文章编号：2095-2163(2019)04-0082-05 中图分类号：TP391 文献标志码：A

基于条件深度卷积生成对抗网络的语音增强研究

褚 伟

（华东交通大学 电气与自动化工程学院，南昌 330013）

摘 要：语音交互技术日益在现实生活中得到广泛的应用，由于干扰的存在，现实环境中的语音交互技术远没有达到令人满意的程度。为了提高现实环境中语音交互性能，本文提出了一种基于条件深度卷积生成对抗网络（C-DCGAN）的语音增强模型，这是在 GAN 的基础上加入卷积层和条件信息。C-DCGAN 利用卷积层提取语音特征，同时利用条件信息，生成高质量的语音。通过 TIMIT 数据集、NOISEX-92 噪声库、Aurora2 噪声库及环境噪声数据集对所提出的语音增强模型进行验证。结果表明，与谱减法、DNN 等语音增强方法相比，C-DCGAN 模型在 *PESQ* 和 *STOI* 指标上均有提高，表明本文提出的模型能取得良好的语音增强效果。

关键词：语音增强；条件卷积生成对抗网络；深度学习；带噪语音

Research on Speech Enhancement Model Based on Conditional Deep Convolutional Generative Adversarial Networks

CHU Wei

（School of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China）

【Abstract】 Voice interaction technology is increasingly widely used in real life. Due to the existence of interference, voice interaction technology in real environment is far from satisfactory. In order to improve the performance of speech interaction in real environment, a speech enhancement model based on conditional deep convolutional generative adversarial network (C-DCGAN) is proposed, which adds convolution layers and conditional information to GAN. C-DCGAN uses convolution layers to extract speech features, and uses conditional information to generate high-quality speech. TIMIT database, NOISEX-92 database, Aurora2 database and environmental noise database are used to validate the proposed speech enhancement model. Results show that compared with spectral subtraction and DNN, the C-DCGAN model improves both *PESQ* and *STOI*, and demonstrates that the proposed model can achieve good speech enhancement effect.

【Key words】 speech enhancement; conditional deep convolutional generative adversarial networks; deep neural networks; noisy speech

0 引 言

语音增强是从被干扰的语音信号中提取出纯净的语音信号或者去除复杂的背景噪声，用来改善受噪声污染的语音的质量，提高语音清晰度和可懂度。语音增强作为信号处理中的一个重要研究领域，近些年来受到国内外研究者的广泛关注和重视。

当下的各类相关研究指出，深度神经网络的隐含层数目多，可以更好地提取语音信号中的结构化信息和高维信息。与此同时，这些研究也引发了学界对基于深度学习的语音增强技术的探索热潮。Xu 等人[1]提出了一种基于深度神经网络的语音增强方法。与基于 MMSE 的方法相比，该方法的性能得到了显著的改善，而且能够很好地抑制非平稳噪声。Koizumi 等人[2]提出了一种基于深度神经网络的源增强训练方法，实验表明，该方法可以显著提高语音质量的客观评价指标。基于深度神经网络的方法需要人工提取语音特征，忽略了语音信号时域上的相位信息。但是经分析可知，相位信息对于语音的感知质量是重要的[3]。

GAN 是当前人工智能研究的热点，Goodfellow 等人[4]提出了生成对抗性网络（GAN），并在 MNIST 数据集、CIFAR-10 数据集上进行了实验，结果表明，该方法能应用于图像样本生成。Pascual 等人[5]第一次将生成对抗性网络应用在语音增强中，对模型进行端到端的训练，并证实了模型的有效性。Mirza 等人[6]引入了生成对抗性网络的条件形式，在生成器和判别器中都添加了条件信息。研究结果显示，该模型能够生成以类标签作为条件的 MNIST 数字。

综合前文论述可知，本文采用条件深度卷积生成对抗网络（C-DCGAN）进行语音增强，C-DCGAN

作者简介：褚 伟（1991-），男，硕士研究生，主要研究方向：语音增强、语音识别。

收稿日期：2019-05-25

是在 GAN 的基础上加入卷积层和条件信息。本文在 TIMIT 纯净语音数据库和 3 种不同的噪声库中进行了实验。结果表明，与谱减法、DNN 模型相比，C-DCGAN 模型能取得良好的语音增强效果。本文拟对此展开研究论述如下。

