298–2304.

[14] Liao M H, Zhang J, Wan Z, *et al*. Scene text recognition from two-dimensional perspective[C]//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2019: 8714–8721.

[15] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[16] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 21–37.

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012: 1097–1105.

[18] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural Comput*, 1997, **9**(8): 1735–1780.

[19] Graves A, Fernández 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*, 2006: 369–376.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[Z]. arXiv:1409.1556, 2014.

[21] Luo W J, Li Y J, Urtasun R, *et al*. Understanding the effective receptive field in deep convolutional neural networks[C]//*Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016: 4905–4913.

[22] Zhu X Z, Hu H, Lin S, *et al*. Deformable ConvNets V2: more deformable, better results[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 9308–9316.

[23] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2315–2324.

[24] Karatzas D, Shafait F, Uchida S, *et al*. ICDAR 2013 robust reading competition[C]//*Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*, 2013: 1484–1493.

[25] Zhang S Y, Lin M D, Chen T S, *et al*. Character proposal network for robust text extraction[C]//*2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: 2633–2637.

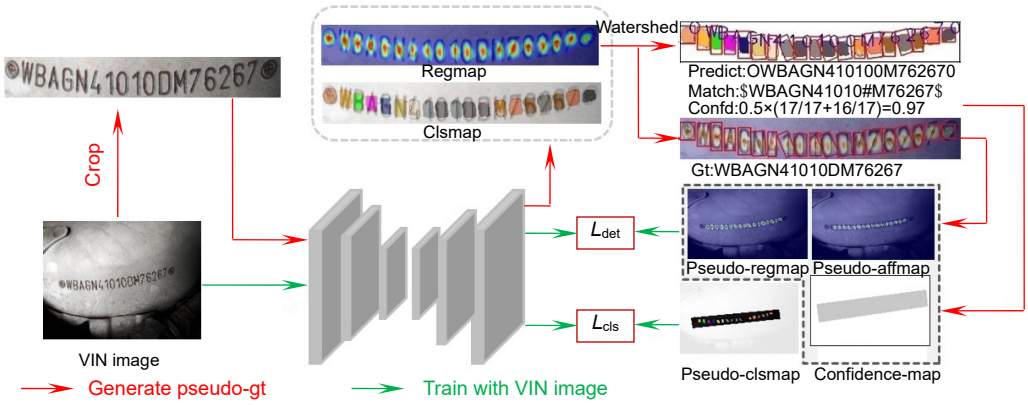
[26] Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations[J]. *IEEE Trans Pattern Anal Mach Intell*, 1991, **13**(6): 583–598.

[27] Kingma D P, Ba J. Adam: a method for stochastic optimization[Z]. arXiv:1412.6980, 2014.

A weakly supervised learning method for vehicle identification code detection and recognition

Cao Zhi[1,2], Shang Lidan[1,2], Yin Dong[1,2*]

[1]School of Information Science Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
[2]Key Laboratory of Electromagnetic Space Information of Chinese Academy of Sciences,
University of Science and Technology of China, Hefei, Anhui 230027, China



Pseudo-gt generation for VIN

Overview: The vehicle identification number (VIN) is a combination of 17 letters and numbers. It is a unique set of numbers on the car. It plays an important role in verifying the unique identity of the vehicle during the annual vehicle inspection. The manual review of VIN consists of two parts: reviewing whether the VIN in the picture matches the actual VIN; reviewing whether the VIN character style (font type) in the picture is consistent with the VIN extension style. With the development of deep learning technology, the use of computer automatic review has become a trend. The automatic review of VIN can use the universal optical character recognition (OCR) technology. Universal OCR detects and recognizes text from non-specific scenes containing text, which is mainly divided into scene text detection and scene text recognition. The development of scene text detection has mainly gone through three stages: the detection of horizontal text, the detection of text at any angle, and the detection of curved text. There are two main ways of scene text recognition: recognition of the whole text based on the RNN structure and recognition of each character based on the segmentation method. However, due to the lack of character-level annotations, both the text detection method and the text recognition method treat the entire text line as a whole. Since the verification of the character style of VIN needs to detect a single character, we propose a framework to detect and recognize a single character at the same time. In order to solve the problem of lack of character-level annotations in the VIN dataset, we propose a weakly supervised learning algorithm for the framework, which can achieve end-to-end training of the framework. The single character detection and recognition framework proposed in this paper is mainly composed of three parts, namely, backbone, text detection branch, and character branch. Backbone is used to extract the feature F that combines the location and semantic information of the picture. Text detection branch is used to decode single-character position information from F. Character branch is used to extract the category information of a single character from F. The weakly supervised learning algorithm is used to estimate the single-character pseudo-labels, thus completing the training of the framework. The final experimental results show that our framework can not only detect and recognize a single character without character-level annotations, but also achieve good results in detection accuracy and recognition accuracy.

Cao Z, Shang L D, Yin D. A weakly supervised learning method for vehicle identification code detection and recognition[J]. *Opto-Electron Eng*, 2021, **48**(2): 200270; DOI: 10.12086/oee.2021.200270

Foundation item: Key Research and Development Plan Projects in Anhui Province (1804a09020049)
* E-mail: yindong@ustc.edu.cn