引用格式：曹中辉，黄志华，葛文萍，等. 注意力机制对生成对抗网络语音增强迁移学习模型的影响[J]. 声学技术, 2021, **40**(1): 77-81. [CAO Zhonghui, HUANG Zhihua, GE Wenping, et al. Influence of attention mechanism on generative adversarial network speech enhancement transfer learning model[J]. Technical Acoustics, 2021, **40**(1): 77-81.] **DOI**: 10.16300/j.cnki.1000-3630.2021.01.012

注意力机制对生成对抗网络语音增强迁移学习模型的影响

曹中辉，黄志华，葛文萍，黄 浩
(新疆大学信息科学与工程学院，信号检测与处理新疆维吾尔自治区重点实验室，新疆乌鲁木齐 830001)

摘要：基于深度学习的语音增强模型对训练集外语言语音和噪声进行降噪时，性能明显下降。为了解决这一问题，提出一种引入注意力机制的生成对抗网络(Generative Adversarial Network, GAN)语音增强迁移学习模型。在生成对抗语音增强模型的判别模型中引入注意力机制，以高资源场景下的大量语音数据训练得到的语音增强模型为基础增强模型，结合低资源场景下的少量语音训练数据，对基础增强模型进行权重迁移，提升低资源场景下语音增强模型的增强效果。实验结果表明，采用注意力机制的生成对抗语音增强迁移学习模型，对低资源场景下的带噪语音和集外噪声可以进行有效的降噪。

关键词：生成对抗网络(GAN)；语音增强；迁移学习；跨语言语音增强；注意力机制
中图分类号：H107 **文献标志码：**A **文章编号：**1000-3630(2021)-01-0077-05

Influence of attention mechanism on generative adversarial network speech enhancement transfer learning model

CAO Zhonghui, HUANG Zhihua, GE Wenping, HUANG Hao
(*College of Information Science and Engineering, Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjaing University, Urumqi* 830001, *Xinjiang, China*)

Abstract: The deep learning based speech enhancement model encounters the problem of enhancement performance degradation when de-noising the unseen languages and noise in training sets. In order to solve this problem, a generative adversarial network (GAN) speech enhancement transfer learning model with attention mechanism (called ATGAN speech enhancement model) is proposed in this paper. The attention mechanism is introduced into the discriminator of GAN speech enhancement model. Based on the well-trained model obtained with high-resource materials and combining a small amount of speech training data in low-resource condition, the weight transfer of the basic enhancement model trained with low-resource data is carried out to improve the enhancement effect in low-resource condition. Experiments show that the use of ATGAN speech enhancement model can effectively enhance the de-noising effect of low-resource noisy speech.

Key words: generative adversarial network (GAN); speech enhancement; transfer learning; cross-language speech enhancement; attention mechanism

0 引 言

语音增强[1]是从带噪信号中恢复出原始信号的一种信号处理技术。谱减法、维纳滤波等基于统计模型的方法是语音增强领域中广泛使用的经典方法[2-4]，但是传统语音增强方法对于非平稳噪声的增强效果有限。近些年来，深度学习技术在语音增强领域取得显著进步，基于降噪自编码器，深度神经网络(Deep Neural Network, DNN)、卷积神经网络(Convolutional Neural Network, CNN)、长短时记忆网络(Long Short-Term Memory, LSTM)的语音增强方法先后被提出[5-8]，这些基于深度神经网络的增强模型能有效抑制非平稳噪声。2014 年，Goodfellow 等[9]提出生成对抗网络(Generative Adversarial Network, GAN)。2017 年，Santiago 等[10]将 GAN 应用在语音增强上，提出一种端到端的 GAN 语音增强框架(Speech Enhancement GAN, SEGAN)，在客观和主观测评指标上均优于传统维纳滤波方法。Daniel 等[11]提出条件 GAN (Conditional

收稿日期：2019-12-06; **修回日期：**2020-02-03
基金项目：新疆维吾尔自治区自然科学基金项目资助(2017D01C044)
作者简介：曹中辉(1996－)，男，新疆库尔勒人，硕士研究生，研究方向为语音信号处理。
通信作者：黄志华，E-mail: echohzh@163.com

