基于多域融合及神经架构搜索的语音增强方法

张睿，张鹏云，孙超利
（太原科技大学计算机科学与技术学院，山西 太原 030024）

摘 要：为进一步提高语音增强模型的自学习及降噪能力，提出基于多域融合及神经架构搜索的语音增强方法。该方法设计了语音信号多空间域映射及融合机制，实现信号实复数关联关系的挖掘；围绕模型卷积池化运算特点，提出了复数神经架构搜索机制，通过设计的搜索空间、搜索策略及评估策略，高效自动地构建出语音增强模型。实验搜索到的最优语音增强模型与基线模型的对比泛化实验中，语音质量客观评价（PESQ）、短时客观可懂度（STOI）两大指标较最优基线模型均最大提升 5.6%，且模型参数量最低。

关键词：语音增强模型；复数空间域映射；多域融合；复数神经架构搜索；低成本评估
中图分类号：TP18
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024018

Speech enhancement method based on multi-domain fusion and neural architecture search

ZHANG Rui, ZHANG Pengyun, SUN Chaoli
College of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Abstract: In order to further improve the self-learning and noise reduction ability of speech enhancement model, a speech enhancement method based on multi-domain fusion and neural architecture search was proposed. The multi-spatial domain mapping and fusion mechanism of speech signals were designed to realize the mining of real complex number correlation. Based on the characteristics of convolution pooling of the model, a complex neural architecture search mechanism was proposed, and the speech enhancement model was constructed efficiently and automatically through the designed search space, search strategy and evaluation strategy. In the comparison and generalization experiment between the optimal speech enhancement model and the baseline model, the two indexes of PESQ and STOI increase by 5.6% compared with the optimal baseline model, and the number of model parameters is the lowest.

Keywords: speech enhancement model, complex spatial domain mapping, multi-domain fusion, complex neural architecture search, low-cost evaluation

---

收稿日期：2023−10−11；**修回日期：**2023−12−19
通信作者：张鹏云，zpyyyny@126.com
基金项目：国家自然科学基金资助项目（No.62372319）；教育部人文社会科学研究基金资助项目（No.23YJCZH299）；山西省重点研发计划基金资助项目（No.202102020101002）；山西省基础研究计划基金资助项目（No.20210302123216）；太原科技大学研究生联合培养示范基地基金资助项目（No.JD2022004）；太原科技大学研究生教育创新基金资助项目（No.SY2023040）
Foundation Items: The National Natural Science Foundation of China (No.62372319), Humanities and Social Science Research Project of Ministry of Education (No.23YJCZH299), The Key Research and Development Project of Shanxi Province (No.202102020101002), Basic Research Project of Shanxi Province (No.20210302123216), Project of Graduate Joint Training Demonstration Base of Taiyuan University of Science and Technology (No.JD2022004), Graduate Education Innovation Project of Taiyuan University of Science and Technology (No.SY2023040)

0 引言

语音通信等任务[1]广泛应用于日常生活中，但语音中的噪声会覆盖语音中的关键信息，降低语音感知质量和可理解性，使语音质量难以保证，影响用户对语音内容的理解，进而降低用户通信体验，对日常生活造成严重影响。为有效消除噪声干扰，语音增强技术被提出，它通过对带噪语音信号进行处理，尽可能移除噪声信号，恢复出干净语音，大幅减少了语音通信中的噪声干扰、提高了语音信噪比，使语音通信质量更好。因此，语音增强对语音的相关任务至关重要。

近年来，深度学习成为语音领域的研究热点[2]，基于深度学习的语音增强方法得到快速发展。早期研究的语音增强方法主要是基于时频域的分析方法，更关注时频域中与振幅相关的训练目标而忽略了相位，即关注实部忽略虚部，单纯地利用语音信号实部特征进行语音增强，导致语音信号严重偏离正常值，造成降噪效果差等问题。最近一些研究开始重视虚部信息，在实值网络中进行训练，分别预测实部和虚部后进行叠加。微软提出了 Phasen 语音增强模型[3]，使用双流模型分别处理实部和虚部信息，各项指标都取得了不错的效果。Tan 等[4]的卷积递归网络（CRN, convolutional recurrent network）集成了卷积编解码器结构和长短期记忆（LSTM, long short-term memory）网络，已被证明有利于处理复杂的目标。这些方法虽利用了实部与虚部的信息，但实部与虚部也需分离计算且不受复乘规则的限制，没有充分利用实部与虚部的内在关联性。为将实部和虚部共同计算以充分利用其关联性，Choi 等[5]根据 U-net 提出 DCUNet 网络[6]，在 U-net 基础上设计了复数批归一化和复数 ReLU 块来实现该思想，复数模块通过复数乘法来模拟实部和虚部之间的相关性，输入的复数数据可以直接进行运算，不需要将实部虚部分开估计。Hu 等[7]设计了 DCCRN，借鉴 DCUNet 的复数思想并对 CRN 进行大量修改得到复数 Conv2d 层，并提出复数 LSTM 来代替传统的 LSTM，进一步更新了 CRN。此类基于复数的方法充分利用了实部和虚部，很大程度地保留了有效的语音特征，提高了语音增强效果，但这些方法仍然只是基于语音信号的时频域进行分析，也同样缺乏高效的轻量化多域复数特征融合模块，除此之外，这些深度模型所取得的卓越性能大部分是因为研究人员精心设计了较深的模型体系结构和层结构，这给模型自动化带来了巨大的挑战。

与人工设计的复杂网络相比，探索灵活机动的模型体系结构更符合当前技术发展的需要。因此，近年来出现了大量的神经架构搜索（NAS, neural architecture search）方法。早期专家尝试使用循环神经网络作为控制器并使用强化学习控制其参数来搜索−评估−更新模型，该方法需要超过 60 年的 GPU 计算日，因为其需要将所有搜索到的模型进行完全训练评估。后续，在具有 13 个操作的搜索空间中专注搜索正常 Cell 和缩减 Cell 这 2 种细胞来解决评估时间长的问题[8]，该方法将评估时间减少到 2 000 个 GPU 计算日。Baker 等[9]使用了与之相同的方法设计了 MetaQNN，但仍需 96 个 GPU 运算日，消耗的计算资源仍然很大且无法扩展该方法的使用范围。为降低评估时间，Beeche 等[10]提出高效神经架构搜索（ENAS），使用经过策略梯度训练的控制器，在大型计算图中搜索最佳子图来发现最优体系结构，ENAS 强制所有生成的体系结构共享参数来减少评估时间，在不到一个 GPU 计算日内完成有效评估，但造成性能下降等问题。Liu 等[11]提出渐近式神经架构搜索（PNAS），使用一个 LSTM 做代理模型来指导模型结构搜索，输入模型结构的变长字符串描述，输出预测的验证精度，但与 ENAS 具有相同问题。上述提到的各类方法本质上均是在离散空间中搜索及评估，它们将目标函数看作黑盒，但从已有研究可知，若搜索空间连续且目标函数可微，那么基于梯度信息的搜索评估策略将更加快速，因此 Huang 等[12]提出基于梯度的可微分架构搜索（DARTS），将搜索空间转换为连续空间，目标函数看作可微函数，使用基于梯度的优化方法搜索评估最优模型；与之类似的还有 Luo 等[13]提出的另一种基于梯度的方法，先将模型嵌入连续空间，该空间中每个点对应一个模型，在该空间上可定义准确率的预测函数，以目标函数进行基于梯度的优化，找到更优模型的嵌入表征，优化完成后再将这个嵌入表征映射回模型。

这类基于梯度的方法的优点之一就是搜索评估效率高，结合一些如权重共享的加速手段，消耗可少于一个 GPU 计算日，但对语音增强应用来说评估时间仍然很长。虽然 Mellor 等[14]提出了低成本评估策略，在不完全训练模型的情况下根据模型初始特征对模型性能进行评分，Lopes 等[15]也提出了

类似的策略，使用局部线性算子的雅可比矩阵来评估不同模型的性能，但上述方法的评分只是粗略评估，且该方向研究还相对较少，需进一步深入探索。综上所述，目前语音增强方法仍存在一些挑战。

1) 特征提取上的挑战。目前鲜有 NAS 专家为语音增强设计专用的复数搜索空间，搜索空间中也同样缺乏高性能复数特征提取模块，造成模型计算时间过长、效率过低。除此之外，现有语音增强方法更侧重于时频域中语音信号的振幅和相位信息，忽视了语音信号其他空间域的信息表达，导致收集的数据样本单一，特征间相关性表达不充分，将会限制后续语音增强模型效果，进而阻碍语音增强技术的进一步发展。

2) 搜索性能上的挑战。目前 NAS 中大多数搜索策略对搜索目标函数的设计具有较高要求，常用的基于梯度下降的搜索策略需对目标函数进行复杂的设计，常常面临内存短缺且搜索性能较低的问题。主流模型性能评估策略需对候选模型进行完全训练，计算成本高，而低成本评估策略也存在评估性能弱等问题，在算力薄弱或时间不足时将难以保证语音增强效果。

围绕上述 2 个问题，本文提出了一种基于多域融合及神经架构搜索的语音增强方法，它同时考虑了高质量语音增强模型的特征增强及模型自主构建学习能力。本文主要贡献如下。

1) 多域映射及轻量化复数特征融合机制。为提高语音信号的特征丰富度，将一维语音信号映射至多个空间域中生成基础域和辅助域，并设计了复乘规则简单的轻量化复数特征融合机制，对基础域和辅助域进行融合，增强一维信号特征表达，得到更丰富的语音特征。

