一种轻量级的不规则场景文本识别模型

产世兵,刘宁钟,沈家全
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要:场景文本识别是近年来极具挑战性的任务,不同于规则的文档文本图像,场景图像中的文本具有形态多变和弯曲等特点,识别起来很有难度。该文提出了一种轻量级的场景文本识别模型(ISTR-LW),不同于现有的场景文本识别模型具有参数量大的缺点,该模型在特征序列提取中引入了经过改变后的轻量级网络 PeleeNet,不仅大幅度减少了模型的参数量,还加快了网络预测的速度;在循环网络层中获取标签分布时,引入了 Dense Block 模块,加快了网络训练的收敛速度;在获取最终识别结果时,引入了注意力机制,获得需要关注的重点区域,提高了模型文本识别的准确度;引入了薄板样条插值转换,通过修正不规则的文本,改善了不规则的文本识别率低的问题。ISTR-LW 模型是一个端到端的文本识别模型,在 Synth90K、Street View Text 和 ICDAR 等公开数据集上进行了实验,取得了不错的效果。

关键词:场景文本识别;卷积神经网络;轻量级网络;循环神经网络;空间变换网络
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2020)11-0020-05
doi:10.3969/j.issn.1673-629X.2020.11.004

A Lightweight Model for Irregular Scene Text Recognition

CHAN Shi-bing,LIU Ning-zhong,SHEN Jia-quan
(School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China)

Abstract:Scene text recognition is a challenging task in recent years. Unlike regular document text image,the text in scene image has the characteristics of changeable shape and bending,so it is quite difficult to recognize. A lightweight model for irregular scene text recognition (ISTR-LW) is proposed. Different from the existing scene text recognition model,which has a large number of parameters,we introduce the changed lightweight network PeleeNet into the feature sequence extraction of the model,which not only greatly reduces the number of parameters of the model,but also speeds up the network prediction. The Dense Block module is introduced to obtain the label distribution in the recurrent neural network,which greatly accelerates the convergence of the network. The attention mechanism is introduced to obtain the final recognition results,which improves the accuracy of model text recognition. The thin-plate spline transformation improves the low accuracy rate of irregular text by correcting irregular text. ISTR-LW model is an end-to-end text recognition model. Experiments are carried out on Synth90k,Street View Text,ICDAR and other public data sets to obtain better results.

Key words:scene text recognition;convolutional neural network;lightweight network;recurrent neural network;spatial transformation network

0 引 言

随着科学技术的不断发展,人们的生活方式也不断地向现代化、科技化、智能化转变。如今,人们获取信息的方式越来越多样化、智能化,特别是随着各种智能设备的普及,人们随时随地可以使用智能手机、数码相机、iPad 等设备获取信息。这些移动设备给人们获取图像提供了优越的平台,人们利用这些移动设备可以获取到各种内容丰富的场景图像,增加了图像的多样性。

在自然场景图像中,包含了丰富的文本语义信息,这些语义信息是帮助人们更加深入理解场景的重要依据。近年来,场景文本识别已经在工业、商业和民用上占据了重要的地位。不同于工整的文档文本图像,场景图像文本表现得十分不同,它具有背景多样性、字体不统一、间隔不等、颜色不同和字符大小不一等特点,场景图像中的文本区域还可能会产生变形。这些特点加大了对场景文本识别的难度。使用传统的基于字符分割和单字符识别的方法[1-3],可以很好地识别文档

收稿日期:2019-12-05 **修回日期**:2020-04-09
基金项目:国家自然科学基金(61375021)
作者简介:产世兵(1994-),男,硕士,研究方向为计算机视觉和模式识别;刘宁钟,教授,博导,研究方向为计算机视觉和模式识别;沈家全,博士,研究方向为计算机视觉和模式识别。

文本,但是并不适用于多样的场景文本识别。

随着深度学习技术的快速发展以及广泛应用,和深度卷积神经网络[4-5](deep convolutional neural network,DCNN)在目标检测的优良表现,人们开始使用 DCNN 和常用于语音识别的循环神经网络[6-7](recurrent neural network,RNN)来建立场景文本识别模型,并在场景文本识别中取得了不错的效果,更加激励了人们对场景文本识别的研究[8-10]。

