协同优化效率与可靠性的递进式神经架构搜索方法

张 睿*,魏晓楠,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: 神经架构搜索作为自动化深度学习模型构建的核心技术,旨在搜索面向特定任务的最优网络结构,然而现有方法在以下方面存在不足:搜索空间中卷积与池化操作耦合,导致解空间冗余;高维编码采用整体优化方式,模块间协作不足,易陷入局部最优;评估策略依赖单一或少数指标,难以准确反映架构真实性能,易对搜索方向产生误导。上述问题相互交织,高潜力架构易被淘汰,搜索所得架构性能与最优值存在差距,限制了神经架构搜索(Neural Architecture Search, NAS)技术在资源受限及高可靠性场景中的应用。针对上述不足,本文提出一种递进式协同的神经架构搜索方法。该方法包含三个模块:在搜索空间层面,设计递进式表征降维可分离架构搜索空间,依据卷积与池化的功能特性将二者解耦,以缩小解空间规模并保留特征表征,为后续搜索与评估提供输入;在搜索策略层面,提出高维解耦式单元自适应搜索策略,将高维架构编码按卷积段、池化段、深度段进行分段解耦,对不同模块分别采用两点交叉、单点交叉及自适应变异等单元级定向遗传操作;在评估策略层面,构建多维低成本模型性能评估策略,从鲁棒性、各向同性、不确定性、规整性四个维度分别评估候选架构,并通过排名驱动的非线性几何融合机制将各维度分数整合为综合指标。上述三个模块依次衔接,形成递进式协同框架。实验结果表明,在 NAS-Bench-201 基准搜索空间上,所提方法在 CIFAR-10、CIFAR-100 及 ImageNet16-120 数据集上的肯德尔相关性系数分别为 0.712 0、0.705 2 与 0.698 1,其中在 CIFAR-10 上较 NASWOT 方法提升 20.13%。在工业焊缝缺陷检测与医疗 APTOS-2019 视网膜病变分级两项任务中,所得最优架构较现有无训练 NAS 方法平均精度分别提升约 18.67% 与 6.12%,搜索耗时分别为 227.97 s 与 559.34 s,推理延迟分别为 0.41 ms 与 0.46 ms。该研究为构建高效、低成本且具备跨领域泛化能力的自动化模型设计技术提供了参考。

关键词: 神经架构搜索;自动化模型设计;进化算法;递进式搜索;低成本评估

基金项目: 国家自然科学基金(No.62372319);教育部人文社会科学研究项目(No.23YJCZH299);山西省重点研发计划(No.202302140601012);山西省专利转化专项计划(No.20250021);山西省基础研究计划(No.202403021221142);山西省市场监督管理局科技计划(No.2025KJ019)

中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112(2026)04-1792-14

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260222

Progressive Neural Architecture Search Method for Collaborative Optimization of Efficiency and Reliability

ZHANG Rui*, WEI Xiaonan, SUN Chaoli
(*College of Computer Science and Technology*, *Taiyuan University of Science and Technology*, *Taiyuan*, *Shanxi* 030024, *China*)

Abstract: Neural architecture search (NAS), a core technology for automated deep learning model construction, aims to discover task-specific optimal network architectures. However, existing methods suffer from three critical limitations: coupled convolution and pooling operations in search spaces generate redundant solution spaces; holistically optimized high-dimensional encodings lack inter-module collaboration and easily fall into local optima; evaluation strategies relying on single or limited metrics fail to accurately reflect true architecture performance, biasing search directions. These intertwined issues eliminate high-potential architectures, create performance gaps between searched and optimal architectures, and restrict NAS deployment in resource-constrained and high-reliability scenarios. To address these deficiencies, this paper presents a progressive collaborative neural architecture search method. This method consists of three modules: a progressive representation dimensionality reduction separable architecture search space is designed at the search space level, where convolution and pooling are decoupled by their functional properties to reduce the solution space scale and retain feature representations, thus supplying input for subsequent search and evaluation. A high-dimensional decoupled unit-adaptive search strategy is proposed at the search strategy level, which decomposes high-dimensional architecture encodings into convolution segments, pooling segments and depth segments, and applies unit-level directed genetic operations including two-point

