文章编号：2095-2163(2021)03-0209-06 中图分类号：TP183 文献标志码：A

基于听觉掩蔽生成对抗网络的单通道语音增强方法

杜志浩，韩纪庆
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：为提高语音识别系统在复杂声学场景下的识别率，出现了以单通道语音增强（Monaural Speech Enhancement）技术作为前端处理的鲁棒语音识别系统。尽管现有的单通道语音增强技术能够提高混响干扰下的识别率，却未能显著提升宽带非平稳噪声干扰下的系统识别率。为此，本文提出基于听觉掩蔽生成对抗网络的单通道增强方法，通过听觉掩蔽增强模型和判别器构成的对抗过程，来使增强后的语音特征满足目标语音的概率分布。实验结果表明，就语音识别率而言，所提出的听觉掩蔽生成对抗网络超越了现有的增强方法，能够相对减少19.50%的词错误率，显著提升语音识别系统的噪声鲁棒性。

关键词：听觉掩蔽；生成对抗网络；单通道语音增强；鲁棒语音识别

Adversarial generative network based on auditory masking for monaural speech enhancement

DU Zhihao, HAN Jiqing
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 To improve the accuracy of speech recognition system in the complex acoustic scene, monaural speech enhancement method is involved into the robust automatic speech recognition (ASR) system as a front-end processing. Although monaural speech enhancement has improved the recognition performance under the reverberant conditions, it failed to improve the accuracy of speeches interrupted by the wide-band non-stationary noises. To overcome this problem, the paper proposes the adversarial generative network based on auditory masking for monaural speech enhancement. Through the adversarial process between a discriminator and a masking-based enhancement model, the proposed method can make the enhanced speech features follow the distribution of target speeches. Experimental results show that, the proposed method outperforms current enhancement method in terms of recognition accuracy. It achieves 19.50% relative word error rate (*WER*) reduction for a robust ASR system, which indicates that the proposed method can further improve the noise robustness.

【Key words】 auditory masking; adversarial generative network; monaural speech enhancement; robust speech recognition

0 引 言

在语音识别领域，通常使用含有噪声的语音和纯净语音共同训练自动语音识别（Automatic Speech Recognition，ASR）系统，从而提高其在噪声环境下的识别率。为进一步提高 ASR 系统在复杂声学场景下的识别率，出现了将单通道语音增强模型作为前端处理的识别系统。这类系统先使用增强模型尽可能地去除含噪语音中的噪声干扰，接着将增强后的语音送入 ASR 系统，从而得到识别文本。为使增强后的语音和识别系统所要求的输入尽可能地匹配，通常需要利用增强后的语音重新训练 ASR 系统中的声学模型[1-2]，或者将声学模型和增强模型堆叠，进行联合训练[3-4]。然而，上述这 2 种方法在增强模型改变时需要重新训练语音识别系统，对实际应用而言，这通常是不合理的。首先，重新训练语音识别系统非常耗时，其次，增强模型一般运行在终端设备，而识别系统则通常运行在云端设备，可能无法对两者进行联合优化。

近年来，出现了基于生成对抗网络（Adversarial Generative Network，GAN）[5] 的单通道语音增强方法。该方法通过构建增强模型和判别器之间的对抗过程，来使增强后的语音满足目标语音分布。基于 GAN 的增强方法能够显著提升增强后语音的可懂度和感知质量[6]。受此启发，鲁棒语音识别领域也出现了基于 GAN 的前端处理方法，以尽可能地减少增强模型输出与识别系统所要求输入之间的不匹配程度，从而直接提高增强后语音的识别率，而不需要联合训练或重新训练声学模型[7-8]。通过增强后语音特征和目标语音特征之间的对抗训练，文献[7]

基金项目：国家重点研发项目（2017YFB1002102）。

作者简介：杜志浩（1993-），男，博士研究生，主要研究方向：单通道语音增强、鲁棒语音识别；韩纪庆（1964-），男，博士，教授，博士生导师，主要研究方向：语音信号处理、音频信息处理。