但是,目前在文本识别应用广泛的 DCNN,例如 VGG[11]、ResNet[12]等,这些网络的参数量很大,需要消耗大量内存和计算量,在训练时,需要巨大的样本量。RNN 在场景文本识别中具有广泛的应用,但是在训练的时候,经常有梯度消失和梯度爆炸的情况。为了避免这些问题,文本识别一般采用改进的 RNN 模型,比如长短记忆网络[13-14](long short-term memory,LSTM)。LSTM 在反向传播中避免小梯度的乘法运算,降低了梯度消失的概率,它是文本识别中最常用的模型之一。但是,LSTM 在时间序列上仍然是个深度网络,训练过程中的过拟合和梯度爆炸问题并没有根本解决,网络收敛较慢,模型训练比较困难。

1　相关工作

场景文本识别技术发展迅速,早期传统的基于字符分割的技术利用二值化或者滑动窗口等方法,将单个字符从背景中分别分割出来,然后再识别分割出来的字符。例如,Novikova 等人的 Extremal Regions[15]和 Bissacco 等人的 Niblack 自适应二值化算法[16]来对图片进行二值化,分割字符,此算法根据字符区域的长宽比自动调整 Niblack 窗口的大小来进行字符的分割,然后再进行字符识别。然而,场景文本图片中的背景十分复杂多变,文本的字体、颜色、大小、形态等都不统一,人们很难从这样恶劣的环境下提取出单个字符。基于滑动窗口的方法 random terns[17]和 integer programming[18],直接使用多尺度滑动窗口策略从图片中定位字符。在字符识别阶段,人们利用单个字符的语义信息来对字符进行分类。其中比较常用的方法有支持向量机[19](support vector machine,SVM)、Bayesian inference[20]和条件随机场[21](conditional random field,CRF),它们需要人工提取大量的特征,代价很大,而且不同的场景需要的特征不同,导致人工特征的鲁棒性低,泛化能力差。

近年来,基于神经网络的深度学习方法发展十分迅速,相比于传统的人工提取特征的方法,它不需要人工设计特征,可以通过学习的方式自动获取对象的特征,在目标检测、语音识别等方面取得了不错的效果。受到语音识别技术的启发,人们开始使用 Encoder-Decoder 框架进行文本识别。首先获取一张文本图片的特征序列,然后将这些特征序列转换成字符串。在 CRNN[22]中,作者利用 VGG 网络提取输入图片稳定的图片特征,然后以特征图的每列为单位生成特征序列,这些特征序列被传入 RNN 层,利用 RNN 结构提取每个特征序列的上下文特征,最后利用 CTC(connectionist temporal classification)结构获取最终的字符串。在 DTRN[23]模型中,用 CNN 滑动窗口来提取序列特征,并利用 RNN 网络来提取特征序列的上下文特征,最终获取字符串。这些基于 CNN+RNN+CTC 的网络模型是场景文本识别中的主流框架。

该文提出的 ISTR-LW 模型,通过引入轻量级网络 PeleeNet[24]、Dense Block[25]以及注意力机制[26](attention mechanism,AM),不仅加快了网络训练的收敛速度,而且提高了网络预测的速度以及准确度。此外,还采用了空间变换网络[27](spatial transformer network,STN),将弯曲的文本变换成规则的文本,改善了场景文本变形的问题。

该文主要的创新点有:

(1)提出了一种新颖的对不规则文本具有鲁棒性的场景文本识别轻量级模型;

(2)在特征序列提取阶段,引入了轻量级网络 PeleeNet,减小了 ISTR-LW 模型的大小,加快了 ISTR-LW 网络的识别速度;

(3)在循环网络层加入 Dense Block 模块,加快了 ISTR-LW 训练收敛速度,降低了网络的训练难度。

2　ISTR-LW 模型与方法

这一节将介绍 ISTR-LW 模型的整体框架和 4 个组成部分,包括 STN、特征序列提取层、循环网络层和注意力机制模型。ISTR-LW 模型框架如图 1 所示。