GAN, cGAN)结构进行语音增强，测评结果在主观语音质量评估(Perceptual Evaluation of Speech Quality, PESQ)指标上优于基于最小均方误差的短时幅度谱增强方法(Short-time Spectral Amplitude Minimum Mean Square Error, STSA-MMSE)和基于DNN的理想比值掩模(Ideal Ratio Mask, IRM)增强算法。2018年，Li等[12]将GAN应用在语音去混响上，与权重预测误差(Weighted Prediction Error, WPE)系统和基于DNN的去混响方法相比，PESQ和语音混响调制能量比(Speech to Reverberation Modulation Energy Ratio, SRMR)值更高。现有增强方法虽然取得有效的增强效果，但均采用单一语言数据对增强模型进行训练，并未探讨单一语言增强模型对新语言语音的增强效果。2014年，Xu等[13]对基于DNN语音增强框架进行模型迁移实现了跨语言语音增强，对于低资源新语言语音的增强效果优于低资源单语言语音训练出的模型。2017年，Santiago等[14]用SEGAN迁移学习模型对新语言带噪语音进行去噪，采用英语单语言增强模型对网络进行参数初始化，低资源语音采用韩语和加泰罗尼亚语，以迁移学习的方式训练SEGAN，对低资源带噪语音的去噪效果与直接用低资源语音数据训练的SEGAN相比，在评测指标分段信噪比(Segmental Signal Noise Ratio, SSNR)上提升了10 dB，PESQ值提升了将近1。

研究表明，在卷积神经网络中引入注意力机制可进一步提升网络的分类准确性[15-16]。本文提出一种在生成对抗网络中引入注意力机制的迁移学习模型(Attention Transfer Learning Generative adversarial Network, ATGAN)，有效提高了低资源语言场景下少量语音的去噪效果。

1 GAN语音增强

GAN是一种基于生成对抗思想训练的神经网络模型，由生成模型(Generator) G 和判别模型(Discriminator) D 两部分组成。GAN的结构图如图1所示。

G 将随机噪声生成尽可能符合真实数据分布的数据，D 负责区分输入数据是 G 生成的数据还是真实数据。对于给定的真实数据 x，D 为其打上标签1；对于给定的生成数据 $G(n)$，D 为其打上标签0。在对抗训练过程中，传给 D 的生成数据 $G(n)$，则尽可能让 D 为其打上标签1。D 将判决结果误差传递给 G 模型，直到 D 对于给定数据预测为真的概率逼近0.5，达到纳什均衡。这一过程可表示为[10]

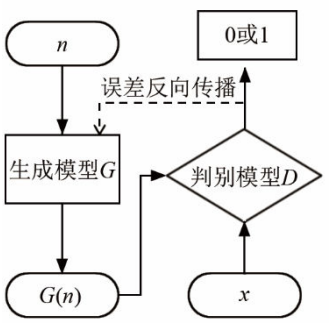


图1 生成对抗网络(GAN)的结构图
Fig.1 GAN structure diagram

$$\min_D \max_G f(G,D)=\mathrm{E}_{x\sim p_{\mathrm{data}}(x)}[\log_2 D(x)]+\mathrm{E}_{n\sim p_n(n)} \log_2 \{1-D[G(n)]\} \tag{1}$$

其中：n 表示噪声，x 为真实数据。为了更好地控制生成数据的质量，常在 G 和 D 中加入条件 y，此时目标函数为

$$\min_D \max_G f(G,D)=\mathrm{E}_{x,y\sim p_{\mathrm{data}}(x,y)}[\log_2 D(x,y)]+\mathrm{E}_{n\sim p_n(n),y\sim p_{\mathrm{data}}(y)} \log_2 \{1-D[G(n,y),y]\} \tag{2}$$

GAN语音增强模型中的 G 即为语音增强部分，可由CNN或者LSTM网络构成。干净语音为 x，n 为带噪语音，达到均衡后的 G 输出即为增强后的语音。

2 引入注意力机制的生成对抗网络语音增强迁移学习模型

迁移学习是将模型在某一领域学到的知识迁移到相近或者不同领域的技术。迁移学习使模型能够在已有知识的基础上快速有效解决新目标域的问题，其在机器学习和数据挖掘领域具有重要研究价值[17]。本文提出一种在GAN网络中引入注意力机制的GAN语音增强迁移学习模型(ATGAN)，进一步提高GAN语音增强迁移学习模型对低资源带噪语音的去噪效果。ATGAN语音增强模型注意力机制示意图如图2所示。