收稿日期：2020-11-03

中的增强方法降低了混响干扰下语音识别系统14%~19%的相对字错误率。在文献[8]中,经过对抗训练的增强模型能够显著提升纯净语音训练的ASR系统,却未能进一步提高鲁棒ASR系统的识别率。

为了进一步提高鲁棒ASR系统在宽带非平稳噪声干扰下的识别率,本文提出基于听觉掩蔽生成对抗网络的单通道语音增强方法,并将其作为鲁棒ASR系统的前端处理过程,以尽可能地去除背景噪声的干扰,从而提高识别率。所提出的方法由基于听觉掩蔽的特征增强模型和区分增强后语音和目标语音特征的判别器构成。特征增强模型的首要目标是以含噪语音的声学特征为输入,来对目标语音相应的理想比率掩膜(Ideal Ratio Mask,IRM)进行预测,而后利用听觉掩蔽效应得到增强后的声学特征。增强模型的次要目标则是通过欺骗判别器,使增强后的声学特征尽可能地满足目标语音的概率分布,从而减少与目标语音特征之间的差异,进而提高增强后语音特征的识别率。

1 基于听觉掩蔽生成对抗网络的增强方法

首先给出基于听觉掩蔽的有监督语音增强方法,而后对所提出的听觉掩蔽生成对抗网络(Generative adversarial network based on auditory masking, GANAM)进行介绍。GANAM主要由2部分构成,分别是:基于听觉掩蔽的特征增强模型E,以及用于区分增强后特征和目标语音特征的判别器D。图1给出了所提出方法的结构示意图。

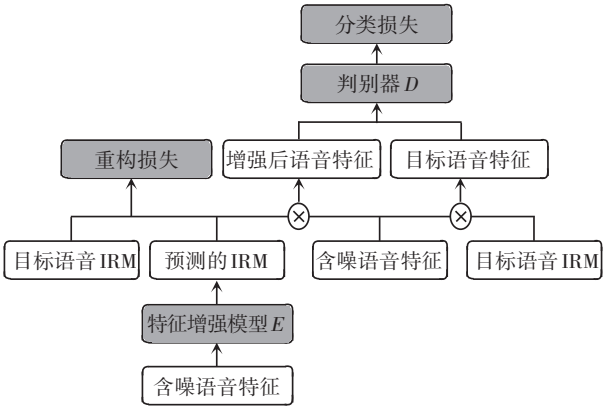

图1 基于听觉掩蔽生成对抗网络增强方法的示意图

Fig. 1 The schematic diagram of GANAM-based enhancement method

1.1 基于听觉掩蔽的有监督增强方法

基于听觉掩蔽的有监督增强方法能够有效提升含噪语音的可懂度和感知质量[9],同时还能够通过与声学模型的联合训练提高ASR系统的识别率[3]。因此,本文也采用基于听觉掩蔽的增强模型。理想比率掩膜IRM[6]是基于听觉掩蔽的增强模型最常用的学习目标之一,其定义如下:

$$M(t,f)=\left[\frac{S(t,f)}{Y(t,f)}\right]_0^1, \tag{1}$$

其中,M表示理想比率掩膜IRM;S和Y分别表示目标语音和含噪语音的时频特征;t和f分别表示时间帧序号和频带编号;$[\cdot]_0^1$表示对数值进行截取,将其限制在0~1之间。

理想比率掩膜可以定义在不同的时频特征上,包括能量谱FFT、对数能量谱log-FFT、梅尔能量谱fbank,以及对数梅尔能量谱log-fbank,其提取过程如下:

(1)对给定的语音波形进行短时傅里叶变换,得到复数谱。

(2)对复数谱的每个时频单元取模,再平方,得到能量谱FFT。

(3)对能量谱的每个时频单元进行对数压缩,即可得到对数能量谱log-FFT。

(4)使用梅尔滤波器组对能量谱进行滤波,得到梅尔能量谱fbank。

(5)对梅尔能量谱的每个时频单元取对数,得到对数梅尔能量谱log-fbank。

前期的实验表明,使用对数梅尔能量谱log-fbank能够获得较好的增强性能,因此本文后续的实验都以log-fbank作为声学特征。

当特征域确定后,即可训练增强模型E以含噪语音的声学特征Y为输入,来对目标语音相应的比率掩膜进行估计,即:

$$\hat{M}=E(Y), \tag{2}$$

在训练时,将理想比率掩膜真实值和预测值之间的均方误差(Mean Square Error, MSE)作为损失函数,通过最小化该损失函数,来对增强模型E的参数θ进行求解:

$$\theta=\arg\min_{\theta} L_{MSE}(M,\hat{M};\theta)=\arg\min_{\theta}\frac{1}{T}\frac{1}{F}\sum_{t=1}^{T}\sum_{f=1}^{F}(M(t,f)-\hat{M}(t,f))^2, \tag{3}$$

其中,T和F分别表示语音帧总数和频带总数。

在测试时,只需将预测出的比率掩膜$\hat{M}$与含噪语音声学特征Y中对应的时频单元相乘,便可得到

增强后的语音特征 $\hat{S}$：

$$\hat{S}(t,f)=\hat{M}(t,f)\cdot Y(t,f). \tag{4}$$

1.2　听觉掩蔽生成对抗网络

听觉掩蔽生成对抗网络 GANAM 在有监督语音增强模型的基础上，另外添加了区分声学特征纯净与否的判别器，从而能够借助其与增强模型形成的对抗过程，使增强后的声学特征更具判别性。

由于目标语音的数值范围较大，直接将其作为正例会增加生成对抗网络的训练难度。因此，在 GANAM 中，判别器 D 将含噪语音特征 Y 与理想比率掩膜 M 的乘积作为正例 $\bar{S}$：

$$\bar{S}(t,f)=Y(t,f)\cdot M(t,f), \tag{5}$$

而将比率掩膜预测值 $\hat{M}$ 增强后的语音特征作为反例 $\hat{S}$：

$$L_D=-\mathsf{E}_{\bar{S}\sim P_{\bar{S}}}[\log D(\bar{S})]-\mathsf{E}_{\hat{S}\sim P_{\hat{S}}}[1-\log D(\hat{S})], \tag{6}$$

在判别器尽可能区分增强后语音和目标语音特征的同时，增强模型则试图通过改变其预测的比率掩膜，来欺骗判别器 D，以获得较高的分数：

$$L_{adv}=\mathsf{E}_{\hat{S}\sim P_{\hat{S}}}[1-\log D(Y\odot E(Y))], \tag{7}$$

其中，$\odot$ 表示对应元素相乘。

单纯以最小化对抗损失 L_{adv} 作为增强模型 E 的训练目标，会使预测出的比率掩膜缺少约束，因为其只需使增强后的声学特征满足目标语音的概率分布即可。这样一来，增强后的语音和目标语音特征之间就会缺乏对应关系。因此，本文将比率掩膜的均方误差与对抗损失相结合，提出对抗多任务损失函数：

$$L_E=L_{MSE}+\lambda L_{adv}=\frac{1}{T}\frac{1}{F}\|M-\hat{M}\|_2^2+\lambda\,\mathsf{E}_{\hat{S}\sim P_{\hat{S}}}[1-\log D(\hat{S})]. \tag{8}$$