2) 复数神经架构搜索机制。为降低语音增强模型设计的人为影响，提高模型构建的自适应性，根据模型编解码部分的卷积池化特性，提出了一种面向语音增强的轻量化基于联合 Cell 的可分离复数搜索空间及编码策略，并基于上述可分离搜索空间设计了一种高性能自适应全局/局部协同特征感知的搜索策略。最后为进一步加快对训练成本较高的语音增强模型的性能评估的速度，降低计算开销，提出了一种低成本模型性能评估策略。

1　方法整体框架

本文方法框架如图 1 所示，主要包含 2 个部分，分别为多域映射及轻量化复数特征融合机制和复数神经架构搜索机制。

多域映射及轻量化复数特征融合机制将一维语音信号映射至多个空间域中生成基础域和辅助域，以提高语音信号的特征丰富度，并在此基础上为模型设计了轻量化复数特征融合机制将多域信息进行高性能融合。

复数神经架构搜索机制包含：1) 专门为语音增

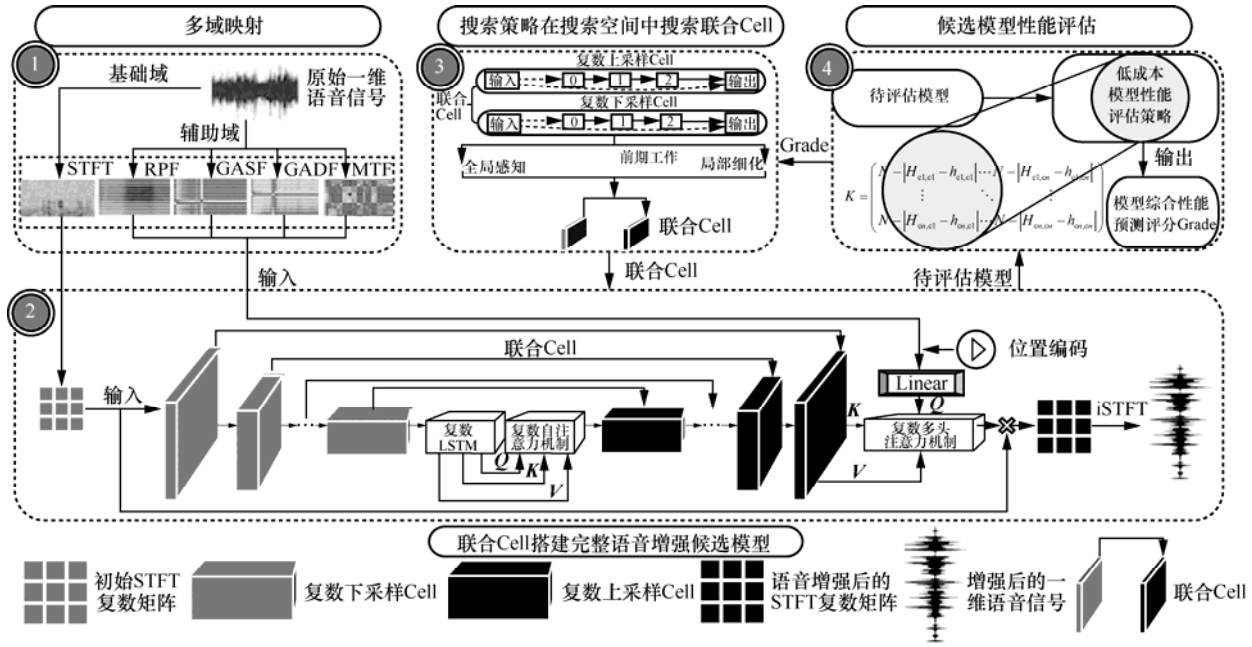


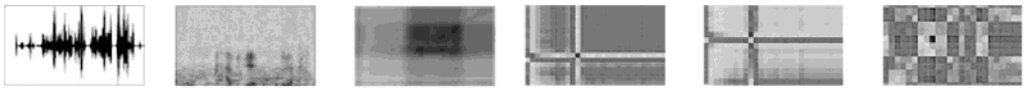
图 1　本文方法框架

强设计的轻量化基于联合Cell的可分离复数搜索空间，该可分离空间考虑到卷积的全局感知与池化的局部细化特点，将联合 Cell 分为复数卷积子空间及复数池化子空间，与现有的基于 Cell 的搜索空间相比搜索更加灵活，并为这种卷积池化分离搜索的搜索空间设计了一种新的编码策略；2) 一种高性能自适应全局/局部协同特征感知的搜索策略，可在上述搜索空间中根据种群状态权衡开发和勘探，以进行全局感知或局部细化搜索最优联合 Cell，并以此搭建最优语音增强模型；3) 一种低成本模型性能评估策略，根据模型初始化特性对模型进行评分，以此近似预测模型的最终性能，进一步提高候选模型的评估效率。下文详细介绍了本文提出的基于多域融合及神经架构搜索的语音增强方法的细节。

2 方法详细介绍

2.1 多域映射及轻量化复数特征融合机制

通常语音增强模型只是利用单一时频域进行分析，存在提取特征单一、特征属性关联性不足等问题，因此本文依据前期研究[16]将语音一维信号样本映射入多个空间域中以得到更多的语音信息。数据样本高维映射共有基础域的短时傅里叶变换（STFT, short-time Fourier transform），辅助域的递归图域（RPF, recurrence plot field）、格拉姆角和场域（GASF, Gramain angular summation field）、格拉姆角差场域（GADF, Gramain angular difference field）和马尔可夫转移场域（MTF, Markov transition field）。多域映射图像如图 2 所示。



(a) 时域 (b) STFT (c) RPF (d) GASF (e) GADF (f) MTF

图 2 多域映射图像（基础域和辅助域）

将一维信号映射入多个空间域后，为实现多空间域的特征融合，本文在语音增强模型中设计了一种轻量化的复数特征融合机制——复数多头注意力机制，如图 3 所示，可将基础域（STFT）与表现较好、表征能力较强的辅助域相结合，从语音信号中提取更有效的语音特征，充分利用不同域的特点实现深层浅层不同类型特征的融合。

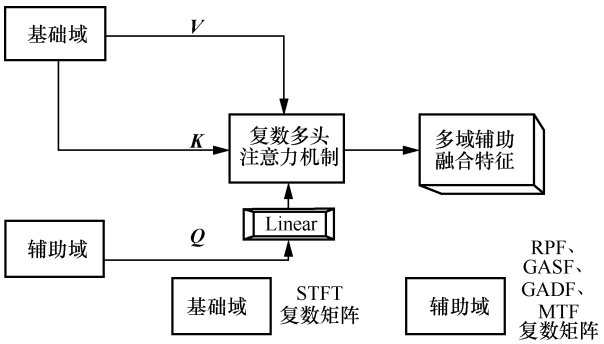


图 3 轻量化复数特征融合机制

图 4 及式(1)、式(2)为复数多头注意力机制具体计算方式。首先将 $\boldsymbol{Q}$ 、 $\boldsymbol{K}$ 运算得到的 $\boldsymbol{W}$ 的实部矩阵 $\boldsymbol{W}_{\mathrm{r}}$ 和虚部矩阵 $\boldsymbol{W}_{\mathrm{i}}$ 分别进行 Softmax 计算后转换为概率值，再将概率值分别与 $\boldsymbol{V}_{\mathrm{r}}$ 和 $\boldsymbol{V}_{\mathrm{i}}$ 对应相乘，概率值可以反映 $\boldsymbol{V}$ 值的重要程度，最后将得到的实部矩阵和虚部矩阵叠加得到输出 **Output**，并通过维度重构将 **Output** 的维度重构到与输入相同。图 4 中 2 条实线 $\boldsymbol{Q}_{\mathrm{i}}$ 与 $\boldsymbol{K}_{\mathrm{i}}^{\mathrm{T}}$ 和 $\boldsymbol{Q}_{\mathrm{r}}$ 与 $\boldsymbol{K}_{\mathrm{r}}^{\mathrm{T}}$ 代表式(1)中的 $\boldsymbol{Q}_{\mathrm{r}}\times\boldsymbol{K}_{\mathrm{r}}^{\mathrm{T}}-\boldsymbol{Q}_{\mathrm{i}}\times\boldsymbol{K}_{\mathrm{i}}^{\mathrm{T}}$ 操作，$\times$ 表示矩阵乘法。

$$\boldsymbol{W}=\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}}=\left(\boldsymbol{Q}_{\mathrm{r}}\times\boldsymbol{K}_{\mathrm{r}}^{\mathrm{T}}-\boldsymbol{Q}_{\mathrm{i}}\times\boldsymbol{K}_{\mathrm{i}}^{\mathrm{T}}\right)+\mathrm{j}\left(\boldsymbol{Q}_{\mathrm{r}}\times\boldsymbol{K}_{\mathrm{i}}^{\mathrm{T}}+\boldsymbol{Q}_{\mathrm{i}}\times\boldsymbol{K}_{\mathrm{r}}^{\mathrm{T}}\right) \tag{1}$$

$$\textbf{Output}=\mathrm{Softmax}(\boldsymbol{W})\times\boldsymbol{V} \tag{2}$$

2.2 复数神经架构搜索机制

2.2.1 基于联合Cell的可分离复数搜索空间及编码策略

目前语音增强模型中最重要的上/下采样编解码部分均为人工设计，为进一步降低人为影响，提高模型构建的自适应性及语音增强性能，也为了将 NAS 应用到语音增强领域，针对语音信号非线性及非平稳性所导致的特征提取困难等问题，本节为模型中的上/下采样编解码部分设计了一种基于联合 Cell 的可分离复数搜索空间，如图 5 所示。

