DOI:10.19651/j.cnki.emt.2106213

基于深度学习的声源无网格定位及量化方法

王言彬 徐长秋 毛富哲
(渤海造船厂集团有限公司船舶设计研究院 葫芦岛 125000)

摘 要: 波束形成技术是一种常用的声源定位方法,但在多数研究中声源强度往往没有考虑到。为准确定位和量化复杂环境下的单一点声源,在常规波束形成图的基础上,提出一种基于残差网络的声源定位及其强度估计方法,旨在精确预测点声源的位置和强度。研究采用 Acoular 软件模拟时间信号,对神经网络进行训练得到预测模型,并通过计算机仿真对深度神经网络能否从麦克风阵列数据中得到单一点源的精确描述进行了验证。结果表明,该方法不仅能够快速准确地给出单一点声源的位置和强度,其中距离误差 $e_{dist}/\Delta x \approx 0.15$,水平误差均值 $\bar{e}_{level} \approx 0.002$ dB,且对于较大的频率有更好的预测效果。

关键词: 深度学习;残差网络;声源定位;波束形成
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 510.1050

Meshless localization and quantization of sound source based on deep learning

Wang Yanbin Xu Changqiu Mao Fuzhe
(Bohai Shipyard Group Co., Ltd., Huludao 125000, China)

Abstract: Beamforming is a particular method for sound source localization. However, in most of the relating researches, the intensity of the source is often ignored. Therefore, this research proposes a method of sound source localization and intensity estimation based on the residual network based on the conventional beamforming map, aiming to accurately predict the position and intensity of the point sound source. Acoular software is adopted to simulate time signals, the neural network is trained by the simulated signals, and the prediction model can be obtained then. It is verified by computer simulation whether the deep neural network can obtain an accurate description of a single point source from the microphone array data. The results show that the proposed method predicts the location and the intensity of the sound source fast and effectively with distance error $e_{dist}/\Delta x \approx 0.15$ and level error $\bar{e}_{level} \approx 0.002$ dB. Moreover, the proposed method behaves better prediction effect for higher frequencies.

Keywords: deep learning; residual network; sound source localization; beamforming

0 引 言

近年来,相控麦克风阵列技术在声源定位和表征方面的价值已得到肯定。声源的空间分布及其强度可以用常规的波束形成方法加以评估。然而,视觉表征往往呈现出错误的源强度重构和有限的空间分辨率,尤其是当波长较大时这种现象更为明显。针对该问题,相关学者已经给出了大量的研究[1]。这些方法大多是基于模型的反演来实现的,旨在重构基于波束形成图的真实源分布,同时需要很高的计算能力和对声场特性的精确描述,其中一些方法还依赖于复杂的数学求解策略。

近年来,深度学习在图像识别、人工智能及其虚拟现实等领域取得了飞速的发展[2-5],而基于神经网络的麦克风阵列数据的声源表征研究却寥寥无几。现有的研究大多是关于医学超声影像方面的研究[6-8]。Reiter 等[6]采用 Alex-Net[7]结构证明了根据传播波阵面的光声图像来可靠估计点源位置是可行的。Allman 等学者对该方法进行了更深入的研究[8-9]。也有学者在传统的目标检测方法中加入了卷积神经网络结构来识别混响环境中的点声源[10]。区别于传统目标检测对目标进行分类并确定其在图像中的位置这一主要目的,该方法可以在定位声源的同时,对光声图像中的真实声源和虚拟声源加以区分。

基于上述研究背景,本文采用一种深度神经网络模型从麦克风阵列数据中得到单一点声源位置及强度的精确描

收稿日期:2021-03-29

述。以残差神经网络为基础结构，实现了声源位置和强度的预测，并通过仿真实验验证了模型的精度。

1　方法理论

设常规波束形成图像为 $\boldsymbol{B}=\boldsymbol{R}^{51\times51}$ 到声源向量 $\boldsymbol{y}=[x,\ y,\ p^2]^{\mathrm{T}}\in\boldsymbol{R}^3$ 之间的映射关系为 $\mathcal{F}_M:\boldsymbol{R}^{51\times5}\rightarrow\boldsymbol{R}^3$。其中 $[x,\ y]$ 表示声源的位置，p^2 则表示声源的功率。输入数据的维度(51×51)由所使用的样本数据决定。样本数据将在后文中详细介绍。