收稿日期:2026-03-13;录用日期:2026-04-08;
*通信作者:张睿

crossover, single-point crossover and adaptive mutation to different modules separately. A multi-dimensional low-cost model performance evaluation strategy is constructed at the evaluation strategy level to assess candidate architectures from four dimensions, namely robustness, isotropy, uncertainty and regularity, and merges multi-dimensional scores into a comprehensive indicator through a rank-driven nonlinear geometric fusion mechanism. The three modules connect sequentially to form a progressive collaborative framework. Experimental results demonstrate that on the NAS-Bench-201 benchmark search space, the proposed method yields Kendall correlation coefficients of 0.712 0, 0.705 2 and 0.698 1 on CIFAR-10, CIFAR-100 and ImageNet16-120 datasets respectively, with a 20.13% relative gain over NASWOT on CIFAR-10. For industrial weld defect detection and APTOS-2019 medical retinopathy grading, the derived optimal architecture boosts average accuracy by approximately 18.67% and 6.12% respectively over existing training-free NAS methods including NASWOT, ZiCo and MSTF-NAS, with search time of 227.97 s and 559.34 s, and inference latency of 0.41 ms and 0.46 ms correspondingly. This work provides a reference for developing efficient, low-cost automated model design techniques with strong cross-domain generalization.

Keywords: neural architecture search; automated model design; evolutionary algorithm; progressive search; low-cost evaluation

Foundation Item(s): National Natural Science Foundation of China (No. 62372319); Humanities and Social Sciences Research Project of the Ministry of Education (No. 23YJCZH299); Key Research and Development Program of Shanxi Province (No. 202302140601012); Special Program for Patent Transformation of Shanxi Province (No. 20250021); Basic Research Program of Shanxi Province (No.202403021221142); Science and Technology Program of Shanxi Administration for Market Regulation (No.2025KJ019)

0　引言

随着深度学习技术的迅猛发展,兼具低延迟与高鲁棒性的自动化模型构建技术,已成为电子信息领域的重要研究方向之一[1],神经架构搜索(Neural Architecture Search,NAS)作为深度学习模型的自动化设计范式,可根据预定义的搜索空间搜索出特定任务下的优秀架构。

在NAS的众多实现方式当中,基于进化计算的NAS方法,因其兼具非可微问题的求解能力以及复杂解空间的全局探索能力,受到了学术界的广泛关注[2]。Liang等人[3]对进化计算当中的种群生成策略进行了优化调整,并且利用基于随机森林的后代选择机制,依靠降本增效的核心设计理念取得了阶段性进展。Zou等人[4]设计了一种两阶段进化单元的基架构搜索方法,采用预先确定网络框架再改动单元结构的策略,在多个图像分类数据集中取得了良好的效果。蒋鹏程等人[5]提出了一种基于新型损失函数训练的多层感知机(Multi-Layer Perceptrons,MLPs)得分预测器,用于取代种群个体评估时的真实训练过程,并设计了两阶段演化的搜索策略,最终在多个数据集上证实了所提方法的有效性。

尽管上述对进化计算进行改进的NAS方法在特定任务当中取得了良好的效果,但其基于真实训练的评估方法造成了效率的先天瓶颈。为此,研究者们开始提出基于无训练的NAS方法,引发了广泛关注[6]。Lopes等人[7]提出了一种低约束的宏神经架构搜索方法(Less Constrained Macro-NAS,LCMNAS),该方法凭借现有的卷积神经网络(Convolutional Neural Network,CNN)信息产生加权有向图搜索空间,并通过进化计算生成高质量的候选架构,最后利用初始化阶段的架构信息估算候选架构性能,提升了搜索策略的效率。Onzo等人[8]将无训练评估方法与进化算法相结合,通过基于随机梯度和动量的代理模型作为回归预测器来评估候选架构,加快了NAS的收敛速度。张睿等人[9]提出了可分离的复数搜索空间和自适应全局-局部协同的搜索策略,通过构建能最大化平衡特征差异性和准确率映射关系的度量矩阵选择出语音增强任务中的最优架构。Wu等人[10]利用搜索空间当中不同区域的期望值,引导搜索策略探索高潜力区域,并且将加噪输入和原始输入的相关性进行关联,从而对候选架构的最终性能进行快速预测,使得在大幅降低计算效率的同时,高效得出了最优架构。Yamasaki等人[11]基于不同输入图像下候选架构的深层特征与输入图像的相似度进行量化表征,对候选架构性能进行评估,实现了低成本评估候选架构性能的需求。