具体来说，卷积和池化注重特征学习的不同方面，卷积侧重于特征图整体的深度特征学习，也就是全局感知，而池化更多地侧重于卷积层特征降维，而不进行深度特征提取，也就是局部细化。因此本文将候选操作划分为 2 个子空间，即复数卷积

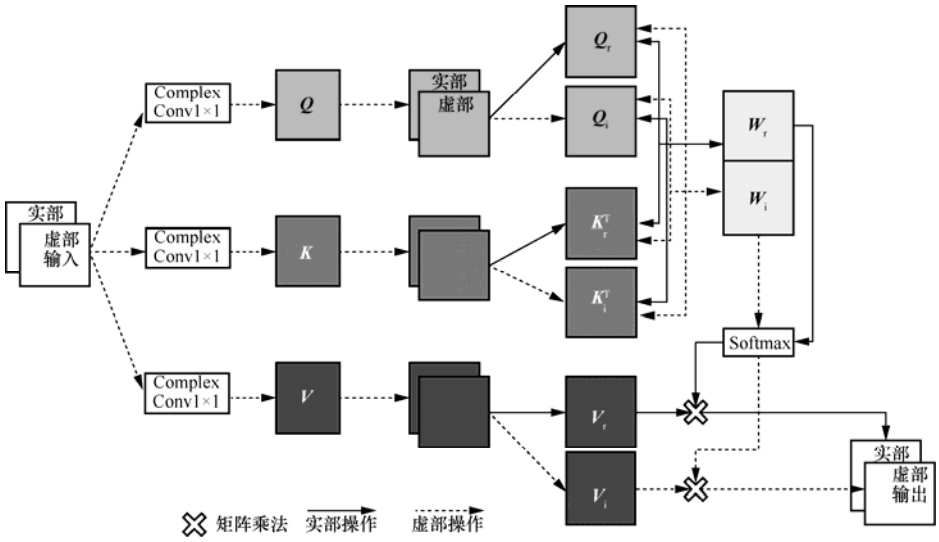


图 4　复数多头注意力机制具体计算方式

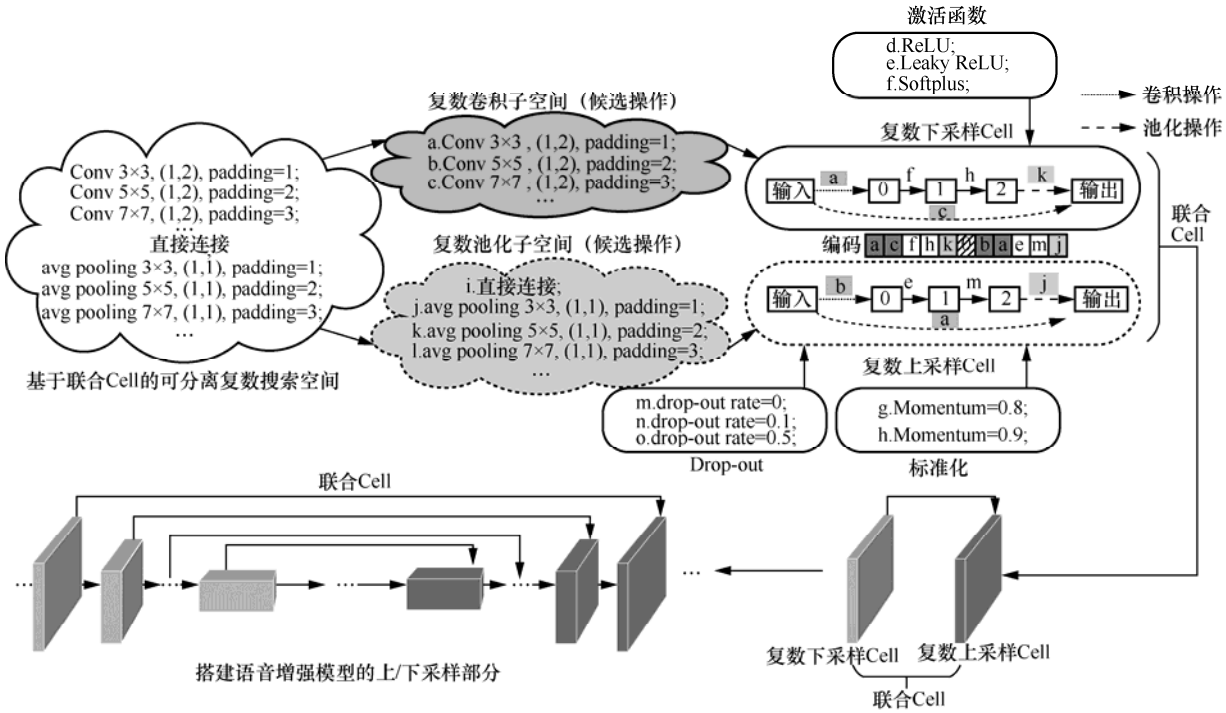


图 5　基于联合 Cell 的可分离复数搜索空间、联合 Cell 拓扑及编码方式

子空间和复数池化子空间，分别包含不同卷积核的复数卷积候选操作及不同滑动窗口的复数池化候选操作，之后将每个 Cell 中待搜索节点同样分离为卷积节点及池化节点，充分利用了卷积及池化的特性，降低搜索空间复杂度和体积，且该设计可自适应地对复杂性较强的非平稳语音信号进行深层特征提取或特征降维。由于搜索空间的合理划分，后续提出的搜索策略可以高效有效地进行。

为保证语音增强模型编解码时特征映射维度相同，以及语音信号在特征提取后的维度恢复，本文将一个复数上采样 Cell 与一个复数下采样 Cell 结合共同作为一个联合 Cell（如图 5 所示），后续将仅对一个联合 Cell 进行搜索，再将搜索到的最优联合 Cell 由外向内搭建数次，这种对称的体系结构确保了模型编解码部分的输出与输入具有相同维度及大小，最后结合 2.1 节轻量化复数特征融合机制即可构建完整的语音增强候选模型（如图 1 所示）。

在模型特征提取时，为了在频率方向上利用更多的上下文信息，所有搜索空间中复数卷积操作步长设为(时间维度,频率维度)=(1,2)。换句话说，该方法可在复数下采样 Cell 间逐层将特征图频率维度的大小减半，在复数上采样 Cell 间将其逐层加倍，而不改变特征图时间维度的大小。模型部分参数如表 1 所示，其中每个复数上采样 Cell 的输入通道数由于跳连接增加了一倍，每一层的输入大小和输出大小以通道数×时间维度 T×频率维度格式给出。

表 1　　模型部分参数

操作层	输入大小	输出大小
reshape_1	$T\times 257$	$1\times T\times 257$
复数下采样 Cell_1	$1\times T\times 257$	$16\times T\times 129$
复数下采样 Cell_2	$16\times T\times 129$	$32\times T\times 65$
复数下采样 Cell_3	$32\times T\times 65$	$64\times T\times 33$
复数下采样 Cell_4	$64\times T\times 33$	$128\times T\times 17$
复数下采样 Cell_5	$128\times T\times 17$	$256\times T\times 9$
reshape_2	$256\times T\times 9$	$T\times 2\,304$
复数 LSTM 及其他模块	$T\times 2\,304$	$T\times 2\,304$
reshape_3	$T\times 2\,304$	$256\times T\times 9$
复数上采样 Cell_5	$512\times T\times 9$	$128\times T\times 17$
复数上采样 Cell_4	$256\times T\times 17$	$64\times T\times 33$
复数上采样 Cell_3	$128\times T\times 33$	$32\times T\times 65$
复数上采样 Cell_2	$64\times T\times 65$	$16\times T\times 129$
复数上采样 Cell_1	$32\times T\times 129$	$1\times T\times 257$
reshape_4	$1\times T\times 257$	$T\times 257$

为进一步方便后续搜索策略的搜索，对联合 Cell 拓扑进行了编码，编解码方式与文献[17]中常用的“01”字符串编码仅表示神经节点之间的连接不同，本文在其基础上，提出了一种新的编码方式来指示联合 Cell 的信息流，考虑使用字符串和“0”字符来分别描述运算符信息和分离点，其中“0”字符代表复数下采样 Cell 与复数上采样 Cell 信息流的分离。改进的编码策略允许结构信息的完整表示，从而促进了搜索过程中操作符的组合。

2.2.2　高性能自适应全局/局部协同特征感知的搜索策略

基于进化搜索（ES, evolutionary search）[18]的方法因其较强的优化能力和易于实现的特点而越来越受到专家的关注。因此，为了进一步提高搜索策略在搜索空间的搜索效率，也为了更准确地寻找到高质量的语音增强体系结构，本节为 2.2.1 节可分离语音增强搜索空间配套设计了一种具有全局感知和局部细化特性的高性能自适应全局/局部协同特征感知的搜索策略，可根据种群状态对联合 Cell 中的复数卷积节点及复数池化节点分别进行全局感知及局部细化搜索，而不是让每个节点随机选择卷积或池化操作。与现有的统一搜索策略相比，所提策略由于卷积和池化空间的分离，可以更快速有效地发现性能良好的语音增强体系结构，并有效降低搜索成本。搜索策略流程如图 6 所示，搜索策略算法如算法 1 所示。

算法 1　搜索策略算法

定义　复数卷积操作 $O_{\text{complexcon}}$，复数池化操作 $O_{\text{complexpool}}$，其他操作 O_{other}，n 张相似图片的小批次数据 D_{val}