1　C-DCGAN 语音增强模型

本文采用条件深度卷积生成对抗网络（C-DCGAN）模型，将条件信息 c 加入 GAN 的生成器中，条件信息将引导样本数据的生成。与原始的条件生成对抗网络（CGAN）不同，本文所用的判别器中不需要连接条件信息 c。在判别器和生成器中使用卷积层替换池化层，使判别器和生成器变换为全卷积层，利用卷积层提取特征的能力训练网络，改善生成样本的效果。

判别器 D 和生成器 G 使用公式（1）中的目标函数 $V(G,D)$ 来进行极小极大博弈，其数学形式具体如下：

$$\min_G \max_D V(D,G) = E_{x\sim P_{data(x)}}[\log D(x)] + E_{z\sim P_{z(z)}}[\log(1 - D(G(z \mid c)))] \quad (1)$$

其中，$E(\cdot)$ 为期望的计算；x 采样于真实数据分布 $P_{\mathrm{data(x)}}$；z 采样于先验分布 $P_{z(z)}$；映射空间 $G(z;\theta_g)$ 构建于先验噪声分布 $P_{z(z)}$。

C-DCGAN 模型采用交替优化的方法进行训练，对此可表述为：先固定生成器 G，优化判别器 D，使得判别器 D 的判别准确率最大化，即使 D 判别训练样本为 1 和判别生成样本为 0 的概率最大化；然后固定判别器 D，优化生成器 G，使得 D 的判别准确率最小化，即 $\log(1 - D(G(z \mid c)))$ 最小化。在训练过程中，同一轮参数更新中，每优化 k 次判别器，优化 1 次生成器。算法的研发设计流程详见如下。

算法 1　条件深度卷积生成对抗网络算法流程。用小批量随机梯度下降算法训练网络，用于判别器的步骤 k 是一个超参数，文中设置 $k = 2$

for 训练次数 do

　for　k steps　do

　从噪声分布 $p_z(z)$ 中获得 m 个小批量噪声样本 $\{z(1),\ldots,z(m)\}$

　从数据生成分布 $p_{data}(x)$ 中获得 m 个小批量样本 $\{x(1),\ldots,x(m)\}$

　在生成器中加入条件信息 c

　用随机梯度下降法最大化判别器：

$$\nabla_{\theta d}\frac{1}{m}\sum_{i=1}^{m}[\log D(x^{(i)}) + \log(1 - D(G(z^{(i)} \mid c)))]$$

　end for

　从噪声分布 $p_z(z)$ 中获得 m 个小批量噪声样本 $\{z(1),\ldots,z(m)\}$

　在生成器中加入条件信息 c

　用随机梯度下降法最小化生成器：

$$\nabla_{\theta d}\frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z^{(i)} \mid c)))$$

end for

C-DCGAN 模型的工作原理如图 1 所示。由图 1 可知，首先，通过纯净语音数据集和噪声集在多种信噪比下构造混合语音数据集，然后，在 GAN 的基础上加入卷积层，同时在生成器中加入条件信息，从而得到 C-DCGAN 模型。最后，混合语音通过 C-DCGAN 模型生成增强语音，实现语音增强。