从上述分析可知,现有基于无训练评估的NAS方法仍存在许多有待解决的问题。首先,这类方法的搜索空间设计通常为卷积与池化耦合的方式,在模型中不区分两者的作用和功能,这样会导致候选架构的解空间冗余,从而增大了搜索策略的寻优成本;其次,现有无训练评估方法的性能评估主要是通过少数几个维度甚至单个维度来预测模型性能,导致候选架构评分严重偏离了模型真实性能,进而对搜索策略寻优

方向产生误导;最后,现有主流搜索策略均将搜索空间的高维编码作为一个整体进行搜索,而忽略了编码中不同模块的协作和依赖关系,一方面会造成算法优化效率低下,另一方面容易造成算法陷入局部最优。

为此,本文提出一种协同优化效率与可靠性的递进式神经架构搜索方法(progressive NAS method for Collaborative Optimization of Efficiency and Reliability, COER-NAS),核心贡献如下:

(1)针对传统搜索空间中,卷积与池化操作紧密耦合产生的特征丢失、解空间冗余引发搜索策略低效的问题,本文提出递进式表征降维可分离架构搜索空间,通过对架构搜索空间进行功能上的分解以实现卷积层与池化层的解耦,为后续搜索策略和评估策略提供基本的架构支撑。

(2)针对高维进化搜索算法性能低效、解集多样性不足、无法适配搜索空间架构编码的问题,本文提出高维解耦式单元自适应搜索策略,通过基于精英保留和锦标赛选择的后代择优策略,结合编码分段解耦及定向遗传操作,提升了搜索过程的稳定性与效率,为评估策略提供了优秀的候选架构。

(3)针对现有评估策略无法全面反映候选架构真实性能对搜索策略优化方向产生误导的问题,本文设计了多维低成本评估策略,通过对架构的不同维度进行全方位评估,并对评估结果进行非线性融合以提高预测精度,从而对搜索策略进行精确引导。该策略与搜索空间、搜索策略形成协同的自动化架构搜索体系。

1 方法整体框架

针对传统NAS方法中存在的搜索空间操作耦合导致的解空间冗余、搜索策略效率低下引发的搜索耗时过长,以及评估策略预测精度不足造成的搜索方向偏差的问题,本文提出COER-NAS,其整体结构如图1所示。设计递进式表征降维可分离的架构搜索空间,基于基本操作的固有特性构建候选架构,缩小了解空间的大小,降低了搜索策略的优化复杂度。设计高维解耦式单元自适应架构搜索策略,基于搜索空间对候选架构的分块编码,执行单元级分块遗传操作,从而提升收敛速度。最后,设计多维低成本模型性能评估策略,通过非线性几何融合方法实现多维评估分数的有效整合,提升模型性能预测的准确性,进而引导搜索策略向最优方向精准收敛。三大模块递进协同,在兼顾效率与可靠性的前提下,得出最优的模型架构。