1.1　模型结构

二维的常规波束形成(conventional beamforming，CB)图由与空间有关的数据构成。卷积神经网络(convolutional neural network，CNN)是一种可以处理与空间有关数据的神经网络结构[11]。通常，CNN 多用于图像处理相关研究中。它的卷积层由过滤器组成，通常称为内核，在多维输入数据上执行卷积或互相关操作。本文采用残差网络(ResNet)[12-13]建立神经网络模型。残差网络是 CNN 的中结构，相关研究已经证明该网络模型可以有效地进行样本分类和回归，同时可用于处理小维度图像。而波束形成图的维度往往小于图像识别研究中的图像维度[14]，这为使用 CNN 来处理 CB 图像提供了可行性。由于残差网络存在额外的快捷链，其网络层数与 CNN 不同。快捷连接是附加的处理路径，输入数据可以与过滤器路径并行地从输入层流到输出层。快捷连接中并未对数据进行修改，因此称之为标识映射。在输出层，标识映射与过滤后的数据合并。因此，残差网络只学习从输入层到输出层之间的残差，这种网络结构解决了深度学习中训练误差增加和精度饱和的问题。本文所用残差网络模型为 TensorFlow GitHub 库中体统的正式模型[15]，网络输出层为被预测声源位置和强度。表 1 所示为该模型每一步的处理信息。网络的输入层是一个由 26 个波束形成图线性卷积滤波器组成的卷积层，每个滤波器包含 3×3 个数值。在卷积层的输入层对波束形成图进行补 0，卷积后数据维度仍然保持 51×51。然后对 26 个输入图像进行变换，并合并为一个 26 通道的新图像(51×51×26)，这个新图像将作为后续处理的输入。后续 3 个步骤包含了 3 个连续剩余块的数据转化，其中每一个块由 3 个剩余单元构成。本文采用了 He 等[12]提出的优化剩余单元，各单元间数据的变换方式为：首先采用批量归一化方法对每个单元的输入数据进行缩放和移位[15]；然后采用 ReLU 函数对归一化的输入数据进行像素非线性变换，其中 $\mathrm{ReLU}(x)=\max\{x,\ 0\}$；接下来是一个权值层，将数据与一定数量的滤波器进行卷积。剩余单元将上述 3 个步骤重复两次之后，将标识映射的输入赋加到转换后的数据，然后输出；其中剩余块的权重层由不同数量的过滤器组成，对于剩余块 1 而言，每个权重层有 26 个滤波器进行卷积、剩余块 2 的每个权重层有 52 个滤波器进行卷积、最后一个剩余块的每个权重层则有 104 个滤波器进行卷积。每次卷积后剩余层中卷积的步长变为原来的 2 倍，这将使前两个维度上输出数据的维度减少为输入数据维度的 1/2。而剩余块输出数据的第 3 维取决于相应的剩余块中使用滤波器的数量。因此，最后一个剩余块的输出为 104 个 7×7 维的图像。接下来，平均池化层对数据进行进一步降维，该层不对数据进行训练，仅对上一层输出的 104 个图像分别求取其均值，构成一个 104 维的向量。最后在输出层对向量进行线性变换，将一个 3×104 维的可训练权重矩阵与向量相乘，即可得到声源的方位 $\{x,\ y\}$ 与功率 p^2。

表 1　神经网络结构中的块

块	输出维度	核数量	核维度
卷积层	51×51×26	26	3×3
剩余块 1	26×26×26	$\begin{bmatrix}26\\26\end{bmatrix}\times3$	$\begin{bmatrix}3\times3\\3\times3\end{bmatrix}\times3$
剩余块 2	13×13×52	$\begin{bmatrix}52\\52\end{bmatrix}\times3$	$\begin{bmatrix}3\times3\\3\times3\end{bmatrix}\times3$
剩余块 3	7×7×104	$\begin{bmatrix}104\\104\end{bmatrix}\times3$	$\begin{bmatrix}3\times3\\3\times3\end{bmatrix}\times3$
平均池化层	104×1	1	7×7
输出层	3×1	1	3×104