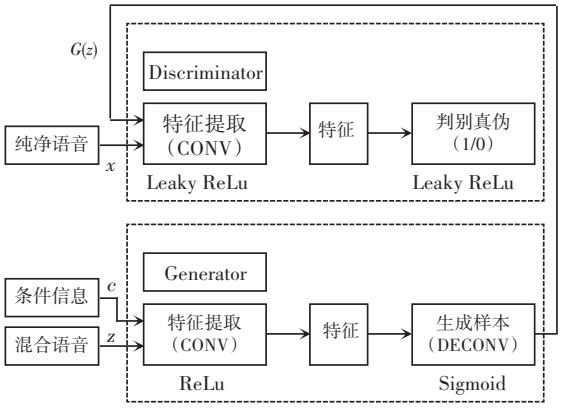


图 1　C-DCGAN 模型原理

Fig. 1　C-DCGAN model principle

2　实验过程与结果分析

2.1　数据集

本次研究使用 TIMIT 语音数据库[7]，NOISEX-92 噪声库[8]、Aurora2 噪声库[9] 和环境噪声数据库[10]。其中，TIMIT 数据集的采样率为 16 kHz，一共包含 6 300 个句子，由 630 个人分别轮流说出给定的 13 个句子组成。NOISEX-92 噪声库包含 15 种常见噪声类型。Aurora2 噪声数据库由 8 种噪声组成。环境噪声数据库是由 100 种常见的环境噪声组成。本文选取 TIMIT 训练集中所有的句子，选取环境噪声库中的 100 种噪声，从 Aurora2 噪声库中选取餐厅嘈杂声（Restaurant）、机场声（Airport）、火车声（Train）、汽车引擎声（Car）、街道声（Street）这 5 种噪声，按信噪比 -5 dB、0 dB、5 dB、10 dB、15 dB、20 dB 混合得到带噪语音，再从中随机选取 100 h 混合语音作为训练集。随机选取 TIMIT 测试集中的 200 个句子，选取 NOISEX-92 噪声库中餐厅内嘈杂

噪声(Babble)、坦克内部噪声(Tank)、高频信道噪声(HFchannel)、驾驶舱噪声(Destroyerengine)这 4 种在训练集中未出现的噪声,按信噪比-5 dB, 0dB, 5 dB,10 dB,15 dB,20 dB 混合得到带噪语音测试集。

2.2 评价指标

本次研究使用的评价指标包括:语音质量听觉评估(*PESQ*)[11] 和短时客观可懂度(*STOI*)[12]。其中,*PESQ* 用来衡量语音质量,取值范围为 -0.5 ~ 4.5, 得分越高说明语音感知效果越好。*STOI* 主要是为了衡量语音的可懂度,其取值范围为 0 ~ 1,得分越高表示语音质量具有越好的可懂度。

2.3 实验环境

本文实验的硬件环境为:TITAN Xp 实验平台,i7-9700k@3.6 GHz CPU,32 G 内存,500 G 固态硬盘。软件环境为:Ubuntu 16.04 操作系统、TensorFlow 框架,编程选用 Python 语言,编辑器为 PyCharm。

2.4 模型参数

为了评估模型的性能,本文实验仿真比较了谱减法、DNN、C-DCGAN 三种语音增强模型。研究可得阐释分述如下。

(1)谱减法模型如下:首先,估计噪声信号的幅度谱。然后,将带噪语音进行傅里叶变换,得到带噪语音的幅度谱。再用带噪语音的幅度谱减去估计出来的噪声幅度谱,就求得了语音的幅度谱估计。最后,利用估计的幅度谱和带噪语音的相位来重构语音信号,而由重构得到的语音信号就是语音增强的结果。

(2)DNN 模型参数如下:先对语音信号进行分帧处理,采用 256 点的汉明窗进行加窗分帧,帧移为 128 点。然后将分帧处理后的语音进行离散傅里叶变换,获得语音的幅值,对幅值取自然对数得到对数能量谱。隐含层数为 3,每个隐含层有 1 024 个神经元。在训练过程中,最初的 10 次迭代过程中,学习速率为 0.1,而在此后的各次迭代时学习速率下降 10%。动量速率 w 为 0.9,迭代次数为 1 000 次。

(3)C-DCGAN 模型参数如下:学习率设为 0.000 2, $batch\_size = 128$, $epochs = 1\ 000$,采用随机梯度下降算法。在训练过程中,每 500 ms 提取约 1 s语音(16 384 个样本)。为避免出现过拟合,在生成器的全连接层加入 Dropout,Dropout 率为 0.5,判别器的全连接层后加入 Dropout,Dropout 率为 0.8。为了防止梯度消失,除了生成器模型的输出层及其对应的判别器模型的输入层外,其它层都使用了批量归一化。