给定输入特征图 F，通过通道注意力模块，得到通道注意力权重 $C(F)$，然后与输入特征图相乘，

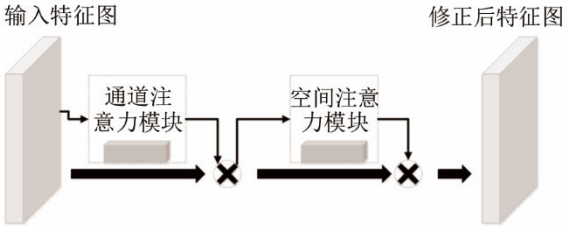


图2 ATGAN语音增强模型注意力机制示意图
Fig.2 Attention mechanism in ATGAN speech enhancement model

结果 F_1 送入空间注意力模块，得到空间注意力权重 $S(F_1)$，与中间输入 F_1 相乘，得到修正后的特征图 F_2，数学描述为

$$F_1=C(F)\otimes F, \tag{3}$$

$$F_2=S(F_1)\otimes F_1 \tag{4}$$

式(3)、(4)中的⊗表示点乘。C 表示通道注意力模块映射函数，S 表示空间和注意力模块映射函数。

生成模型 G 由 22 层包含跳跃连接的对称 U 型卷积和反卷积层构成[18]。网络结构如图 3 所示。

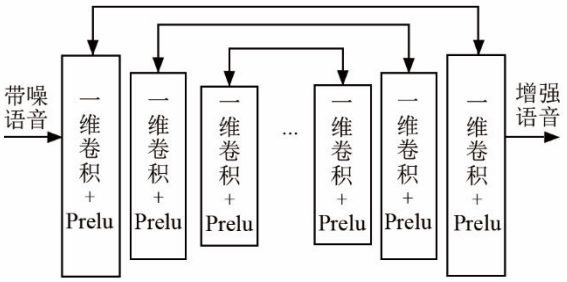


图 3　ATGAN 语音增强模型生成模型 G 的结构

Fig.3　Structure of generator G in ATGAN speech enhancement model

音频数据经过预处理，送入 G 的维度为 16 384×1，卷积操作为一维卷积，激活函数为 Prelu，卷积核宽为 31，步长为 2。卷积部分结束输出维度为 8×1 024，然后从相应维度的标准正态分布中采样，与卷积结果拼接，送入与卷积部分对称的反卷积网络。

D 的结构如图 4 所示，由编码和注意力模块组成，编码部分为 9 层下采样卷积层，由一维反卷积和正则化层构成，激活函数为 Lrelu，卷积核大小为 31，步长为 2。下采样结束得到 8×1 024 维度的编码结果，送入注意力模块，经过最大池化和平均池化以及 sigmoid 函数操作，得到经通道注意力权重修正后的特征图，然后结果经过最大池化和平均池化处理后进行拼接，再送入一维卷积，卷积核大小为 7，个数为 1，得到经空间注意力权重修正的特征图，最后得到更为准确的分类结果，流程图如图 5 所示。