其中，$\|\cdot\|_2^2$ 表示 L_2 范数的平方，λ 是用来平衡重构损失和对抗损失的超参数，一般通过实验事先选取，本文取 0.000 1。

1.3　听觉掩蔽对抗训练有效性分析

本节对听觉掩蔽对抗训练的损失函数 L_D 进行分析。为了符号的简明性，记目标语音特征的分布为 P_s，增强后语音特征的分布为 P_e。当增强模型固定时，样本 x 在判别器 D 上的损失为：

$$L_D=-P_s(x)\log D(x)-P_e(x)\log(1-D(x)), \tag{9}$$

令其关于 D 的导数为 0，即可得到最优判别器 D^* 的分类面：

$$D^*(x)=\frac{P_s(x)}{P_s(x)+P_e(x)}, \tag{10}$$

事实上，该分类面也是最小贝叶斯分类误差的分类面。在训练时，通常会以最小化损失 L_D 为目标多训练几次判别器 D，再对增强模型进行更新。因此，增强模型最大化对抗损失 L_D 也是在最大化该最优分类器 D^* 的损失：

$$\begin{aligned}-L_D&=P_s(x)\log D^*(x)+P_e(x)\log(1-D^*(x))=\\&P_s(x)\log\frac{P_s(x)}{P_s(x)+P_e(x)}+P_e(x)\log\left(\frac{P_e(x)}{P_s(x)+P_e(x)}\right),\end{aligned} \tag{11}$$

将式(11)进行简单变换，可得：

$$\begin{aligned}-L_D&=P_s(x)\log\frac{P_s(x)}{\frac{1}{2}[P_s(x)+P_e(x)]}+\\&P_e(x)\log\frac{P_e(x)}{\frac{1}{2}[P_s(x)+P_e(x)]}-2\log 2=\\&\mathsf{E}_{P_s}\log\frac{P_s(x)}{\frac{1}{2}[P_s(x)+P_e(x)]}+\\&\mathsf{E}_{P_e}\log\frac{P_e(x)}{\frac{1}{2}[P_s(x)+P_e(x)]}-2\log 2=\\&2JS(P_s\parallel P_e)-2\log 2.\end{aligned} \tag{12}$$

其中，$JS(\cdot\parallel\cdot)$ 表示 2 个分布之间的 Jensen-Shannon 散度(JS 散度)。

由此可见，对抗训练最大化判别器的损失，实际上是在最小化目标语音和增强后语音特征分布之间的 JS 散度。因此，对抗训练能够使增强后的语音特征尽可能地满足目标语音的概率分布，从而有效提高 ASR 系统的识别率。

2　实验设置与评价指标

2.1　数据集

本文使用第 3 届语音分离与识别公开挑战赛 CHiME-3[10] 所提供的语音数据，来对所提出的方法进行评价。由于该数据集包括多个通道的语音数据，因此这里仅采用第 5 个通道的数据来进行单通道语音增强和识别任务的训练和测试。CHiME-3

数据集由模拟合成和真实录制的 2 部分数据构成。对于模拟合成的数据,其理想比率掩膜使用含噪语音和参与合成的目标语音计算得到;对于真实录制的数据,则使用录制到的远讲语音和近讲语音计算得到理想比率掩膜。为方便处理,所有的语音文件都采样到 16k Hz。另外,为模拟无混响的背景噪声干扰,还将纯净语音和噪声按照 0 dB、3 dB 和 6 dB 等信噪比进行混合,以扩充增强模型的训练集。

2.2 评价指标

通过计算增强后语音在鲁棒语音识别系统上的词错误率(Word Error Rate,WER),来评价增强模型的性能。一般而言,词错误率越低表示增强模型的性能越好,反之,词错误率越高则表示增强模型的性能越差。

本文使用 CHiME-3 挑战赛中官方提供的鲁棒语音识别系统对增强模型进行评价。该系统由深度神经网络(Deep Neural Network,DNN)和隐马尔科夫模型(Hidden Markov Model,HMM)构成。对其声学模型而言,首先训练高斯混合模型(Gaussian Mixture Model,GMM)和 HMM 构成的混合系统,来进行音素和语音帧之间的强制对齐,这里采用经过决策树聚类的三音素作为识别的基本单元。而后使用每帧的对数梅尔能量谱和三音素类别标签训练深度神经网络 DNN。为获得良好的初始化参数,先使用受限玻尔兹曼机对神经网络进行逐层初始化,而后再进行输入特征和标签对应的有监督分类训练。为使该声学模型尽可能地鲁棒,训练集含有多种声学场景下的语音数据,包括纯净语音、近讲语音、模拟的含噪语音,以及真实录制的含噪语音。通过这种多条件的训练方式,声学模型的噪声鲁棒性能够获得极大提升[11]。