2.5 结果分析

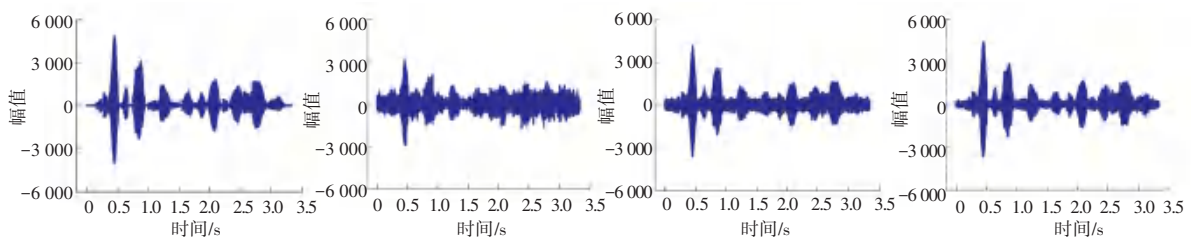
谱减法、DNN 和 C-DCGAN 三种模型在含有 105 种噪声的训练集中训练,在含有 4 种不可见噪声的测试集中的测试结果见表 1。由表 1 可以看出,C-DCGAN 模型相对于谱减法,*PESQ* 和 *STOI* 的平均值分别提高 0.25 和 0.05。C-DCGAN 模型相对于 DNN 模型,*PESQ* 和 *STOI* 的平均值分别提高 0.13 和 0.03,表明 C-DCGAN 模型明显优于谱减法和 DNN 模型,语音感知效果和语音可懂度得到了提高。

表 1 谱减法、DNN 和 C-DCGAN 模型在 105 种噪声条件下的评估结果

Tab. 1 Evaluation results of spectral subtraction, DNN and C-DCGAN models under 105 noise conditions

SNR	带噪语音		谱减法		DNN		C-DCGAN	
	PESQ	*STOI*	*PESQ*	*STOI*	*PESQ*	*STOI*	*PESQ*	*STOI*
20	2.87	0.96	3.12	0.96	3.34	0.97	3.46	0.98
15	2.54	0.93	2.95	0.93	3.13	0.95	3.24	0.96
10	2.18	0.88	2.74	0.90	2.85	0.92	2.95	0.94
5	1.87	0.79	2.45	0.81	2.52	0.85	2.63	0.90
0	1.51	0.65	2.07	0.74	2.15	0.78	2.33	0.82
-5	1.28	0.55	1.70	0.62	1.76	0.65	1.95	0.68
Ave	2.04	0.79	2.51	0.83	2.63	0.85	2.76	0.88

选取 TIMIT 中训练集的 sa1.wav 纯净语音文件,其内容为“She had your dark suit in greasy wash water all year”,选取 NOISEX-92 中 babble 噪声。将纯净语音和噪声按信噪比 $SNR = 0$ 的方式生成带噪语音,再对模型进行测试。纯净语音和增强语音的波形如图 2 所示。从图 2 可以看出,经过谱减法增强后的语音能够减少噪声信号,但产生了较为明显的失真,影响了听觉感受。经过 DNN 模型增强后的语音能够在相当程度上减少噪声信号,但还会残留一定的噪声信号。经过 C-DCGAN 模型增强后的语音最接近纯净语音信号。