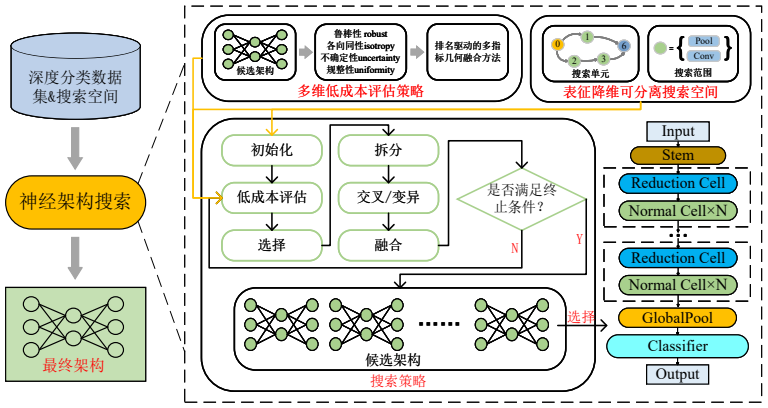

图1 方法整体框架示意图

Figure 1 Overall framework of the proposed method

2 本文方法

2.1 递进式表征降维可分离架构搜索空间

在传统的搜索空间设计当中,卷积与池化操作常被耦合于同一Cell单元内,这种未充分利用二者特性的耦合构建方式往往会造成解空间冗余,进而影响后续搜索策略的收敛效率。针对这一问题,本文提出递进式表征降维可分离的架构搜索空间。通过利用卷积的全局特征提取能力构建正常单元以及利用池化的局部信息保留能力构建降采样单元,实现了高效特征提取与信息保留,为高质量架构的构建奠定了基础。

如图2所示,输入图像经stem预处理模块后,进入包含三阶段的特征提取与挖掘流程($N_{stage}=3$),每个阶段首先通过降采样单元压缩特征图尺寸以保留关键信息,该单元定义了7种步幅为2的基本降采样操作(涵盖2×2、3×3、4×4最大池化、平均池化以及条形池化),节点数$N_{reduction}=4$,各节点对应一种基本操作并经并行处理后融合为输出特征图。随后,将采样后的特征图送入重复堆叠的特征提取单元,以挖掘当前尺度下的关键特征信息,该单元定义了10种基本操作(涵盖3×3、5×5、7×7深度可分离卷积,

3×3、5×5、7×7空洞卷积,并行的3×3、5×5、7×7多尺度卷积,两个3×3串行的深度可分离卷积,以及恒等操作与零操作),节点数$N_{\text{normal}}=5$,其中0号与6号节点分别为输入和输出节点,内部每个节点均需与序号更小的节点建立连接并选择一种运算操作。第三阶段输出的特征图经全局池化与分类器处理,输出最终的类别预测。

对于某个特定的特征提取阶段,其编码长度为$2\times N_{\text{normal}}+N_{\text{reduction}}$,值得注意的是,特征提取单元需要同时保存每个节点的基本操作和其前驱节点的序号;整个架构的基因型长度为$N_{\text{stage}}\times(2\times N_{\text{normal}}+N_{\text{reduction}})+N_{\text{stage}}$,在本文的参数设定下,编码长度为45位,最后三位的编码为三阶段中特征提取单元的堆叠次数。

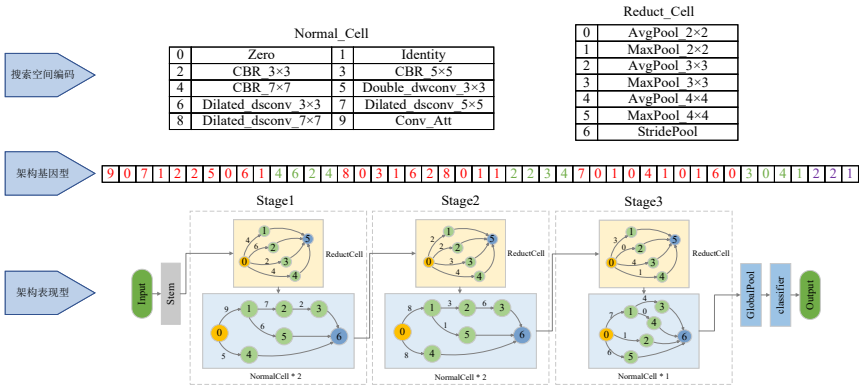

图2　递进式表征降维可分离架构搜索空间

Figure 2　Progressive representation dimensionality reduction separable architecture search space