输出　最佳语音增强模型

1) $O\leftarrow$通过 $O_{\text{complexcon}}$、$O_{\text{complexpool}}$ 及 O_{other} 构建搜索空间；

2) $f\leftarrow 0$；

3) $p_f\leftarrow$基于 O 初始化 Pb 个个体；

4) 使用低成本模型性能评估策略及 D_{val} 评估 p_f 中个体的性能；

5) 循环

6) P_1、$P_2\leftarrow$使用轮盘赌从 p_f 中选择 2 个父代；

7) K_1、$K_2\leftarrow$对 P_1、P_2 进行交叉及变异；

8) $Q_f\leftarrow$分离 K_1、K_2 中的每个个体并通过全局感知搜索产生 M 个新的复数卷积字符串；

9) $Q_f\leftarrow$融合生成 M 个新个体并使用 D_{val} 对个体性能进行评估；

10) $Q_f\leftarrow$若 Q_f 中某一个体性能超过父代中的个体，则进行局部细化搜索更新 Q_f 中的所有个体并使用 D_{val} 对所有个体进行评估；

11) $p_{f+1}\leftarrow$通过精英策略在 $P_f\cup Q_f$ 中选择 Pb 个较优个体；

12) $f\leftarrow f+1$；

13) until 达到设定的最大运行次数。

14) 从 Q_f 中选择一个最好个体并以此搭建最优语音增强模型。

具体来说，该搜索策略首先通过轮盘赌随机选择方法[19]从种群中选择 2 个父解并编码，即 P_1 和 P_2。接下来 P_1 和 P_2 将进行交叉、变异生成子代 K_1 和 K_2。此外，对于新个体生成的解决方案，本文将子代 K_1 和 K_2 的编码分离为复数卷积字符串、复数

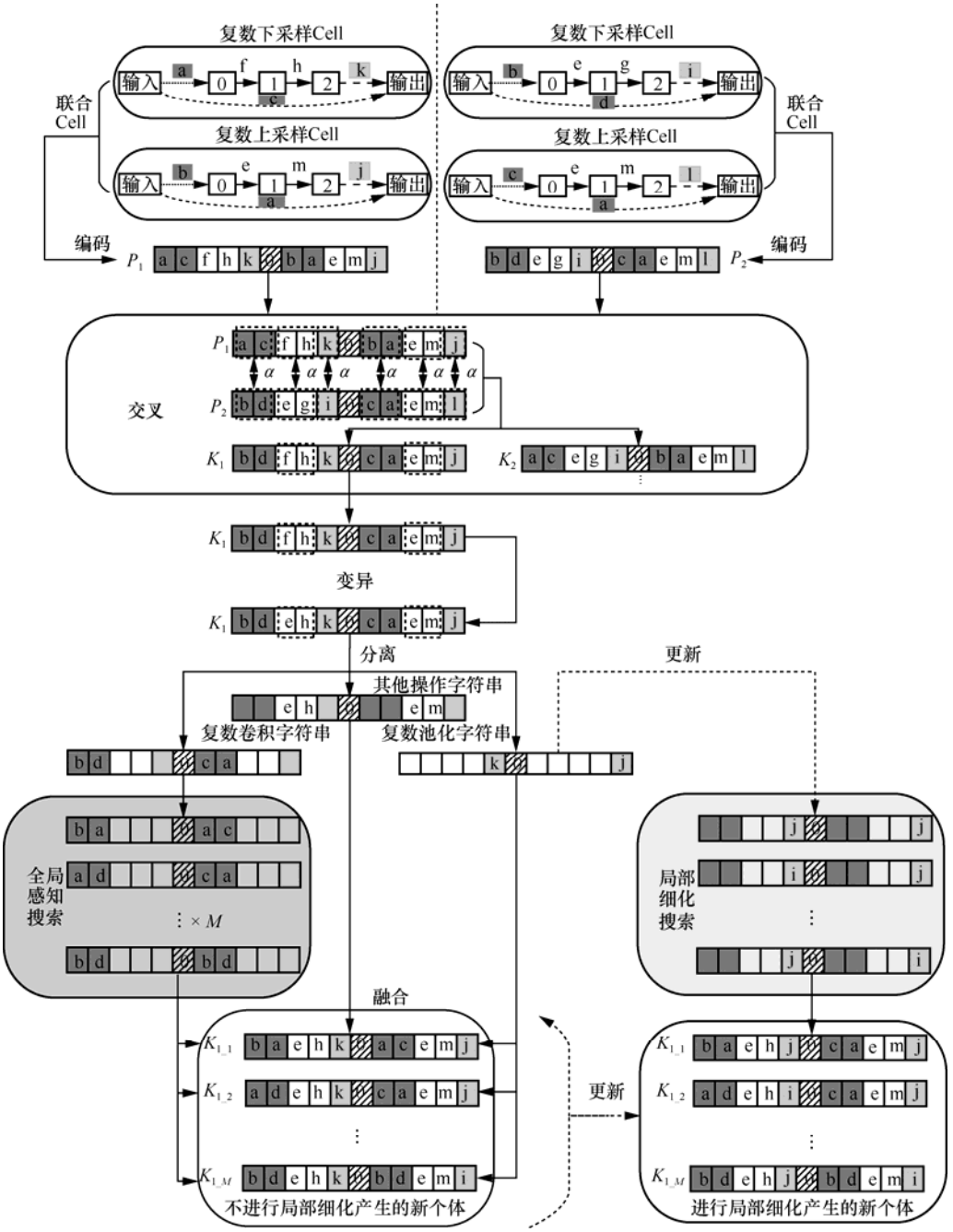


图 6　搜索策略流程

池化字符串及其他操作字符串。然后利用子代的复数卷积字符串，进行 M 次全局感知搜索，探索卷积空间，生成 M 个新的复数卷积字符串。全局感知搜索结束后，将上述 3 种字符串进行融合，判断生成新个体的语音增强性能，当新个体中存在某个个体的性能优于父代时，则证明当前种群所处搜索范围性能较优，因此会对新个体的池化空间进行局部细化搜索来更新所有新个体信息。通过局部细化扩大种群在较优搜索空间的搜索范围，不但有机会找到当前空间的峰值点，也可有效防止个体陷入局部最优。在局部细化发现的最佳个体将存活下来，并被用于取代种群中的父代个体。

搜索过程中，变异运算包含 2 种突变算子，如图 7 所示。在子代 K_1 和 K_2 的其他操作字符串部分，突变被触发时随机选择执行，以不同尺度进一步探索搜索空间。从图 7 中可以看出，第一个突变算子主要关注联合 Cell 中某一复数上或下采样 Cell 的变化，并保持另一 Cell 的操作不变；第二个突变算子

试图将一个 Cell 的信息流改变到另一个 Cell 相应的神经节点上。

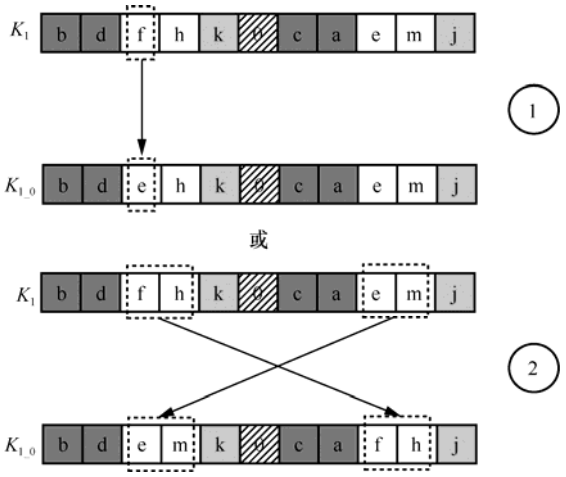


图 7　突变算子

在上述的 M 次全局感知搜索阶段中，考虑到这里的联合 Cell 的编码本质上是一个字符串编码，本文提出了一个多点搜索的变体来对子代 K_1 和 K_2 的复数卷积子空间进行全局感知搜索，如图 8 所示。对于某个个体，首先根据编码“0”字符来识别联合 Cell 中复数上/下采样 Cell 的字符串信息流的分离点。从图 8 中可以看出，算法已经识别出了一个分离点，字符串中的每个段都代表复数上采样 Cell 或复数下采样 Cell 中隐藏的神经节点。接下来，每个段将在预定义的概率下进行交换，以生成多个新个体的复数卷积字符串。

2.2.3　低成本模型性能评估策略

模型性能评估策略是指准确、高效地度量经过上述搜索策略搜索到的联合 Cell 所搭建模型的性能，尽管 ES 较梯度下降具有全局优化能力，但值得注意的是，ES 在其迭代搜索过程中需要对大量的候选模型完全训练，而深度语音增强模型的训练本身就是一个计算成本极高的任务。因此，为进一步降低 ES 过程中语音增强模型性能评估的计算开销。本节提出一种低成本模型性能评估策略（LC-MPES, low-cost model performance evaluation strategy），通过观测不同性能模型下输入的原图及对应特征图间的差异性与模型真实准确率的隐含关系，设计搭建性能评估矩阵，实现不需要对候选模型完全训练即可对其性能进行精细化快速近似评估。