其中卷积层的输入维度为 51×51×1；任意后续层的输入维度都对应于前一层的输出维度。

1.2　训练和测试数据集

由于测量数据库中的样本含量不足以对神经网络进行训练，本文采用 Acoular 软件[16]来模拟时间信号，并计算 CB 图像。数据集的特性按照文献[17]中的描述设置，用蒙特卡罗方法创建了不同的声源星座来研究波束形成算法的性能。数据集的特性如表 2 所示，训练数据和测试数据的特性并不完全相同，以提升网络的泛化行。考虑到 PSF 是移变的，若无显式训练则无法假定该算法可以将一个区域内的局部映射属性传递给其他区域。同时，表 2 还给出了用于模拟真实值的测试数据在传感器位置上的微小偏差。移变的大小则根据 Herod 的研究成果确定[18]。传感器的偏差根据二元正态分布绘制，其标准差为 1/3 倍两个传声器之间的最小距离($d_{\min}=0.04d$)，其中 $d=0.686$ m 为孔径。本文采用无量纲的亥姆霍兹常数来表示频率。

$$\mathrm{Hc}=\frac{f\cdot d}{c}\tag{1}$$

式中：f 和 c 分别表示频率和声速。

1.3　模型优化与评价

采用无监督学习的方法对模型进行优化。在监督学习中，模型通过学习来逼近给定样本对之间的潜在映射关系，即训练。一个样本对包含输入数据和模型的期望输出，即目标。通常，训练数据集中样本对的子集 S 在每个训练步

表 2　数据集属性

属性	训练数据	测试数据
传感器阵列	64 个传感器,孔径 d =0.686 m	
聚焦栅	x, $y\in[-0.5d, 0.5d]$, $z=0.5d$, $\Delta x=0.02d$	
信号	宽带白噪声(互不相关)	
采样率	He=40	
时域样本数量	512 000	
块大小	1 024	
块重叠	50%	
窗函数	汉宁窗	
CSM 主对角线	去除	
导向矢量	Ⅲ	
评价依据	第 3 倍频带	
频率范围	$He_{min}=1$, $He_{max}=16$	
第三倍频带数	13	
声源分布	正态分布($\delta=0.168\,8$)	正态分布($\delta=0.168\,8$)
声源位置数	10 000	613
声图像数	130 000	7969
传感器干扰	无	正态分布($\delta=0.04d$)

骤中进行评估。估计输出与期望目标之间的差值由损失函数 L 来度量。在训练中,损失函数 L 是一批样本中模型预测与目标值之间误差的平均和。

$$L(\theta)=\frac{1}{S}\sum_{s=1}^{|S|}e(\boldsymbol{y}^{(s)},\hat{\boldsymbol{y}}^{(s)}),\quad \boldsymbol{y}=[x,y,p^2]^{\mathrm{T}} \tag{2}$$

式中:$\boldsymbol{y}$ 和 $\hat{\boldsymbol{y}}$ 分别表示预测值和目标值;e 表示均方根误差,用于衡量模型预测精度。根据计算得到的误差,采用随机梯度优化法对网络的权值进行调整。因此,目标函数的梯度 $\nabla_\theta L$ 是依据每一次迭代 t 中所有网络训练参数 θ,由反向传播法计算的[17]。He 的研究中给出了 Res 网络中反向传播的性能分析[11]。以计算的梯度为依据,在每次迭代中通过优化算法对网络权值进行调整。

网络的性能通常是经过大量的训练后,以测试数据集中的样本来交叉验证。在交叉验证中,通过对整个测试集的误差取均值并评估每例错误来计算交叉验证损失,即测试损失。在优化过程中,额外对距离误差及水平误差进行评价,但二者并未集成于整体优化过程中。距离误差是通过计算声源真实位置与预测位置的欧几里德范数来评价的;水平误差是通过计算真实声压与预测声压的平方对数比评价的。

$$e_{dist}=\|\boldsymbol{x}-\hat{\boldsymbol{x}}\| \tag{3}$$

$$e_{level}=|\Delta L_{p,e,s}|=\left|10\lg\frac{p^2}{\hat{p}^2}\mathrm{dB}\right| \tag{4}$$