2.2　高维解耦式单元自适应架构搜索策略

进化搜索(Evolutionary Search, ES)[12]因其具备强大的全局优化能力,长期以来一直是NAS领域主流的搜索策略构建方式之一。然而,若直接将ES运用到第2.1节进行全局寻优时,将会造成搜索速度缓慢且易陷入局部最优的问题。为此,本文提出高维解耦式单元自适应架构搜索策略(High-Dimensional Decoupled Unit-Adaptive Architecture Search strategy, HD-DUAAS),其核心思想是将第2.1节的高维搜索空间编码进行分块解耦,进行单元级定向遗传,从而提升收敛速度。

种群初始化与父代选择策略。针对传统ES随机初始化导致的初始种群分布不均、有效个体占比偏低的问题,本文引入了佳点集初始化策略[13]生成初始种群P_{curr},提高了初始种群在高维空间当中的分布均匀性。在遗传操作的父代选择中,考虑到单一选择机制易引发种群多样性流失问题,本文设计了融合精英筛选与锦标赛[14]机制的父代选择方案。具体而言,首先根据预设的比例筛选种群当中适应度最佳的精英个体P_{elite};随后,将种群按照预先设定的数量划分为若干小组进行锦标赛选择,选出每个小组当中适应度最优的个体组成集合P_{tour};最后,对两类个体进行合并去重后,构建兼具高适应度与多样性的父代集合P_{parent},保障优质个体在遗传操作当中的优先次序。

高维解耦式单元级分段遗传操作。为解决对高维架构编码直接进行全局遗传操作所面临的模块完整性被破坏及搜索效率低下的问题,本文提出了高维解耦式单元级分段遗传操作,其核心执行流程如算法1所示。具体而言,首先,本文将第2.1节定义的45维架构编码按照卷积段、池化段、深度段进行解耦,根据不同的编码结构,实施针对性的遗传操作。对于卷积与池化段而言,按照单元粒度从架构编码中对其进行拆解,奇数位定义了卷积类型,而偶数位定义了节点连接。对于卷积和池化段的编码,分别采用单元级两点交叉操作与单元级单点交叉操作对其进行迭代,并实施边界受限的变异操作。对于阶段深度段而言,本文主要对精英配对个体进行低概率的一点交叉操作,保证其全局拓扑结构的稳定性。另外,采用平台期自适应变异策略,当算法陷入平台期时,合理提高变异概率以增强全局探索能力。最后,将各个模块段重新组合构成完整的架构编码。本策略的使用,在维持模型向最优架构收敛的同时,有助于提高方法的搜索效率。

种群更新机制。在完成单元级分块遗传操作生成子代集合Q之后,HD-DUAAS首先剔除超出预设边界的无效个体,计算有效子代的适应度值;随后,将其与父代种群P_{curr}合并去重,进入新一代种群P_{next}的构建阶段。首先,算法会强制保留历史的全局最优个体G_{best},以避免优质架构的丢失;随后在剩余的个体选择上,以80%的比例选取适应度最佳的优秀个体保证算法的收敛性,以20%的比例随机抽取个体以维持种群的多样性。以此反复,直到算法达到预先设

定的最大评估次数或连续多轮迭代当中适应度提升小于阈值T_{thres}，且平台期计数器达到了耐心值Pat时停止算法迭代，输出当前的全局最优架构G_{best}。

算法1 高维解耦单元级分段子代生成算法

输入：Pat(平台期耐心值)，PlatCnt(平台期计数器)，P_{elite}(精英个体集合)，P_{tour}(锦标赛选择个体)，$P_{\text{cross}}^{(\text{conv})}$、$P_{\text{cross}}^{(\text{pool})}$、$P_{\text{cross}}^{(\text{depth})}$(卷积/池化/深度段交叉概率)，$P_{\text{mut}}^{(\text{conv})}$、$P_{\text{mut}}^{(\text{pool})}$、$P_{\text{mut}}^{(\text{depth})}$(卷积/池化/深度段变异概率)

输出：Q(当前阶段的遗传结果/子代集合)

$P_{\text{parent}}=P_{\text{elite}}\cup P_{\text{tour}}$;

$Q=\varnothing$;