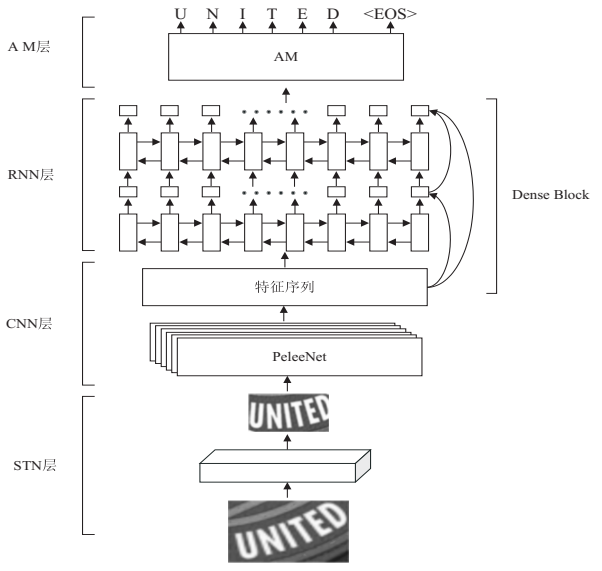

图 1　ISTR-LW 模型

2.1 STN

在场景文本图像中,如图 2(a)所示,存在很多弯曲的文本。如果直接将弯曲的文本输入到网络中,则特征提取阶段就需要学习到形状不变的特征,会增加网络的负担。在框架的初始阶段,为了处理一些弯曲的文本图片,引入了 TPS[28] 变换,作为空间变换网络 STN 的一种变体,如图 2 所示,来解决弯曲文本难以识别的问题。TPS 转换,首先通过定位网络在原图上预测一系列基准点,然后基于预测的基准点通过网格生成器计算转换矩阵,并生成一个关于原图的采样网格,最后采样器结合网格和输入图片,通过采样网格上的点转换得到最终的规则的图片。

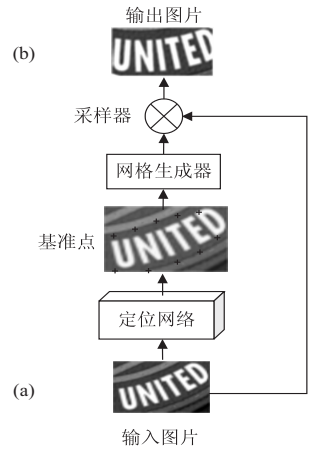

图 2 STN 网络

2.2 特征序列提取层

在传统的场景文本识别中,为了学习到很好的特征,使用参数量巨大的 VGG 网络,不仅计算量巨大,而且内存消耗巨大。为了减少特征序列提取层的计算成本,引入高效的轻量级网络 PeleeNet。图 3 为经过改变后的满足文本特征序列提取的轻量级网络。在文中的 PeleeNet 中,摒弃了原文中的第 4 阶段,将原文中的每个阶段的 2×2 的平均池化层改成 1×2 的平均池化层,并且在最后添加一个 1×1 的卷积层。在不考虑维度的情况下,最终获得一个 $L\times1$ 的特征向量,用 $X=x_1,x_2,\cdots,x_L$ 表示提取的特征序列,每个 x_i 代表特征序列的每一列。由于在特征序列提取时,并不会改变文字的空间结构,所以特征序列顺序与文本的顺序一致。

2.3 循环网络层

将卷积层提取出的特征序列 $X=x_1,x_2,\cdots,x_L$ 输入到循环网络层中,每个 x_i 对应一个 h_i,生成一系列序列 $H=h_1,h_2,\cdots,h_L$。在循环网络层中,以双向长短记忆[22](bidirectional LSTM,Bi-LSTM)为单元,获取文本的左右两边的语义信息。但是,Bi-LSTM 在时间序列上是深度网络,容易产生梯度消失和梯度爆炸的问题,网络收敛较慢,模型训练比较困难。为了缓解这一问题,如图 4 所示,受到 Dense Block 模块的思想启发,在每个输入和输出之间以级联的方式建立一条直接的关联通道。这一设计的引入,有效缓解了梯度消失和梯度爆炸的问题。