语音识别系统的语言模型为华尔街日报(Wall Street Journal,WSJ)5000 词的 trigram 模型,这里使用 Kaldi 工具集中的加权有限状态机(Weighted Finite-State Transducer,WFST)对其进行建模。在对增强模型进行评价时,ASR 系统的声学模型和语言模型将固定不变,仅改变前端增强模型。

2.3 模型结构

所提出的听觉掩蔽生成对抗网络 GANAM 是一种学习范式,对增强模型的具体结构并没有特殊要求。因此这里采用单通道语音增强算法中常用的双向循环神经网络(Recurrent Neural Network,RNN)。为避免长时建模可能产生的梯度消失问题,该 RNN 网络采用长短时记忆单元(Long Short-term Memory Unit,LSTM)作为隐层单元。增强模型共包含 4 个隐层,而每个隐层则由 512 个 LSTM 单元构成。GANAM 中的判别器与声学模型的网络结构类似,其输入为前后各扩展 12 帧、共 25 帧声学特征拼接而成的向量,而其输出则是经过 sigmoid 函数归一化后的概率得分。判别器 D 共包含 3 个全连接层,每层由 1 024 个线性整流(Rectified Linear Unit,ReLU)神经元构成。

2.4 对比方法

为客观评价所提出方法的性能,本文将其与最近提出的 2 种基于生成对抗网络的增强方法进行比较,可得到如下研究结论:

(1)MappingGAN 是文献[7]提出的增强方法。与本文基于听觉掩蔽的增强方法不同,其增强模型以含噪语音的声学特征为输入,直接预测目标语音的特征;其判别器则尽可能地发现增强后语音和目标语音之间的差异。该方法可以有效提高混响干扰下鲁棒 ASR 系统的识别率,但对于宽带非平稳噪声的干扰还未进行评估。

(2)PairGAN 与本文的方法类似,也是基于听觉掩蔽的增强方法[6]。不同的是,PairGAN 将含噪语音与比率掩膜构成的二元组作为正例或反例,而不是将增强后的语音或目标语音特征作为正例或反例。该方法能够有效提高增强后语音的可懂度和感知质量,但对鲁棒 ASR 系统识别率的影响还有待研究。

3 实验结果及分析

3.1 性能对比

表 1 给出了 MappingGAN、PairGAN 以及所提出的 GANAM 方法增强后语音的词错误率。从表 1 中可以看出:

(1)不管是验证集(dt)还是测试集(et),基于听觉掩蔽的增强方法都能够进一步降低模拟合成语音(simu)和真实录制(real)语音在鲁棒 ASR 系统上的词错误率。这说明,前端增强方法是提升语音识别系统噪声鲁棒性的可行途径。

(2)与有监督方法相比,所提出的 GANAM 在所有评测条件下都能够显著降低增强后语音的词错误率,从而说明,GANAM 能够使增强后的语音特征更具判别性。

(3)与现有的增强方法 MappingGAN 和 PairGAN 相比,GANAM 增强后的语音特征获得了更低的词错误率。这就表明,相比于其他的对抗训练

策略，基于听觉掩蔽的生成对抗网络能够更加有效地提取和利用目标语音声学特征的概率分布。

（4）通过对比 PairGAN 和有监督方法可以看出，不恰当的对抗策略非但不能提高 ASR 系统的识别率，甚至还会降低 ASR 系统的识别性能。此外，PairGAN 的实验结果还表明，提高增强后语音的可懂度、感知质量等主观指标，和提高语音识别率这一客观指标是 2 个不同的问题，能够提高主观指标的增强方法并不一定能够提高识别率。