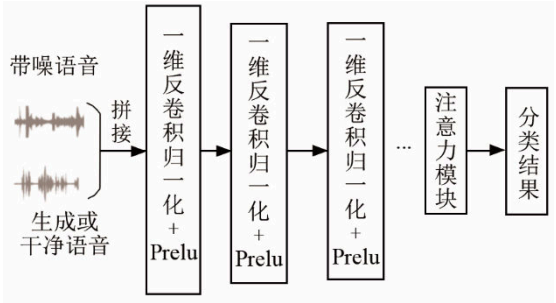


图 4　ATGAN 语音增强模型判别模型 D 的结构

Fig.4　Structure of discriminator D in ATGAN speech enhancement model

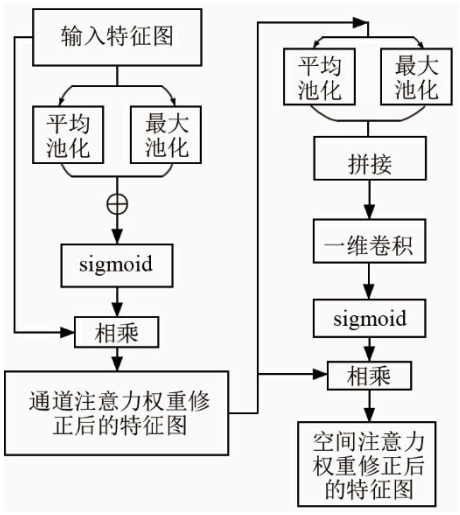


图 5　注意力模块结构流程图

Fig.5　Structure flowchart of attention module

参考文献[19]，损失函数(loss)设计如下：

$$\min_G f(G)=\frac{1}{2}\mathrm{E}_{n\sim p_n(n),y\sim p_{\mathrm{data}}(y)}\log_2(D(G(n,y),y)-1)^2 \tag{5}$$

$$\min_D f(D)=\frac{1}{2}\mathrm{E}_{x,y\sim p_{\mathrm{data}}(x,y)}(D(x,y)-1)^2+\frac{1}{2}\mathrm{E}_{n\sim p_n(n),y\sim p_{\mathrm{data}}(y)}(D(G(n,y),y)^2 \tag{6}$$

生成对抗网络引入注意力机制后，通过高资源(文中的资源是指训练模型数据资源的丰富程度，高资源指训练数据充足的场景，低资源是指训练数据非常少的场景，直接采用低资源场景下的训练模型无法达到较好的增强效果)语音数据训练得到网络权重参数更为合理的预训练模型，然后采用低资源场景下的少量语音数据，对预训练模型进行权重迁移，得到引入注意力机制的 GAN 语音增强迁移学习模型。

3　实验与讨论

3.1　数据集准备与网络参数设置

为了评估和分析本文提出的 ATGAN 语音增强模型对低资源语音的去噪效果，采用英语数据训练的模型迁移到对维吾尔语进行增强的 ATGAN 上。英语数据集采用 Voice Bank 语料库[20]，训练集由 28 位说话人组成，包括 14 位男性、14 位女性；为了获得带噪语音数据集，从 Demand 数据集中选择 kitchen，field，washing，station，river，park，hallway，meeting，restaurant，traffic，metro 11 种噪声[21]，分别以 0、5、10、15 dB 的信噪比与干净语音合成，得到带噪语音训练集，共 11 572 条。维吾尔语数据集采用 THUYG-20[22]，带噪语音训练集的合成方法

及噪声条件与英语带噪语音一致，共 300 条维吾尔语带噪语音；测试集从 Demand 数据集中选择 bus，cafeteria，square，living，office 5 种噪声类型(不在训练集内)，以 2.5、7.5、12.5、17.5 dB 的信噪比与干净语音合成得到。

ATGAN 网络参数设置如下：学习率为 0.000 2，批大小为 100，迭代期数(epoch)大小为 340。优化算法采用 RMSprop 算法[23]。

为了评估 ATGAN 语音增强模型的去噪效果，我们采用对数谱距离(Log Spectral Distance, LSD)，PESQ、短时客观可懂度(Short-Time Objective Intelligibility, STOI) 3 种客观评价指标，LSD 越小，表明增强效果越好，PESQ 和 STOI 越大，表明增强效果越好。

3.2 ATGAN 语音增强模型去噪效果

为了评估 ATGAN 语音增强模型的去噪性能，基线模型采用迁移学习 SEGAN(TSEGAN)作为对比实验算法，实验结果如表 1～3 所示。从表中可看出，ATGAN 语音增强模型增强效果均优于 TSEGAN 模型，ATGAN 可进一步提升对低资源带噪语音的增强效果，语音的客观质量、感知效果和可懂度均有提高。分析认为，在迁移学习生成对抗网络中引入注意力机制，经语音数据训练得到的预训练模型的权重参数更为合理，然后进行权重迁移，注意力机制有助于生成模型重点关注和捕获噪声与纯净语音之间的关系，降低语种因素对重建干净语音的影响。