该策略使用平均哈希值[20]来计算 2 张图片的差异性，具有性能好、计算速度快等特点。平均哈希值等于 2 张图片分别对应的 2 个哈希矩阵之间的汉明距离，2 张同分类相似图片 cm 与 ck 的平均哈希值 $h_{\mathrm{cm,ck}}$（或 cm 与 ck 对应特征图的平均哈希值 $H_{\mathrm{cm,ck}}$）越大，代表 2 张图片差距越大。$h_{\mathrm{cm,ck}}$ 具体计算如式(3)~式(5)所示。

$$\boldsymbol{x}=\text{One - dimensional}(\boldsymbol{D}_{\mathrm{cm}}) \tag{3}$$

$$\boldsymbol{y}=\text{One - dimensional}(\boldsymbol{D}_{\mathrm{ck}}) \tag{4}$$

$$h_{\mathrm{cm,ck}}=\text{Hamming}(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

其中，$\text{One - dimensional}(\boldsymbol{D}_{\mathrm{cm}})$ 表示将 cm 对应的哈希矩阵 $\boldsymbol{D}_{\mathrm{cm}}$ 转为一维向量，$\text{Hamming}(\boldsymbol{x},\boldsymbol{y})$ 为向量 $\boldsymbol{x}$ 与向量 $\boldsymbol{y}$ 的汉明距离。

根据前期研究，性能优异的模型可更好地提取相似图片各自特有的特征信息，导致特征图比原图差异大，因此 2 张差异较小的同分类相似图片 cm 与 ck（$h_{\mathrm{cm,ck}}$ 较小）对应的 2 张特征图之间的差异很大（$H_{\mathrm{cm,ck}}$ 很大）；而性能较差的模型无法有效提取

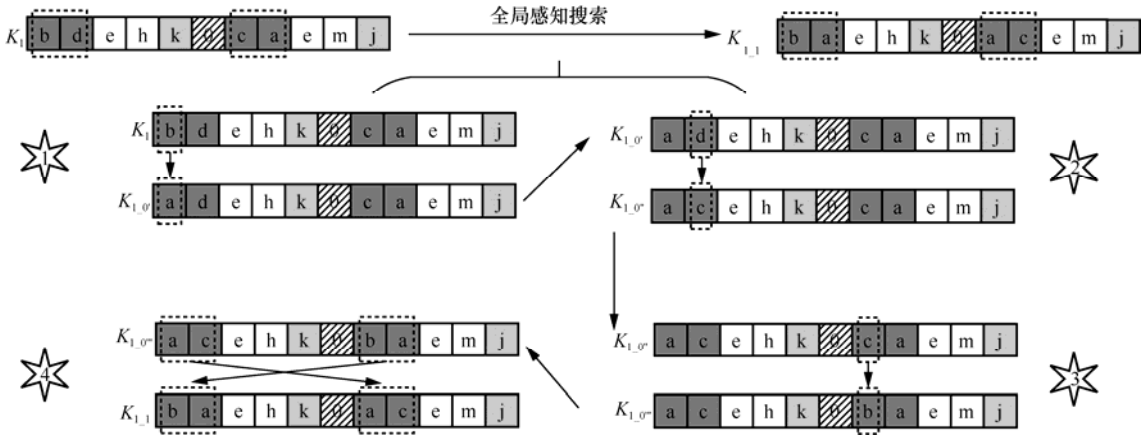


图 8　全局感知搜索

相似图片各自特有的特征信息，导致特征图较原图差异不大甚至相同，因此 2 张差异较小的同分类相似图片 cm 与 ck（$h_{\mathrm{cm,ck}}$ 较小）对应的 2 张特征图之间的差异也较小（$H_{\mathrm{cm,ck}}$ 较小）。故 $H_{\mathrm{cm,ck}}-h_{\mathrm{cm,ck}}$ 越大，则代表越多的图片信息被模型提取。据此，对每个模型评分时仅需将 n 张同分类小批次数据（128、64、32 或 16 张相似图片）输入待评测模型中，在相同位置为每张输入图片提取一张特征图，并构建差异矩阵 $\boldsymbol{U}$ 将数据对（原图及对应特征图）两两进行相似性对比，来计算模型预测分数 Grade，如式(6)和式(7)所示。

$$\boldsymbol{U}=\begin{pmatrix} L-|H_{\mathrm{c1,c1}}-h_{\mathrm{c1,c1}}| & \cdots & L-|H_{\mathrm{c1},cn}-h_{\mathrm{c1},cn}| \\ \vdots & \ddots & \vdots \\ L-|H_{cn,\mathrm{c1}}-h_{cn,\mathrm{c1}}| & \cdots & L-|H_{cn,cn}-h_{cn,cn}| \end{pmatrix} \tag{6}$$

$$\mathrm{Grade}=\ln|\boldsymbol{U}| \tag{7}$$

其中，L 为图片像素值，$H_{cm,ck}-h_{cm,ck}=0(m=k)$ 且 $H_{cm,ck}-h_{cm,ck}\leqslant L(m\neq k)$。根据式(6)可得，高性能模型具有较少的非对角线元素，且在理想情况下，性能非常优异的模型可使差异矩阵 $\boldsymbol{U}$ 主对角线元素均为 L，非主对角线元素均为 0，此时 Grade 达到最大值。因此矩阵越接近对角矩阵，Grade 越高，代表训练后最终准确率越高，利用该特点近似预测未经完全训练模型的最终性能，而不必将整个模型训练完成，极大地降低了模型评估成本。

3　实验

本节进行了全面的实验，以证明所提基于多域融合及神经架构搜索的语音增强方法的有效性。首先验证了本文所提低成本模型性能评估策略的有效性，之后在 2 个基准语音语料数据集 THCHS-30 和 WSJ0 上将本文方法与诸多手工设计和具备自优化能力的模型进行比较，全面评判了所提方法的综合性能和泛化性，最后给出了语音增强效果、模型参数量等方面的比较结果。

3.1　低成本模型性能评估策略性能验证

首先，对该策略的性能进行了定性分析。在 NAS-Bench-101[21] 基准搜索空间的 cifar10 及 NAS-Bench-201[22]基准搜索空间的 cifar10、cifar100 和 ImageNet16-120 数据集下分别采样 1 000 个模型，图 9 展示了在不同数据集与搜索空间下采样模

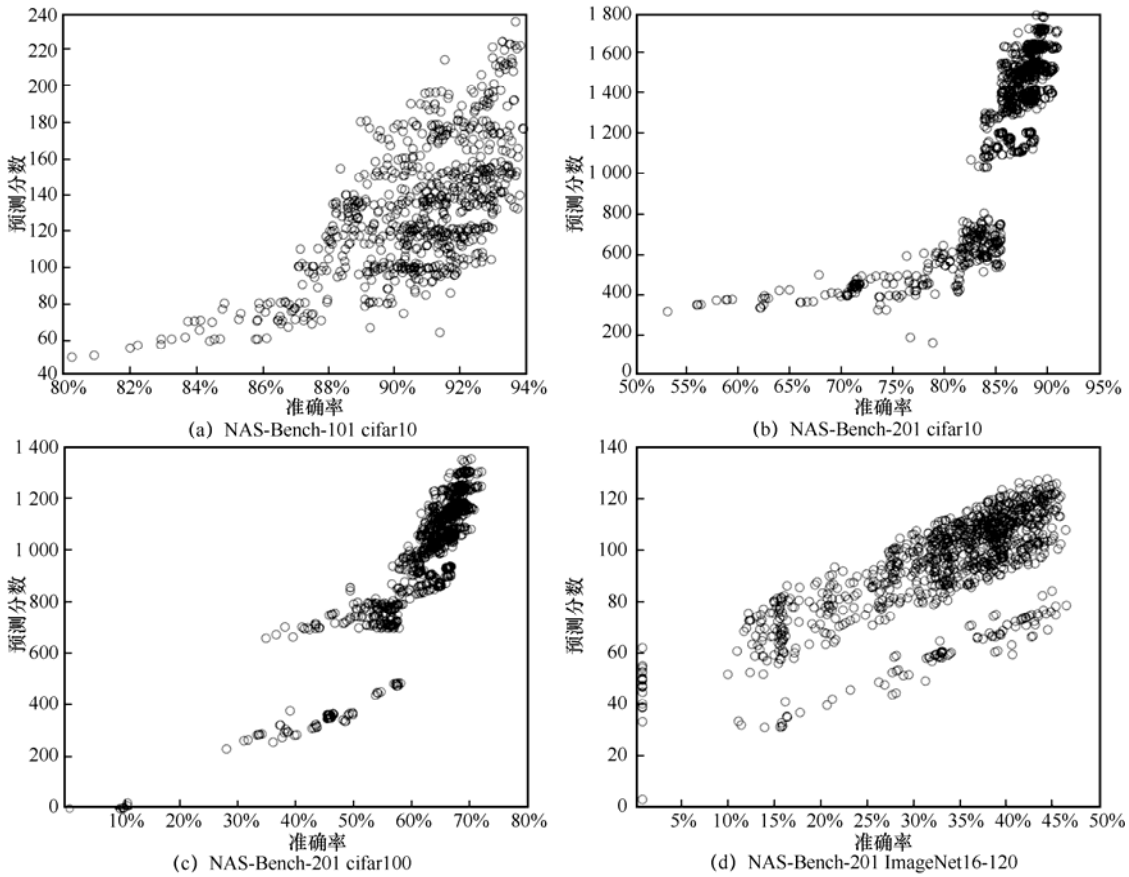


图 9　采样模型的真实准确率与预测分数 Grade 的相关性点状图

型的真实准确率与预测分数 Grade 的相关性点状图，模型真实准确率与预测分数存在较强的正相关关系。

究其根本原因，是本文创建的性能评估矩阵 $\boldsymbol{U}$ 挖掘了多张图片的相似性与差异性的像素信息，并将该信息进行逐一细粒度对比，捕获了原图与其对应特征图像素点间的差异程度，使高性能模型具有较少的非对角线元素且评分更高，表现良好的模型的评分矩阵更加接近对角矩阵，不同性能模型的评分差异更加均匀，评分误差更小，因此所提策略产生的预测分数排名与真实准确率排名具有更高的相关性。为了继续验证这一想法，本文使用 Kyriakides 等[23]提出的 Kendall's tau-b 相关系数度量方法，在 NAS-Bench-201 搜索空间与 ImageNet16-120 数据集下，将本文策略与 NAS-WOT[14]进行定量分析对比，通过 Kendall's tau-b 相关系数 τ（该系数可解决排名并列产生的误差问题）定量计算了不同策略下模型预测评分排名与真实准确率排名的相关性。因为排名会受到噪声的影响，准确率的微小变化会导致排名的巨大变化，因此计算时会将准确率四舍五入到其最接近的整数百分比。实验最终表明，在消除小样本量（N=10 或 100）带来的随机性误差后，本文策略比 NAS-WOT 具有更强的相关性，在样本量 N=1 000 时，本文策略相关性比 NAS-WOT 高 28.6%，这再次证明本文评估矩阵可捕获不同性能模型内在差异的能力，不同策略的模型评分排名与真实准确率排名的 Kendall's tau-b 相关系数如图 10 所示。