表 1　各模型在 CHiME-3 上的词错误率

Tab. 1　The *WER* of different models on CHiME-3　%

模型	dt_simu	dt_real	et_simu	et_real
ASR	12.68	14.19	15.14	25.44
有监督	12.08	12.70	13.44	22.26
MappingGAN	12.29	12.00	13.74	21.70
PairGAN	17.78	16.96	20.75	29.64
GANAM	**11.61**	**12.34**	**12.99**	**20.48**

3.2　判别器模型结构对增强后语音词错误率的影响

对于生成对抗网络而言，判别器的模型结构也会对最终的性能产生影响。为了评估该影响，本节固定增强模型的网络结构不变，分别使用参数量相同的卷积神经网络和循环神经网络，代替所采用的深度神经网络判别器。表 2 给出了不同网络结构的判别器对增强后语音识别率的影响。从表 2 中可以看出，判别器的模型结构确实会对增强后语音的识别率产生显著的影响。与有监督方法（不含判别器的对抗训练）相比，基于卷积神经网络 CNN 和循环神经网络 LSTM 的判别器并不能提升增强后语音的识别率，而基于 DNN 的判别器则在真实录制的测试集 et_real 上带来了 1.78%的词错误率下降。

表 2　判别器网络结构对增强后语音词错误率的影响

Tab. 2　The impact of discriminator network architecture on the *WER* of enhanced speeches　%

模型	dt_simu	dt_real	et_simu	et_real
ASR	12.68	14.19	15.14	25.44
有监督	12.08	12.70	13.44	22.26
CNN	12.08	12.90	13.26	22.29
LSTM	12.01	13.14	13.56	22.19
DNN	**11.61**	**12.34**	**12.99**	**20.48**

3.3　听觉掩蔽生成对抗网络对增强后特征的影响

为探究听觉掩蔽生成对抗网络 GANAM 是如何影响增强后的语音特征，使其识别率得以提升，本节将不同模型增强后的声学特征进行可视化，如图 2 所示。可以看出，有监督增强方法只是在尽可能地最小化增强后语音和目标语音特征之间的差异，而并不关心增强后语音是否满足纯净语音的概率分布，从而使真实含噪语音增强后的特征依然可能含有较多的噪声干扰。而所提出的 GANAM 增强方法则通过对抗训练的方式，来对纯净语音的概率分布进行建模，从而尽可能地去除增强后语音特征中的噪声干扰，得到更为干净的语音特征，进而提升增强后语音的识别率。