阶段	详细		输出
	输入		100 × 32 × 1
阶段 0	Stem Block		25 × 8 × 32
阶段 1	Dense Block	DenseLayer × 3	25 × 4 × 128
	Transition Layer	1 × 1 conv, stride 1 1 × 2 average pool, stride 1 × 2	
阶段 2	Dense Block	DenseLayer × 4	25 ×2 × 256
	Transition Layer	1 × 1 conv, stride 1 1 × 2 average pool, stride 1 × 2	
阶段 3	Dense Block	DenseLayer × 8	25 × 1 ×512
	Transition Layer	1 × 1 conv, stride 1 1 × 2 average pool, stride 1 × 2	
	1 × 1 conv, stride 1		25 × 1 ×256

图 3 特征序列提取网络

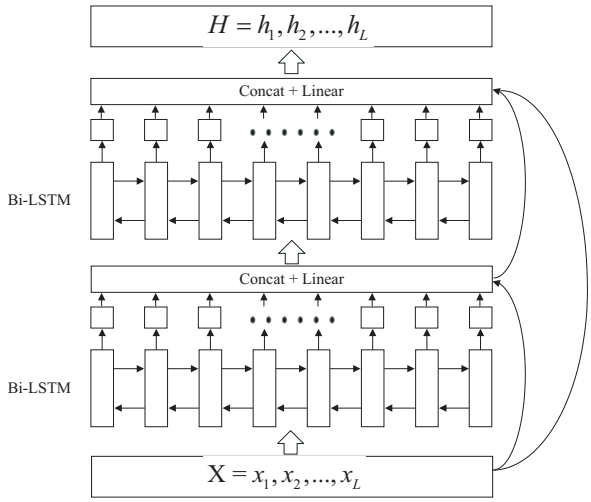

图 4 循环网络层+Dense Block

2.4 注意力机制模型

ISTR-LW 模型在做最终预测时,利用了 RARE[29] 中的注意力机制模型,如图 5 所示,以 GRU[30] 为单位建立的一个预测模型,增加了预测的准确率。在文本区域中,不同的序列对当前字符序列的重要性是不同的,为了获取更加重要的信息,注意力机制模型给不同的序列分配不同的权重,如式(1),计算不同序列的权重:

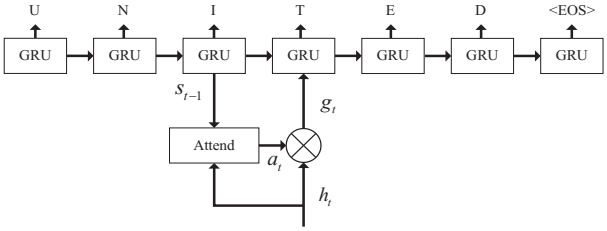

图 5 注意力机制模型

$$a_t=\text{Attend}(s_{t-1},a_{t-1},h) \tag{1}$$

其中,s_{t-1} 为上一个 GRU 的输出。g_t 为当前 GRU 的输

入,g_t 的表示如下:

$$g_t = \sum_{i=1}^{L} a_{ti} h_i \tag{2}$$

2.5　训练过程

最终使用 softmax 函数对文本进行预测,如式(3)所示:

$$y_t = \text{softmax}(W^{\mathrm{T}} s_t) \tag{3}$$

将训练集表示为:$X = \{(I_i, g_i)\}_{i=1,2,\cdots,n}$,其中 I 表示输入的文本,g 表示 ground truth。该文使用最大似然法计算损失函数,如式(4)所示:

$$\text{Loss} = \sum_{i=1}^{N} \log \prod_{t=1}^{|g_t|} p(g_{it} \mid I_i; \theta) \tag{4}$$

其中,$p(.)$ 由 softmax 计算得到,θ 是模型的所有参数。实验表明,优化算法采用 ADADELTA 收敛速度较快。模型参数初始化时,除了 STN 中的定位网络全连接层权值初始化为 0,其他网络权值使用 Kaiming[31] 初始化方法。