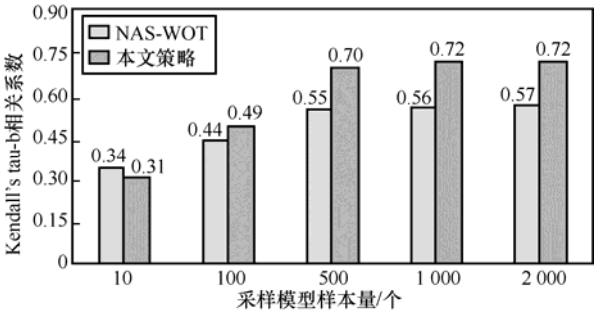


图 10　不同策略的模型评分排名与真实准确率排名的 Kendall's tau-b 相关系数

在上述分析后，为进一步验证本文策略的性能，对本文策略进行了定量度量，表 2 比较了使用 NAS-Bench-201 基准搜索空间与 cifar10、cifar100、ImageNet16-120 数据集的几种策略各运行 500 次（权重共享方法运行 3 次）准确率的平均值±标准差。策略分为 3 个部分：权重共享、非权重共享、低成本（LC-MPES 和 NAS-WOT 与随机搜索[24]结合）。表 2 中“Optimal”也报告了在不同样本量 N 的环境下，抽样到所有待评估模型中最优模型的准确率。

为证明本文策略与 NAS-WOT 策略数据对比的有效性，在不同数据集上将本文策略与 NAS-WOT 策略各自对模型的预测排名进行配对 t 检验（每组数据中模型样本量 N=100，置信区间为 95%），所有情况下均有 I<0.05，证明本文实验中不同策略的性能排名数据差异较大，因此在模型隐含特征挖掘的理论方法上，本文策略与对比策略存在较大差异，数据具有统计学意义。

从表 2 第三列可看出，本文策略需要更少数量级的时间来评估模型，而非权重共享和权重共享都需要较大的时间成本，该策略可在 505.9 s 内评估 2 000 个模型，除了 cifar10 和 cifar100 上的 GDAS 外，在所有数据集上都比权重共享策略获得了较好的结果，同时在时间上也优于 NAS-WOT，原因与创建相关矩阵的时间复杂度有直接关系，而相关矩阵高度依赖于数据点运算难度，本文策略降低了每个相关矩阵的数据运算难度，允许更快的计算。对于因时间复杂度而无法使用较大数据集进行搜索的 NAS 方法，则可直接集成 LC-MPES 进行快速模型评估。因此当多次重复 NAS 时，时效就变得非常重要，本文策略可能在未来能够为不同任务和资源环境廉价地设计专用模型，而每次设置只需要几秒。

与 NAS-WOT 比较可以看出，本文策略优于 NAS-WOT，具有更快的速度和准确性，能够更高效地选择高性能模型。NAS-WOT 策略出现样本量 N 增加但搜索到模型的平均准确率下降的问题，例如在表 2 中 3 个数据集下，N=1 000 的测试集准确率甚至比 N=10 的还要低。这主要是因为 NAS-WOT 的性能评估矩阵的可用参数较少，使其更擅长在小样本的空间（类似的拓扑结构和大小）内对模型进行评分，但在较大样本的空间对模型进行评分时则会产生噪声问题，随着评估样本数量增加，评估性能更加不稳定，该问题对于目前 NAS 中海量的待评估模型样本是致命的。而本文策略则正好相反，随着样本量 N 的增加，逐渐搜索到准确率更高的模型，这是非常重要的，因为最优模型不大可能出现在小样本量中。

3.2　所提方法在语音增强应用上的性能验证

为验证本文方法的综合性能，在实际语音增强

表 2　　不同策略性能对比

策略类型	策略	搜索时间/s	cifar10		cifar100		ImageNet16-120	
			验证集	测试集	验证集	测试集	验证集	测试集
非权重共享	REA	12 000	91.19±0.31	93.92±0.30	71.81±1.12	71.84±0.99	45.15±0.89	45.54±1.03
	RS	12 000	90.93±0.36	93.70±0.36	70.93±1.09	71.04±1.07	44.45±1.10	44.57±1.25
	REINFORCE	12 000	91.09±0.37	93.85±0.37	71.61±1.12	71.71±1.09	45.05±1.02	45.24±1.18
	BOHB	12 000	90.82±0.53	93.61±0.52	70.74±1.29	70.85±1.28	44.26±1.36	44.42±1.49
权重共享	RSPS	7 587	84.16±1.69	87.66±1.69	59.00±4.60	58.33±4.34	31.56±3.28	31.14±3.88
	DARTS-V1	10 890	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
	DARTS-V2	29 902	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
	GDAS	28 926	90.00±0.21	93.51±0.13	71.14±0.27	70.61±0.26	41.70±1.26	41.84±0.90
	SETN	31 010	82.25±5.17	86.19±4.63	56.86±7.59	56.87±7.77	32.54±3.63	31.90±4.07
	ENAS	13 315	39.77±0.00	54.30±0.00	15.03±0.00	15.61±0.00	16.43±0.00	16.32±0.00
低成本	I (本文策略及 NAS-WOT)	—	0.001 2		0.019 0		0.036 0	
	NAS-WOT (N=10)	3.6	89.16±1.56	91.40±1.13	69.26±2.25	69.10±2.06	41.98±4.01	41.20±4.11
	本文策略(N=10)	3.1	89.86±0.12	91.62±0.30	69.77±1.45	69.11±1.72	41.77±3.33	41.30±4.27
	NAS-WOT (N=100)	30.9	89.51±0.78	91.31±1.12	68.13±1.05	69.18±1.41	42.33±3.23	42.48±3.01
	本文策略(N=100)	28.6	89.73±1.16	91.64±0.17	68.28±2.01	69.19±0.82	39.99±3.25	41.52±4.45
	NAS-WOT (N=500)	130.2	88.90±0.61	91.61±1.07	68.52±1.22	68.04±1.41	39.69±2.05	39.77±2.10
	本文策略(N=500)	110.6	88.96±0.35	92.19±1.44	69.03±0.72	69.46±0.71	40.93±2.39	42.15±2.11
	NAS-WOT (N=1 000)	310.3	89.63±0.73	91.30±0.81	68.77±1.21	68.58±1.22	39.21±2.12	39.12±1.78
	本文策略(N=1 000)	256.2	89.97±1.85	92.33±1.01	69.68±0.92	69.56±0.93	41.73±2.03	42.77±1.98
	NAS-WOT (N=2 000)	601.5	89.90±1.44	91.33±0.99	69.33±1.41	69.98±2.22	40.21±2.11	40.32±3.08
	本文策略(N=2 000)	505.9	91.09±2.15	93.95±1.33	69.89±1.48	71.99±1.88	42.53±3.13	43.95±2.53
最优值	Optimal (N=10)	—	90.11±0.75	93.40±0.49	70.13±1.98	70.13±1.98	44.77±1.77	44.77±1.77
	Optimal (N=100)	—	91.11±0.12	94.02±0.11	72.81±0.90	72.81±0.90	46.01±0.47	46.01±0.47
	Optimal (N=500)	—	91.14±0.17	94.10±0.22	72.91±0.64	72.91±0.64	46.02±0.73	46.02±0.73
	Optimal (N=1 000)	—	91.32±0.11	94.20±0.14	72.93±0.41	72.93±0.41	46.62±0.57	46.62±0.57
	Optimal (N=2 000)	—	91.36±1.07	94.25±1.08	72.95±0.47	72.95±0.47	46.68±0.45	46.68±0.45

应用上进行了测试，使用本文策略在搜索空间中对最优联合 Cell 进行搜索并以此搭建最优语音增强模型，搜索过程融合了低成本模型性能评估策略进行加速。实验首先在 THCHS-30 中文数据集[25]上将该语音语料数据集和 café、babble、car 这 3 种不同类型的噪声相融合生成训练集和测试集来对比原带噪语音 Noisy 本文方法，传统方法 LSTM，时域增强方法 ConvTasNet，时频域复数增强方法 CRN、DCUNet、DCCRN，以及团队前期提出的具备自优化能力的 AMDCCRN[16]。该数据集包含 13 000 多条数据，将语音进行拼接，保证每个训练语音均为 10 s，按照 8:2 分配训练集和测试集，训练集在信噪比为−5～20 dB 随机进行语音噪声混合，而测试集在 5 个典型信噪比（0 dB、5 dB、10 dB、15 dB、20 dB）下生成。