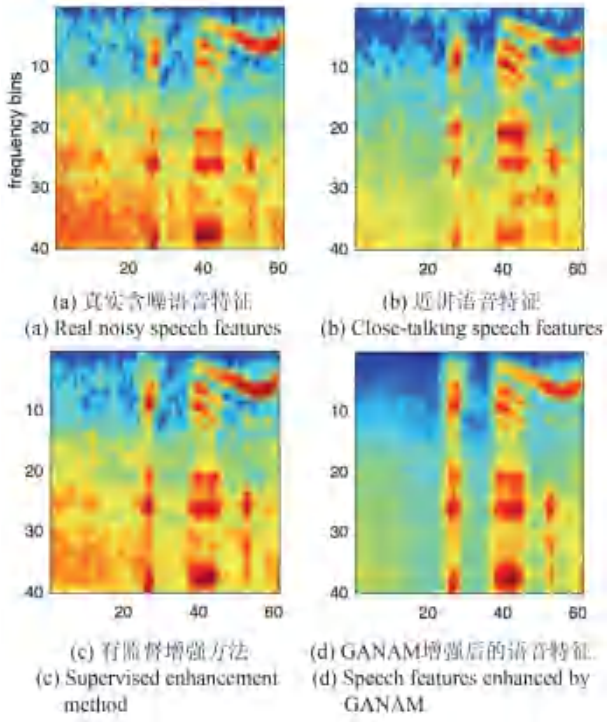

图 2　不同方法增强后的声学特征对比

Fig. 2　The comparison of acoustic features enhanced by different models

4　结束语

本文在基于声学掩蔽有监督增强方法的基础上，通过引入额外的判别器，来对纯净语音的概率分布进行建模，而后利用学习到的概率分布指导增强模型的训练，从而使其增强后的语音特征满足目标语音的概率分布，同时也更具判别性。实验结果表明，将所提出的声学掩蔽生成对抗网络增强方法 GANAM 作为语音识别系统的前端处理，能够降低鲁棒 ASR 系统在真实含噪语音上的词错误率，进一步提高其噪声鲁棒性。

参考文献

[1] HAN Kun, HE Yanzhang, BAGCHI D, et al.Deep neural network based spectral feature mapping for robust speech recognition[C]// 16th Annual Conference of the International Speech Communication

Association (Interspeech). Dresden, Germany: isca - speech organization,2015:2484-2488.
[2] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[M]//VINCENT E, YEREDOR A, KOLDOVSKÝ Z, et al. Latent variable analysis and signal separation. LVA/ICA 2015. Lecture Notes in Computer Science. Cham:Springer, 2015,9237:91-99.
[3] WANG Zhongqiu, WANG Deliang. A joint training framework for robust automatic speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4):796-806.
[4] LIU Bin, NIE Shuai, ZHANG Yaping, et al. Boosting noise robustness of acoustic model via deep adversarial training[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).Calgary, AB, Canada:IEEE, 2018: 3674-3679.
[5] GOODFELLOW I, POUGET - ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// NIPS. Montreal, QC, Canada: NIPS Foundation, 2014:2672-2680.
[6] PANDEY A, WANG Deliang. On adversarial training and loss functions for speech enhancement[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).Calgary, AB, Canada:IEEE, 2018:5414-5418.
[7] WANG Ke, ZHANG Junbo, SUN Sining, et al. Investigating generative adversarial networks based speech dereverberation for robust speech recognition[C]//Interspeech. Hyderabad, India: dblp, 2018:1581-1585.
[8] DONAHUE C, LI Bo, PRABHAVALKAR R. Exploring speech enhancement with generative adversarial networks for robust speech recognition[C]// ICASSP. Calgary, AB, Canada:IEEE, 2018: 5024-5028.
[9] WANG Yuxuan, NARAYANAN A, WANG Deliang. On training targets for supervised speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12):1849-1858.
[10] BARKER J, MARXER R, VINCENT E, et al. The third 'chime' speech separation and recognition challenge: Dataset, task and baselines[C]// 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. Scottsdale, AZ, USA: IEEE, 2015:504-511.
[11] LI Feipeng, NIDADAVOLU P, HERMANSKY H. A long, deep and wide artificial neural net for robust speech recognition in unknown noise[C]// Interspeech. Singapore:dblp, 2014:1-6.

(上接第 208 页)

[5] LI Yanwei, CHEN Xinze, ZHU Zheng, et al. Attention-guided unified network for panoptic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 7026-7035.
[6] JIAO Jianbo, WEI Yunchao, JIE Zequn, et al. Geometry aware distillation for indoor semantic segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, CA, USA:IEEE, 2019:2869-2878.
[7] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision - ECCV 2018. Lecture Notes in Computer Science. Cham: Springer,2018,11211: 833-851.
[8] CHOLLET F . Xception: Deep Learning with depthwise separable convolutions[C]// 2017 IEEE Conference On Computer Vision And Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE, 2017:1800-1807.
[9] HE Kaiming , ZHANG Xiangyu , REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-1916.
[10] CAMPOS C, ELVIRA R, RODRÍGUEZ J J G, et al. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM[J]. arXiv preprint arXiv:2007.11898, 2020.