3　实　验

3.1　实验细节

在 STN 中的定位网络,初始化基准点的数量为 20,图 6 所示为定位网络的结构,最终输出为 2×20=40 大小的向量。在网络中除了输出层以 tan 为激活函数,其他层的激活函数都为 ReLU。

层	详细		输出
输入			100 × 32
Conv1	channel: 64	kernel: 3 × 3	100 × 32
BN1			100 × 32
Pool1	kernel: 2 × 2	stride: 2 × 2	50 × 16
Conv2	channel: 128	kernel: 3 × 3	50 × 16
BN2			50 × 16
Pool2	kernel: 2 × 2	stride: 2 × 2	25 × 8
Conv3	channel: 256	kernel: 3 × 3	25 × 8
BN3			25 × 8
Pool3	kernel: 2 × 2	stride: 2 × 2	12 × 4
Conv4	channel: 512	kernel: 3 × 3	12 × 4
BN4			12 × 4
AVGPool	512 × 12 ×4		512 × 1
FC1	512		256
FC2	256		2 × 20 = 40

图 6　STN 定位网络

ISTR-LW 模型是在 MJSynth 数据集上训练,该数据集约为 890 万张图片。该文采用 ADADELTA 优化算法,衰退率设置为 0.95。批量大小设置为 192,训练次数为 300 K。训练和测试时所有的输入都被缩放为 100×32 的大小。

实验环境是基于 Pytorch 深度学习框架下实现的,CPU 为 Intel(R) Core(TM) i9-9900 @ 3.50 GHz,显卡为 RTX 2080Ti 11 GB 显存,物理内存为 64 GB,操作系统为 Ubuntu 16.04。

3.2　数据集

训练数据集和测试数据集分别介绍如下:

MJSynth:作为唯一训练数据集。该数据集约有 890 万张合成图片作为训练集,9 万个英文单词,约 90 万张合成图片作为测试集。

IC03:作为测试数据集。该数据集包含 1 110 张不规则的场景文本测试图片,排除少于三个字符或者非字母数字的图片,用于文中测试的图片有 860 张。

IC13:作为测试数据集。该数据集大部分取自 IC03 数据集,共有 1 095 张场景图片,排除少于三个字符或者非字母数字的图片,用于文中的测试图片为 857 张。

IIIT5K:作为测试数据集。该数据集包含 3 000 张不规则的测试图片,全是取自于 Google 图片。

SVT:作为测试数据集。该数据集取自 Google Street View,共有 647 张不规则的场景图片用于文中测试。

3.3　实验结果

在测试时,如表 1 所示,分别在上文介绍的 5 个公开的数据集上进行了测试,并且与 PhotoOcr[16]、Jaderberg[32] 等模型进行了对比,可以看出,ISTR-LW 模型相比其他模型大幅减少了参数量。并且,在 IC03 和 IC13 测试数据集上识别准确度有稍微的提升,在其他数据集上,除了比参数量较大的 RARE 模型识别准确率稍微低一些,相比其他模型都有提升。ISTR-LW 模型比 CRNN 模型大小减小了 29%,比 RARE 模型大小减小了 45%。以上数据说明,引入 PeleeNet 网络和 Dense Block 模块的 ISTR-LW 模型,不仅大幅度减少了模型的参数量,还在文本识别精确率上有很大的竞争力。

表 1　多个模型识别准确率数据对比

数据集	PhotoOCR	Jaderberg	CRNN	RARE	RARE(SRN only)	ISTR-LW	ISTR-LW(no STN)
MJSynth	-	-	93.2	94.8	90.1	93.7	93.0
IC03	-	89.6	89.4	90.1	88.7	91.6	89.1
IC13	87.6	81.8	86.7	88.6	87.5	88.6	88.3
IIIT5K	-	-	78.2	81.9	79.7	79.4	77.6
SVT	78.0	71.7	80.8	81.9	81.5	79.4	78.7
模型大小	-	-	8.3 M	10.8 M	9.2 M	5.9 M	4.2 M