在本文实验中，使用语音质量客观评价（PESQ, perceptual evaluation of speech quality）、短时客观可懂度（STOI, short-time objective intelligibility）、参数量作为评价指标。其中，PESQ 通过比较 2 个输入信号的时频域参数的差异，可以得到−0.5～4.5 的客观语音评分值；STOI 用于评估语音的可理解性，范围为 0～1，平均分越高，语音质量越好。模型窗口长度和跳点大小分别为 25 ms 和 6.25 ms，FFT 长度为 512，语音信号都在 16 kHz 下进行采样，使用 Adam 优化器和 SI-SNR 损失函数，学习率为 1×10^{-4}，训练周期为 8，批次大小为 6。基线模型参数及本文方法在不同语音数据下的辅助域选取均遵循团队前期研究[16]，搜索评估策略参数设置如表 3 所示。

表 3　　搜索评估策略参数设置

参数	取值
种群大小	20
进化次数	20
交叉概率	0.5
变异概率	0.5
全局感知次数	5
局部细化次数	5
评估小批次大小	128

THCHS-30 数据集在 café、babble、car 这 3 种不同类型的噪声集下分别搜索到的最优联合 Cell 拓扑如图 11 所示。之后将最优联合 Cell 搭建的模型在不同噪声集下与其他基线模型进行对比，实验结果如表 4～表 6 所示。

不同噪声类型对语音增强效果影响很大，因此为了更好地验证本文方法的有效性，本文使用 3 种不同类型的噪声集来进行全方位的实验。本文方法在模型参数量为最小的情况下，STOI 和 PESQ 指标总体上优于 AMDCCRN、DCUNet、ConvTasNet、DCCRN、LSTM、CRN 基线。在 café 噪声下，本文方法在 PESQ 和 STOI 上都得到了不同程度的提升。原 DCCRN 较 DCUNet 和 ConvTasNet 在部分情况下可能存在性能较差或相同的情况，但本文方法在任何信噪比（SNR）均优于其余基线模型，相较于次优的 AMDCCRN 基线，本文方法在 STOI 指标上最大取得了 0.03 的提升，在 PESQ 指标上最大取得了 0.06 的提升，而相较于 DCCRN 则取得了 0.50 的提升。在 babble 噪声下，本文方法效果较优，

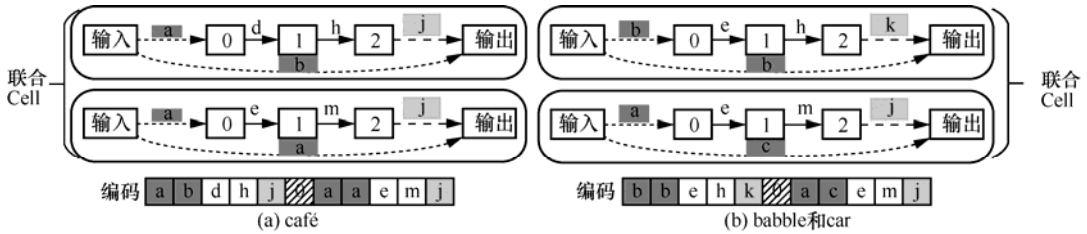


图 11　THCHS-30 数据集在 3 种噪声集下分别搜索到的最优联合 Cell 拓扑

表 4　　THCHS-30 与 café 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.07	0.71	2.11	0.73	2.23	0.77	2.30	0.80	2.39	0.81	—	—
CRN	2.42	0.74	2.45	0.77	2.50	0.80	2.58	0.81	2.64	0.86	6.1×10^6	—
LSTM	2.41	0.73	2.40	0.75	2.49	0.79	2.59	0.81	2.62	0.83	9.6×10^6	—
DCUNet	2.45	0.74	2.48	0.79	2.52	0.81	2.61	0.83	2.70	0.88	3.6×10^6	—
ConvTasNet	2.42	0.74	2.46	0.78	2.51	0.80	2.60	0.80	2.66	0.86	5.1×10^6	—
DCCRN	2.56	0.78	2.57	0.83	2.61	0.84	2.69	0.87	2.70	0.91	3.7×10^6	—
AMDCCRN	2.74	0.81	2.76	0.85	2.85	0.90	2.96	0.92	3.17	0.94	3.6×10^6	GASF
本文方法	2.75	0.82	2.77	0.88	2.86	0.92	3.02	0.93	3.20	0.95	3.5×10^6	GASF

表 5　　THCHS-30 与 babble 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.08	0.72	2.12	0.72	2.25	0.76	2.31	0.80	2.42	0.81	—	—
CRN	2.43	0.76	2.47	0.79	2.52	0.83	2.60	0.84	2.67	0.90	6.1×10^6	—
LSTM	2.42	0.76	2.41	0.77	2.50	0.82	2.59	0.84	2.63	0.86	9.6×10^6	—
DCUNet	2.45	0.75	2.47	0.82	2.53	0.84	2.60	0.85	2.72	0.90	3.6×10^6	—
ConvTasNet	2.43	0.74	2.47	0.80	2.52	0.83	2.60	0.84	2.69	0.91	5.1×10^6	—
DCCRN	2.55	0.79	2.56	0.86	2.61	0.87	2.67	0.90	2.69	0.92	3.7×10^6	—
AMDCCRN	2.75	0.82	2.75	0.87	2.86	0.91	2.97	0.94	3.18	0.95	3.7×10^6	GASF
本文方法	2.75	0.81	2.77	0.85	2.89	0.93	3.01	0.95	3.21	0.95	3.6×10^6	GASF

表 6　　THCHS-30 与 car 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.09	0.71	2.11	0.72	2.24	0.77	2.30	0.79	2.41	0.80	—	—
CRN	2.41	0.75	2.45	0.78	2.51	0.82	2.60	0.84	2.65	0.89	6.1×10^6	—
LSTM	2.42	0.76	2.44	0.77	2.50	0.81	2.59	0.83	2.64	0.87	9.6×10^6	—
DCUNet	2.44	0.75	2.47	0.83	2.52	0.85	2.61	0.87	2.73	0.92	3.6×10^6	—
ConvTasNet	2.42	0.74	2.46	0.81	2.51	0.84	2.62	0.85	2.70	0.92	5.1×10^6	—
DCCRN	2.54	0.79	2.55	0.86	2.62	0.88	2.67	0.90	2.75	0.93	3.7×10^6	—
AMDCCRN	2.74	0.85	2.76	0.89	2.85	0.92	2.99	0.94	3.15	0.95	3.7×10^6	GADF
本文方法	2.74	0.86	2.78	0.90	3.01	0.94	3.05	0.94	3.17	0.95	3.6×10^6	GADF

在 STOI 上较 DCCRN 最大取得 0.06 的提升，在 PESQ 上最大取得 0.52 的提升，研究发现，DCCRN 可能会过度抑制某些片段上的语音信号，导致较差的听力体验。在 car 噪声下，本文方法同样较其余基线模型有所提升。实验表明，本文方法具备较好的模型自适应构建能力，能够在减少人力设计时间成本的基础上，自适应地构建模型，同时也证明了本文模型在不同噪声类型干扰、不同信噪比噪声干扰下均保持良好状态且有较高的可用性。

3.3　语音增强性能泛化性验证

由于单一数据集上作测试难以验证本文方法的有效性以及泛化能力，因此本节在 WSJ0 英文数据集[26]中选择了部分话语（约 50 h），包括 131 个说话人且男女数量均等，同样将该语音语料数据集与 café、babble、car 这 3 种不同类型的噪声融合生成训练集和测试集（训练集及测试集设置方法同 3.2 节）。将搜索到的最优联合 Cell 搭建的模型在不同噪声集下对比本文方法，传统方法 LSTM，时域增强方法 ConvTasNet，时频域复数增强方法 CRN、DCUNet、DCCRN，以及团队前期提出的具备自优化能力的 AMDCCRN[16]。实验结果如表 7～表 9 所示。

从表 7～表 9 中可以看出，在泛化性验证数据集上，本文方法搜索到的最优联合 Cell 所搭建的模型同样也取得了良好效果，在 20 dB 时，STOI 均达到 0.95 的高分，PESQ 最高达到 3.89，且在任何情况下本文方法在参数量上均为最优。虽然 DCCRN 和 DCUNet、ConvTasNet 在不同噪声影响下存在性能相同或相差不大的情况，但本文方法在 THCHS-30 与 WSJ0 两大数据集上的 STOI 和 PESQ 均较其余基线模型有所提升且总体优于 AMDCCRN 模型，与 3.2 节结论相同。该泛化实验进一步验证出本文方法可以提高语音增强效果且泛化能力较强，证明了本文方法在语音增强领域的有效性。

表 7　　WSJ0 与 café 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.06	0.72	2.38	0.74	2.72	0.77	3.05	0.80	3.36	0.82	—	—
CRN	2.85	0.75	3.13	0.77	3.35	0.81	3.57	0.84	3.68	0.87	6.1×10^6	—
LSTM	2.78	0.74	3.09	0.75	3.35	0.80	3.57	0.83	3.70	0.85	9.6×10^6	—
DCUNet	2.83	0.76	3.19	0.79	3.50	0.83	3.71	0.84	3.80	0.88	3.6×10^6	—
ConvTasNet	2.85	0.76	3.16	0.77	3.45	0.80	3.63	0.85	3.73	0.87	5.1×10^6	—
DCCRN	2.86	0.78	3.20	0.82	3.51	0.84	3.72	0.86	3.83	0.90	3.7×10^6	—
AMDCCRN	2.96	0.81	3.30	0.85	3.54	0.89	3.77	0.92	3.85	0.94	3.6×10^6	GASF
本文方法	2.96	0.82	3.32	0.86	3.56	0.90	3.80	0.92	3.87	0.95	3.5×10^6	GASF