式中:符号 $|\cdot|$ 表示绝对值。

1.4　神经网络超参数

在神经网络模型中,除可训练参数外,还存在其他影响模型训练性能的参数,即超参数[19]。在基于梯度的优化算法中,学习速率、学习速率表以及批量大小等超参数通常于优化算法的参数典型相关,且需要提前确定。学习速率 η 在训练步骤中用于调整网络权重的步长,其大小依赖于计算得到的梯度;由于本文在训练中所用优化算法为 Adam 优化算法,该算法在训练过程中对学习速率加以调整,因此与学习速率表无关[20];如前文所述,批量大小 $|s|$ 是训练集的一个样本子集的大小,典型的 $|s|$ 样本容量大约在 1 到几百个之间[20]。

本文 η 和 $|s|$ 由随机搜索确定[21],其中每个超参数的值来自于随机样本空间。学习速率 η 取目标区间 $[10^{-7}, 10^{-2}]$ 对数采样;批量大小 $|s|$ 为二次方值序列均匀采样($|s|\in\{8, 16, 32, 64, 128\}$)。将超参数采样值结合以建立训练的超参数配置。通过对不同超参数配置的网络进行训练并在测试集上对损失函数进行评价,即可确定最佳的超参数组合。

2　仿真实验及结果

每个给定超参数配置 210 568 的 54 特定优化程序都独立地在一个 CPU 集群节点上运行。每个节点包含 4 个 Intel Xeon CPU E5-2620 v4 处理器(32 位)。图 1 所示为一个优化过程中所包含的步骤,共执行 50 个不同超参数配置的优化过程。在每一次优化过程中,均对模型进行 150 次交叉验证。在每一次验证中,使用 16 000 个不同的样本在交叉验证步骤中对模型进行训练,该过程即为一次训练。也就是说,假设训练集中包含 32 个样本,那么每次训练需要进行 500 次迭代。交叉验证步骤给出了测试数据集上的最小损失值。

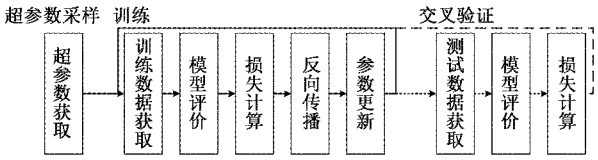

图 1　优化算法流程

当超参数配置为 $|s|=16$、$\eta=1.5\times10^{-4}$ 时,测试数据集的损失 $L=9.7\times10^{-6}$ 最小。在训练过程中,评价每秒约处理 38 张图像,共需要 18 h 可找到全局最小值。共对 715 965 个模型参数进行优化。图 2 所示为训练和交叉验证过程中 L 的变化情况。由图 2 可知,目标函数逐渐衰减并收敛。由于两条曲线随着时间的增加不断衰减,且评价损失在迭代结尾处没有增加,由此可认为已达到全局最小值。

图 3 所示为交叉验证中个体距离和水平误差与频率的关系,更深入地体现了模型的性能。值得注意的是交叉验证数据非常低的距离误差 $e_{dist}/\Delta x\approx0.15$。$e_{dist}/\Delta x<1$ 表示距离误差小于观测网格上两点之间的距离 Δx。图中黑色曲线表示误差,灰色阴影表示标准差。结合图 3 可知,

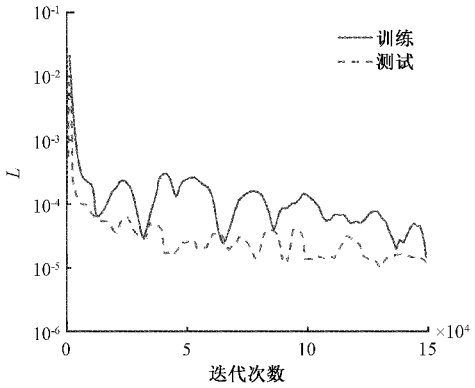

图 2　训练和测试过程中 L 的变化情况

该方法对于点声源具有较高的亚网格精度。在图 3(a)中，对于较大的亥姆霍兹值，距离误差的均值略有增大。然而，对于较小的亥姆霍兹数范围，定位的性能最佳，标准差在频率范围内保持不变。此外，水平误差均值也非常小（$\bar{e}_{level} \approx 0.002$ dB）。同时，在整个测试频段，水平误差较为平坦。图中高频部分的误差以及标准差有增加的趋势，但并不具备统计学显著特征。因此，准确率与频率无关。相较于其他波束形成算法在单一点源情况下具有特定水平误差[14]而言，本文所述的方法可以得到更为准确的水平估计，尤其是在赫尔蒙特数较大时。