表 1 ATGAN 和 TSEGAN 的 LSD 指标比较
Table 1 LSD comparison between ATGAN and TSEGAN

模型	不同信噪比时 LSD 指标			
	2.5 dB	7.5 dB	12.5 dB	17.5 dB
noisy	1.585 2	1.328 6	1.063 4	0.824 8
TSEGAN	1.259 0	1.136 4	1.004 8	0.890 7
ATGAN	1.178 3	1.026 4	0.904 8	0.797 2

表 2 ATGAN 和 TSEGAN 的 PESQ 指标比较
Table 2 PESQ comparison between ATGAN and TSEGAN

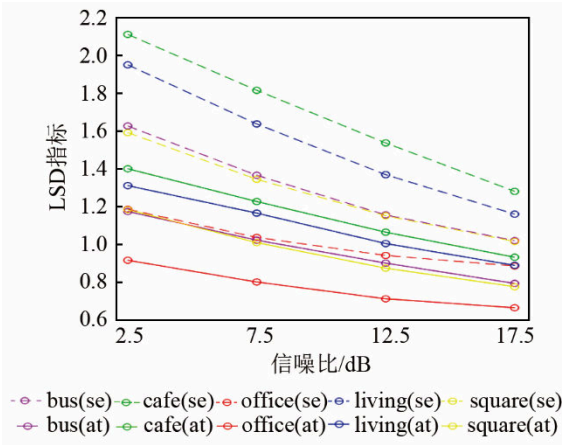
模型	不同信噪比时 PESQ 指标			
	2.5 dB	7.5 dB	12.5 dB	17.5 dB
noisy	1.686 4	2.120 8	2.525 1	2.906 2
TSEGAN	1.926 4	2.430 9	2.835 7	3.124 3
ATGAN	2.079 8	2.532 4	2.922 2	3.254 8

表 3 ATGAN 和 TSEGAN 的 STOI 指标比较
Table 3 STOI comparison between ATGAN and TSEGAN

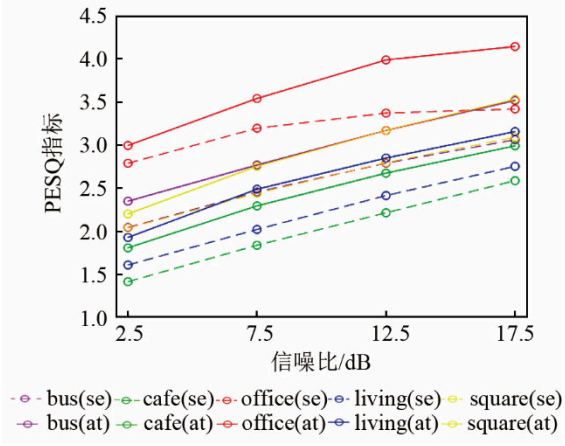
模型	不同信噪比时 STOI 指标/%			
	2.5 dB	7.5 dB	12.5 dB	17.5 dB
noisy	72.02	81.51	88.91	94.26
TSEGAN	78.01	85.60	91.68	95.32
ATGAN	78.54	86.34	91.72	95.36

3.3 ATGAN 语音增强模型的噪声迁移鲁棒性能分析

为了验证 ATGAN 语音增强模型的增强性能对不同类型、不同信噪比噪声的迁移鲁棒性，分析了五种噪声在四种信噪比下的增强效果，结果如图 6 所示，图中实线表示 ATGAN 语音增强模型的去噪结果，图注中用(at)表示，虚线表示直接采用 SEGAN 模型训练的得到的去噪结果，图注中用(se)表示。从图 6 中的 LSD 和 PESQ 指标可看出，对于 bus，office，square 噪声，模型的增强结果最优，而 cafe 噪声的迁移效果最差。通过频谱分析，发现 bus 噪声的能量主要分布在 0～1 000 Hz 频率段，而 cafe 噪声不仅在 0～1 000 Hz 的频率段内能量较高，在 1 000～2 000 Hz 内也具有较高的能量，而且分布更为均匀，这可能是两种噪声迁移去噪效果有差别的原因之一。从测试曲线图中还可看出，信噪比越低，模型的提升效果越明显。



(a) LSD 指标



(b) PESQ 指标

图 6 ATGAN 语音增强模型对不同噪声的去噪效果
Fig.6 Denoising effects of ATGAN speech enhancement model on different noises