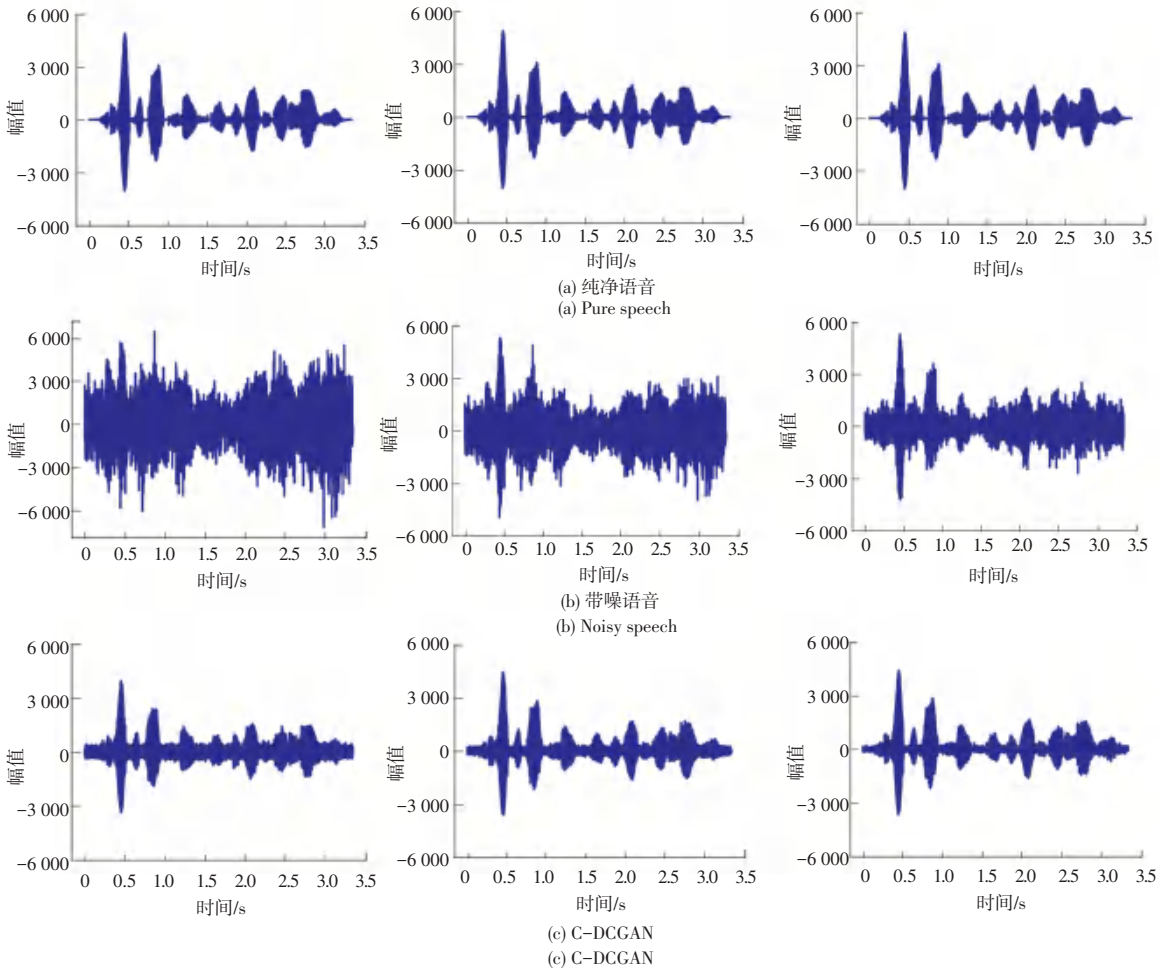
(a) 纯净语音　(b) 谱减法　(c) DNN　(d) C-DCGAN

(a) Pure speech　(b) Spectral subtraction　(c) DNN　(d) C-DCGAN

图 2　babble 噪声条件下 3 种方法对比

Fig. 2　Comparison of three methods under babble noise conditions

为了测试 C-DCGAN 模型在低信噪比下的语音增强性能，在−10 dB、−5 dB、0 dB 等 3 种不同信噪比条件下进行实验。选取 TIMIT 中的 sa1.wav 纯净语音以及 NOISEX-92 中 babble 噪声。将纯净语音和噪声分别在信噪比−10 dB、−5 dB、0 dB 条件下混合，得到带噪语音。并将带噪语音在训练好的 C-DCGAN 模型上进行测试。C-DCGAN 模型测试结果如图 3 所示。图 3(a)表示纯净语音，图 3(b)从左到右分别表示信噪比为−10 dB、−5 dB、0 dB 下的混合语音，图 3(c)从左到右分别表示各个信噪比下 C-DCGAN 模型的语音增强效果。由图 3 可知，C-DCGAN 模型能够在较低信噪比下实现语音增强，并取得良好的效果。



(a) 纯净语音
(a) Pure speech

(b) 带噪语音
(b) Noisy speech

(c) C-DCGAN
(c) C-DCGAN

图 3　3 种信噪比下语音增强效果

Fig. 3　Speech enhancement effect under three signal-to-noise ratios

3 结束语

本文提出了条件深度卷积生成对抗网络(C-DCGAN)模型,利用条件信息以及卷积层提取特征的能力生成高质量的纯净语音,从而实现语音增强。对于深度学习模型,含有大量噪声的训练集对于学习语音特征至关重要。本文在 TIMIT 数据集和不同噪声集中进行了实验,结果表明,相对于谱减法、DNN 模型,C-DCGAN 模型的语音听觉质量和语音可懂度都有提高。

参考文献

[1] XU Yong , DU Jun, DAI Lirong , et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1):65-68.