4 结束语

提出的 ISTR-LW 模型,引入了 PeleeNet 网络和 Dense Block 模块,通过 STN 修正变形图片,通过注意力机制让模型获取需要关注的重点区域,加快了网络的收敛速度,在保证识别准确率的前提下,大大减小了模型的规模大小。并且实现了端到端的训练,可以接受任意长度的文本。在公开数据集上的对比实验表明,ISTR-LW 模型的表现具有很强的竞争力。但是 ISTR-LW 模型仍然存在一些缺陷,例如,对一些弯曲度比较大的图片识别率低;模型的准确度上有些许下降。在以后的工作中,将把工作重心放在处理变形图片和提高准确度上。

参考文献:

[1] 牛小明,毕可骏,唐 军. 图文识别技术综述[J]. 中国体视学与图像分析,2019,25(3):241-256.

[2] 程加乐. 基于特征空间的旋转多字体文字识别[D]. 西安:长安大学,2016.

[3] 吴 锐. 自然场景中文本识别技术研究及实现[D]. 哈尔滨:哈尔滨工业大学,2010.

[4] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in neural information processing systems. San Francisco:Morgan Kaufmann,2012:1097-1105.

[5] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[6] GRAVES A,LIWICKI M,FERNÁNDEZ S,et al. A novel connectionist system for unconstrained handwriting recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008,31(5):855-868.

[7] SU B,LU S. Accurate scene text recognition based on recurrent neural network[C]//Asian conference on computer vision. CHAM:Springer,2014:35-48.

[8] 卢欣辰. 自然场景下的文本识别算法研究[D]. 成都:电子科技大学,2019.

[9] 陈 雨. 自然场景下的端到端文本识别[D]. 南京:南京大学,2019.

[10] 周鹏飞. 自然场景图像中的文本检测与识别技术研究[D]. 西安:西安理工大学,2019.

[11] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International conference on learning representations. San Diego,CA,USA:ICLR,2015:1150-1210.

[12] SZEGEDY C,IOFFE S,VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI conference on artificial intelligence. San Francisco,California,USA:AAAI,2017.

[13] GERS F A, SCHRAUDOLPH N N, SCHMIDHUBER J. Learning precise timing with LSTM recurrent networks[J]. Journal of Machine Learning Research,2002,3(1):115-143.

[14] SCHMIDHUBER J, HOCHREITER S. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[15] NOVIKOVA T,BARINOVA O,KOHLI P,et al. Large-lexicon attribute-consistent text recognition in natural images[C]//European conference on computer vision. Berlin,Heidelberg:Springer,2012:752-765.

[16] BISSACCO A,CUMMINS M,NETZER Y,et al. Photoocr:reading text in uncontrolled conditions[C]//Proceedings of the IEEE international conference on computer vision. Sydney,Australia:IEEE,2013:785-792.

[17] WANG K,BABENKO B,BELONGIE S. End-to-end scene text recognition[C]//International conference on computer vision. Barcelona,Spain:IEEE,2011:1457-1464.

[18] SMITH D L,FIELD J,LEARNED-MILLER E. Enforcing similarity constraints with integer programming for better scene text recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Colorado Springs,CO,USA:IEEE,2011:73-80.

[19] DRUCKER H,WU D,VAPNIK V N. Support vector machines for spam categorization[J]. IEEE Transactions on Neural networks,1999,10(5):1048-1054.

[20] WEINMAN J J,LEARNED-MILLER E,HANSON A R. Scene text recognition using similarity and a lexicon with sparse belief propagation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2009,31(10):1733-1746.

[21] SHI C,WANG C,XIAO B,et al. Scene text recognition using part-based tree-structured character detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Portland,OR,USA:IEEE,2013:2961-2968.

[22] SHI B,BAI X,YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,39(11):2298-2304.

[23] JADERBERG M,VEDALDI A,ZISSERMAN A. Deep features for text spotting[C]//European conference on computer vision. CHAM:Springer,2014:512-528.

[24] WANG R J,LI X,LING C X. Pelee:a real-time object detection system on mobile devices[C]//Advances in neural information processing systems. San Francisco:Morgan Kaufmann,2018:1963-1972.

[25] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu,HI,USA:IEEE,2017:4700-4708.

(下转第 29 页)