4 结 论

本文提出一种引入注意力机制的 GAN 语音增强迁移学习模型，利用已有语言语音训练的增强模型，再结合极少量的新语言语音资源对模型进行训练，可以对新语言低信噪比语音进行有效增强，提高增强后语音的质量。同时，训练 GAN 语音增强模型的时间和所需数据量均大大减少。实验结果表明，ATGAN 语音增强模型相对于 SEGAN 迁移学习模型，去噪后语音的感知质量和可懂度都有进一步提升。本文也讨论了 ATGAN 在不同信噪比下对不同噪声的迁移增强性能，结果表明，ATGAN 对集外噪声有更好的去噪效果。本文结论可为建立低资源新语言场景下的语音增强模型提供参考。在今后的工作中，将进一步研究采用生成对抗网络不同层进行权重迁移对语音增强效果的影响。

参 考 文 献

[1] LOIZOU P C. Speech enhancement[M]. Boca Raton: CRC Press, 2013.

[2] LIM J S, OPPENHEIM A V. All-pole modeling of degraded speech[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978, **26**(3): 197-210.

[3] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, **27**(2): 113-120.

[4] EPHRAIM Y. Statistical-model-based speech enhancement systems[J]. Proceedings of the IEEE, 1992, **80**(10): 1526-1555.

[5] LU X G, TSAO Y, MATSUDA S, et al. Speech enhancement based on deep denoising Autoencoder[C]//Conference of the International Speech Communication Association, 2013: 436-440.

[6] XU Y, DU J, DAI L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, **23**(1): 7-19.

[7] KOUNOVSKY T, MALEK J. Single channel speech en-hancement using convolutional neural network[C]//2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM). Donostia-San Sebastian, Spain. IEEE, 2017: 1-5.

[8] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[C]//Latent Variable Analysis and Signal Separation, 2015.

[9] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2.2014: 2672-2680.

[10] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: speech enhancement generative adversarial network[C]//Conference of the International Speech Communication Association 2017. ISCA: ISCA, 2017: 3642-3646.

[11] MICHELSANTI D, TAN Z H. Conditional generative adver-sarial networks for speech enhancement and noise-robust speaker verification[C]//Interspeech 2017. ISCA: ISCA, 2017: 2008-2012.

[12] LI C X, WANG T Q, XU S, et al. Single-channel speech dere-verberation via generative adversarial training[C]//Conference of the International Speech Communication Association 2018. ISCA: ISCA, 2018: 1309-1313.

[13] XU Y, DU J, DAI L R, et al. Cross-language transfer learning for deep neural network based speech enhancement[C]//The 9th International Symposium on Chinese Spoken Language Pro-cessing. Singapore, Singapore. IEEE, 2014: 336-340.

[14] PASCUAL S, PARK M, SERRÀ J, et al. Language and noise transfer in speech enhancement generative adversarial network[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5019-5023.

[15] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. Salt Lake City, UT, USA. IEEE, 2017: 2011-2023.

[16] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision(ECCV), 2018: 3-19.

[17] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, **22**(10): 1345-1359.

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015, 2015.

[19] MAO X D, LI Q, XIE H R, et al. Least squares generative ad-versarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. IEEE, 2017: 2813-2821.

[20] VEAUX C, YAMAGISHI J, KING S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database[C]//2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). Gurgaon, India. IEEE, 2013: 1-4.

[21] THIEMANN J, ITO N, VINCENT E. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): a database of multichannel environmental noise recordings[C]//Montreal, Canada. ASA, 2013.

[22] 艾斯卡尔·肉孜, 殷实, 张之勇, 等. THUYG-20: 免费的维吾尔语语音数据库[J]. 清华大学学报(自然科学版), 2017, **57**(2): 182-187.
Aisikaer Rouzi, YIN Shi, ZHANG Zhiyong, et al. THUYG-20: a free Uyghur speech database[J]. Journal of Tsinghua University (Science and Technology), 2017, **57**(2): 182-187.

[23] TIELEMAN T, HINTON G. Lecture 6.5-RMSprop: divide the gradient by a running average of its recent magnitude[Z]. COURSERA: Neural Networks for Machine Learning, 2012.