训练结果的有效性可以通过设定不同的交叉验证数据来检验。检验结果如图 4 所示。图中灰色十字表示模拟点

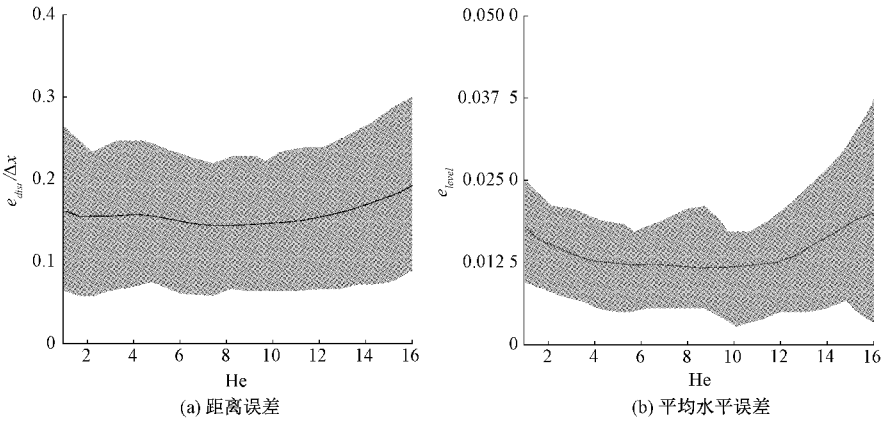

图 3　不同亥姆霍兹常数下神经网络误差

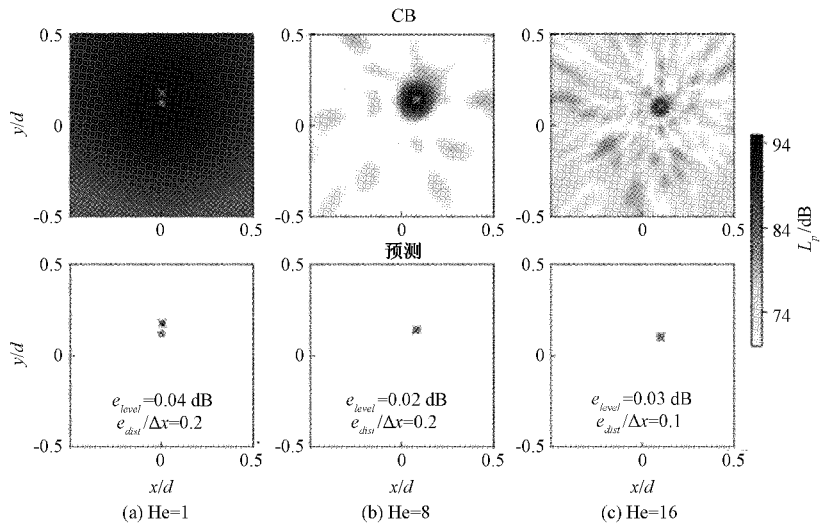

图 4　He=1、He=8 和 He=16 时测试数据集的 CB 图及预测结果

源的真实位置，灰色星形表示图中的最大值，黑点表示预测的点源位置；预测图中标注了水平和距离误差。图 4(a)为当 He=1 时的 CB 图及模型估计值；在 CB 图中，出现了宽阔的主瓣，其中最大值（灰星）与模拟点源的真实位置（灰十字）不重合。这种差异是由于在本研究中选择了用于波束成形的导向向量公式 III，但也有其他引导向量公式，允许将最大级别正确映射到 CB 图中的源位置，但这些公式会导致较大的波束成形源强度误差。图 4(a) 的预测图中黑点为模型估计的点源，其误差小于 1 dB，由此可见该训练模型可以精确表征点源。此外，点源的预测位置对应于其真实位置。由图 4(b)和(c)可知，该方法用于表征具有较高亥姆霍兹数的声源也同样适用；在 He=8 和