[2] KOIZUMI Y , NIWA K , HIOKA Y , et al. DNN-based source enhancement to increase objective sound quality assessment score [J]. IEEE/ACM Transactions on Audio, Speech & Language Processing, 2018,26(10):1780-1792.

[3]PALIWAL K, WÓJCICKI K, SHANNON B. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(4): 465-494.

[4] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. USA:MIT Press, 2014: 2672-2680.

[5] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: Speech enhancement generative adversarial network[J]. arXiv preprint arXiv:1703.09452,2017.

[6] MIRZA M, OSINDERO S. Conditional generative adversarial nets [J]. arXiv preprint arXiv:1411.1784,2014.

[7] GAROFOLO J S. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database[R]. Gaithersburgh, MD: National Institute of Standards and Technology (NIST), 1988.

[8] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX - 92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.

[9]PEARCE D,HIRSCH H G. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions[C]//Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000. Beijing, China:dblp, 2000:1-5.

[10] HU G. 100 nonspeech environmental sounds, 2004 [EB/OL]. [2017-12-04]. http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[11]ITU-T Recommendation P.862 . Perceptual evaluation of speech quality (PESQ) : An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[S]. Geneva: International Telecommunication Union-Telecommunication Standardisation Sector,2001.

[12]TAAL C H , HENDRIKS R C , HEUSDENS R , et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech [C]// 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).Dallas, TX, USA: IEEE, 2010:4214-4217.

(上接第 81 页)

高、技术难度大,本文搭建了基于双目视觉的泊车机器人障碍物检测系统,采用控制变量法完成双目标定以得到较高精度的焦距,利用改进立体匹配算法完成立体匹配,引入 YOLO 卷积神经网络完成障碍物类别检测,最终输出障碍物的类别和距离。由于国内外相关研究较少,智能车库环境下的检测算法不成熟,硬件成本高,如何制造出低成本、高效率的经济型泊车机器人将是未来研究的重点和难点。

参考文献

[1] 徐欣, 周香琴, 江先志,等. 基于物联网技术的小区停车位共享平台的设计与开发[J]. 工业控制计算机, 2018, 31(1):139-141.

[2] 申爱萍. 揭开"最牛泊车机器人"的神秘面纱[J]. 驾驶园, 2017(9):42-43.

[3] 刘爽. 基于二维码识别的自动泊车机器人定位导航技术研究[D]. 武汉: 华中科技大学, 2017.

[4] 研华科技.怡丰机器人携手研华: 以技术力量, 扩展 AGV 市场应用[J]. 自动化博览, 2018, 35(8):76-77.

[5] 魏言华. 基于视觉的车辆后方障碍物检测算法研究与实现[D]. 沈阳: 东北大学, 2008.

[6] 谢若冰. 双目立体成像和显示的 FPGA 视频处理技术研究[D]. 北京: 北京理工大学, 2015.

[7] 叶峰, 王敏, 陈剑东,等. 共面点的摄像机非线性畸变校正[J]. 光学精密工程, 2015, 23(10):2962-2970.

[8] Shubham Shinde, Ashwin Kothari, Vikram Gupta. YOLO based Human Action Recognition and Localization [J]. Procedia Computer Science, 2018, 20(3): 133-134.

[9] 王昊. 基于卷积神经网络的目标检测与识别方法研究[D]. 南京: 南京财经大学, 2016.

[10]施泽浩, 赵启军. 基于全卷积网络的目标检测算法[J]. 计算机技术与发展, 2018, 28(5):55-58.

[11]杨梓豪. 基于区域卷积神经网络的物体识别算法研究[D]. 北京: 北京邮电大学, 2017.

[12]李旭冬. 基于卷积神经网络的目标检测若干问题研究[D]. 成都: 电子科技大学, 2017.