表 8　　WSJ0 与 babble 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.07	0.73	2.39	0.75	2.74	0.77	3.07	0.81	3.38	0.83	—	—
CRN	2.88	0.75	3.16	0.77	3.36	0.80	3.58	0.84	3.69	0.86	6.1×10^6	—
LSTM	2.80	0.75	3.09	0.76	3.37	0.80	3.56	0.83	3.71	0.84	9.6×10^6	—
DCUNet	2.88	0.77	3.17	0.80	3.49	0.84	3.70	0.85	3.83	0.90	3.6×10^6	—
ConvTasNet	2.88	0.77	3.16	0.79	3.46	0.82	3.67	0.84	3.76	0.88	5.1×10^6	—
DCCRN	2.89	0.78	3.20	0.82	3.51	0.85	3.74	0.87	3.84	0.90	3.7×10^6	—
AMDCCRN	2.97	0.82	3.31	0.86	3.56	0.90	3.79	0.91	3.85	0.94	3.5×10^6	GADF
本文方法	2.98	0.81	3.31	0.87	3.57	0.92	3.81	0.94	3.87	0.95	3.5×10^6	GADF

表 9　　WSJ0 与 car 噪声集语音增强结果

模型	SNR=0		SNR=5 dB		SNR=10 dB		SNR=15 dB		SNR=20 dB		参数量	辅助域
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI		
Noisy	2.07	0.72	2.38	0.73	2.72	0.75	3.03	0.80	3.34	0.81	—	—
CRN	2.86	0.75	3.16	0.76	3.35	0.81	3.55	0.84	3.67	0.87	6.1×10^6	—
LSTM	2.80	0.73	3.07	0.75	3.36	0.78	3.56	0.82	3.69	0.85	9.6×10^6	—
DCUNet	2.88	0.77	3.17	0.79	3.37	0.83	3.70	0.86	3.81	0.89	3.6×10^6	—
ConvTasNet	2.89	0.76	3.17	0.78	3.35	0.82	3.68	0.85	3.77	0.88	5.1×10^6	—
DCCRN	2.89	0.76	3.19	0.83	3.49	0.86	3.74	0.86	3.83	0.90	3.7×10^6	—
AMDCCRN	2.96	0.82	3.31	0.85	3.55	0.89	3.79	0.90	3.85	0.95	3.7×10^6	MTF
本文方法	2.98	0.83	3.36	0.87	3.55	0.94	3.83	0.95	3.89	0.95	3.6×10^6	MTF

4　结束语

本文提出了一种基于多域融合及神经架构搜索的语音增强方法。首先，将一维语音信号信息映射至实数/复数域等多个空间域中，提取更丰富的语音特征，增强了一维语音信号特征完备性、关联性的表达，并在此基础上为语音增强模型设计了复数特征融合机制将基础域和表征能力好的辅助域所提取的不同特征相融合。其次，根据语音信号的特点设计了一种基于联合 Cell 的可分离复数搜索空间，利用卷积及池化的特性，将每个联合 Cell 中待搜索节点分离为卷积节点及池化节点，降低搜索空间复杂度和体积，并在此搜索空间基础上结合多域融合机制得到最终的语音增强模型。再次，根据搜索空间的特性，提出高性能自适应全局/局部协同特征感知的搜索策略来搜索语音增强模型所需的最优联合 Cell。最后，为进一步提高模型的评估速度，提出低成本模型性能评估策略，在不需要完全训练模型的情况下，为候选模型进行评分来对模型性能进行精细化近似评估。经过本文方法搜索到的最优语音增强模型在 THCHS-30 和 WSJ0 这 2 个语音语料数据集下与其他语音增强基线模型的鲁棒性和泛化性实验中，在模型参数量较低的情况下，PESQ 和 STOI 两大指标较其他基线模型均有提升，更好地进行了噪声抑制，验证了本文方法的效率和有效性。

参考文献：

[1] 解元, 邹涛, 孙为军, 等. 面向高混响环境的欠定卷积盲源分离算法[J]. 通信学报, 2023, 44(2): 82-93.
XIE Y, ZOU T, SUN W J, et al. Algorithm of underdetermined convolutive blind source separation for high reverberation environment[J]. Journal on Communications, 2023, 44(2): 82-93.

[2] GHOLAMIANGONABADI D, GROLINGER K. Personalized models for human activity recognition with wearable sensors: deep neural networks and signal processing[J]. Applied Intelligence, 2023, 53(5): 6041-6061.

[3] YIN D C, LUO C, XIONG Z W, et al. PHASEN: a phase-and-harmonics-aware speech enhancement network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 9458-9465.

[4] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]//Proceedings of the Interspeech. Hyderabad: ISCA Press, 2018: 3229-3233.

[5] CHOI H S, KIM J H, HUH J, et al. Phase-aware speech enhancement with deep complex U-net[J]. arXiv Preprint, arXiv: 1903.03107, 2019.

[6] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[J]. arXiv Preprint, arXiv: 1505.04597, 2015.

[7] HU Y X, LIU Y, LV S B, et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement[C]//Proceedings of the Interspeech. Hyderabad: ISCA Press, 2020: 2472-2476.

[8] BIAN Y J, SONG Q Q, DU M N, et al. Subarchitecture ensemble pruning in neural architecture search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7928-7936.

[9] BAKER B, GUPTA O, NAIK N, et al. Designing neural network architectures using reinforcement learning[J]. arXiv Preprint, arXiv: 1611.02167, 2016.

[10] BEECHE C, SINGH J P, LEADER J K, et al. Super U-Net: a modularized generalizable architecture[J]. Pattern Recognition, 2022, 128: 108669.

[11] LIU C X, ZOPH B, NEUMANN M, et al. Progressive neural architecture search[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 19-35.

[12] HUANG L, SUN S Q, ZENG J, et al. U-DARTS: uniform-space differentiable architecture search[J]. Information Sciences, 2023, 628: 339-349.

[13] LUO R Q, TIAN F, QIN T, et al. Neural architecture optimization[J]. arXiv Preprint, arXiv: 1808.07233, 2018.

[14] MELLOR J, TURNER J, STORKEY A, et al. Neural architecture search without training[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 7588-7598.

[15] LOPES V, ALIREZAZADEH S, ALEXANDRE L A. EPE-NAS: efficient performance estimation without training for neural architecture search[C]//International Conference on Artificial Neural Networks. Berlin: Springer, 2021: 552-563.

[16] ZHANG R, ZHANG P Y, GAO M R, et al. Self-optimizing multi-domain auxiliary fusion deep complex convolution recurrent network for speech enhancement[J]. Digital Signal Processing, 2023, 134: 103897.

[17] XIE L X, YUILLE A. Genetic CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 1388-1397.

[18] 胡向东, 吕高飞, 白银. 基于优化支持向量回归的工业互联网安全态势预测方法[J]. 电子学报, 2023, 51(2): 446-454.
HU X D, LYU G F, BAI Y. A method of security situation prediction for industrial Internet based on optimized support vector regression[J]. Acta Electronica Sinica, 2023, 51(2): 446-454.

[19] WANG X Y, LIU P B. Image encryption based on roulette cascaded chaotic system and alienated image library[J]. The Visual Computer, 2022, 38(3): 763-779.

[20] ZHANG S X, YANG Y, ZHANG M, et al. A mutli-feature correlation filter tracker with different hash algorithm[C]//Proceedings of the 2021 IEEE 24th International Conference on Information Fusion (FUSION). Piscataway: IEEE Press, 2021: 1-6.

[21] YING C, KLEIN A, CHRISTIANSEN E, et al. NAS-Bench-101: towards reproducible neural architecture search[C]//Proceedings of the 36th International Conference on Machine Learning. New York: PMLR, 2019: 7105-7114.

[22] DONG X Y, YANG Y. NAS-Bench-201: extending the scope of reproducible neural architecture search[J]. arXiv Preprint, arXiv: 2001.00326, 2020.

[23] KYRIAKIDES G, MARGARITIS K. The effect of reduced training in neural architecture search[J]. Neural Computing and Applications, 2020, 32(23): 17321-17332.

[24] LI L, TALWALKAR A. Random search and reproducibility for neural architecture search[C]//Proceedings of Uncertainty in Artificial Intelligence. New York: PMLR, 2020: 367-377.

[25] WANG D, ZHANG X W. THCHS-30: a free chinese speech corpus[J]. arXiv Preprint, arXiv:1512.01882, 2015.

[26] GAROFOLO J, GRAFF D, PAUL D, et al. Linguistic data consortium CSR-I (WSJ0) database[R]. 1993.

[作者简介]

张睿（1987– ），男，山西太原人，博士，太原科技大学副教授、硕士生导师，主要研究方向为智能信息处理、自动机器学习等。

张鹏云（1999– ），男，河北安平人，太原科技大学硕士生，主要研究方向为智能信息处理、自动机器学习等。

孙超利（1978– ），女，浙江诸暨人，博士，太原科技大学教授、博士生导师，主要研究方向为计算智能、机器学习等。