He=16 的 CB 图中,其最大声压级和真实源位置更接近,同时可以观察到更明显的旁瓣;即便如此,该模型仍可以非常精确的对点源进行预测。

3 结 论

本文将卷积神经网络的方法应用于常规波束形成图以表征点声源的位置属性。从理论角度阐述了基于残差神经网络定位单声源方位及强度的原理,并通过仿真实验验证了该方法的优越性。实验结果表明,该方法可以定位任意频率的声源,且精度高于常规波束形成图,且对于声源强度的估计可以取得较高的精度,这是以往研究中经常被忽略。因此,该方法可以有效地估计单声源的方位及强度。

由于本文所用训练及测试数据均为 Acoular 软件的模拟值,因此该方法对于真实场景中的测量数据的预测是否具有较高的精度还有待进一步验证。在未来的训练中,需要高变异度的训练数据来反映真实场景中出现的各种状况;训练数据同时还应该包含非均匀指向性的声源及其空间范围。

参考文献

[1] MERINO R, SIJTSMA P, SNELLEN M, et al. A review of acoustic imaging methods using phased microphone arrays[J]. CEAS Aeronautical Journal, 2019, 10(1): 197-230.

[2] 周晓彦,王珂,李凌燕. 基于深度学习的口标检测算法综述[J]. 电子测量技术, 2017,40(11): 89-93.

[3] 刘云,杨建滨,王传旭. 基于卷积神经网络的苹果缺陷检测算法[J].电子测量技术, 2017, 40(3):108-112.

[4] 王金玉,赵月娇,孔德健,等. 基于小波变换的 HVDC 系统故障检测[J]. 国外电子测量技术, 2017, 36(2): 37-40.

[5] 陈志强,陈旭东,JOSÉ VALENTE DE OLIVIRA,等. 深度学习在设备故障预测与健康管理中的应用[J]. 仪器仪表学报, 2019, 40(9): 206-226.

[6] REITER A, BELL M A. A machine learning approach to identifying point source locations in photoacoustic data [C]. Photons Plus Ultrasound: Imaging and Sensing, 2017,DOI:10.1117/12.2255098.

[7] KRIZGEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[8] ALLMAN D, REITER A, BELL M A, et al. Photoacoustic source detection and reflection artifact removal enabled by deep learning [J]. IEEE Transactions on Medical Imaging, 2018, 37(6): 1464-1477.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv: Computer Vision and Pattern Recognition, 2014: 1-14.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Neural Information Processing Systems, 2015: 91-99.

[11] LECUN Y, BOSER B E, DENKER J S, et al. Handwritten digit recognition with a back-propagation network[C]. Neural Information Processing Systems, 1989: 396-404.

[12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. ArXiv: Computer Vision and Pattern Recognition, 2015, DOI: 10.1109/CVPR.2016.90.

[13] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks [J]. ArXiv: Computer Vision and Pattern Recognition, 2016, DOI: 10.1007/978-3-319-46493-0_38.

[14] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-251.

[15] NASCIMENTO R, FRICKE K, VIANA F. A tutorial on solving ordinary differential equations using Python and hybrid physics-informed neural network [J]. Engineering Applications of Artificial Intelligence, 2020, 96: 1-11.

[16] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. ArXiv: Learning, 2015:1-10.

[17] SARRADJ E, HEROLD G. A Python framework for microphone array data processing [J]. Applied Acoustics, 2017: 50-58.

[18] RUMELHART D E, HINTON G E, WILLIMS R J, et al. Learning representations by back-propagating errors[J]. Nature, 1988, 323(6088): 696-699.

[19] BENGIO Y. Practical recommendations for gradient-based training of deep architectures [J]. ArXiv: Learning, 2012, DOI: 10.1007/978-3-642-35289-8_26.

[20] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. ArXiv: Learning, 2014:1-15.

[21] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012, 13(1): 281-305.

作者简介

王言彬,工学学士,高级工程师,主要研究方向为船舶建造、噪声与振动控制。

E-mail:wangyanbin313@163.com