FOR $(\boldsymbol{p}_1,\boldsymbol{p}_2)\in P_{\text{parent}}$ **DO** //$\boldsymbol{p}_1$,$\boldsymbol{p}_2$为两两不重复的随机配对

$(\boldsymbol{C}_1,\boldsymbol{P}_1,\boldsymbol{D}_1)=\text{split}(\boldsymbol{p}_1)$; $(\boldsymbol{C}_2,\boldsymbol{P}_2,\boldsymbol{D}_2)=\text{split}(\boldsymbol{p}_2)$;

$\boldsymbol{C}'_1=\boldsymbol{C}_1$; $\boldsymbol{C}'_2=\boldsymbol{C}_2$; $\boldsymbol{P}'_1=\boldsymbol{P}_1$; $\boldsymbol{P}'_2=\boldsymbol{P}_2$; $\boldsymbol{D}'_1=\boldsymbol{D}_1$; $\boldsymbol{D}'_2=\boldsymbol{D}_2$;

IF $\text{Random}(0,1)<P_{\text{cross}}^{(\text{conv})}$ **THEN** $\boldsymbol{C}'_1,\boldsymbol{C}'_2=\text{TwoPointCross}(\boldsymbol{C}_1,\boldsymbol{C}_2)$;

END IF //卷积段两点交叉与变异

$\boldsymbol{C}'_1=\text{Mutate}(\boldsymbol{C}'_1,P_{\text{mut}}^{(\text{conv})})$; $\boldsymbol{C}'_2=\text{Mutate}(\boldsymbol{C}'_2,P_{\text{mut}}^{(\text{conv})})$;

IF $\text{Random}(0,1)<P_{\text{cross}}^{(\text{pool})}$ **THEN** $\boldsymbol{P}'_1,\boldsymbol{P}'_2=\text{OnePointCross}(\boldsymbol{P}_1,\boldsymbol{P}_2)$;

END IF //池化段单点交叉与变异

$\boldsymbol{P}'_1=\text{Mutate}(\boldsymbol{P}'_1,P_{\text{mut}}^{(\text{pool})})$; $\boldsymbol{P}'_2=\text{Mutate}(\boldsymbol{P}'_2,P_{\text{mut}}^{(\text{pool})})$;

IF $\boldsymbol{p}_1\in P_{\text{elite}}\wedge\boldsymbol{p}_2\in P_{\text{elite}}\wedge\text{Random}(0,1)<P_{\text{cross}}^{(\text{depth})}$ **THEN** //深度段精英交叉与自适应变异

$\boldsymbol{D}'_1,\boldsymbol{D}'_2=\text{SinglePointCross}(\boldsymbol{D}_1,\boldsymbol{D}_2)$;

END IF

IF $\text{PlatCnt}\geqslant\text{Pat}/2$ **THEN**

$\boldsymbol{D}'_1=\text{Mutate}\left(\boldsymbol{D}'_1,\min(P_{\text{mut}}^{(\text{depth})}+0.01\times\text{PlatCnt},1)\right)$; $\boldsymbol{D}'_2=\text{Mutate}\left(\boldsymbol{D}'_2,\min(P_{\text{mut}}^{(\text{depth})}+0.01\times\text{PlatCnt},1)\right)$;

END IF

$\boldsymbol{c}_1=\text{Merge}(\boldsymbol{C}'_1,\boldsymbol{P}'_1,\boldsymbol{D}'_1)$; $\boldsymbol{c}_2=\text{Merge}(\boldsymbol{C}'_2,\boldsymbol{P}'_2,\boldsymbol{D}'_2)$;

$Q=Q\cup\{\boldsymbol{c}_1,\boldsymbol{c}_2\}$;

END FOR

RETURN Q

2.3 多维低成本模型性能评估策略

在NAS中，候选架构的性能评估需要在尽可能低的计算开销下进行，性能评估的准确性会关系到搜索策略的搜索效率以及最终得到的架构性能。目前，现有的评估策略大多是采用单一或者少数代理指标进行评估，难以全面量化架构的综合性能。为此，本文提出一种多维低成本模型性能评估策略，从扰动鲁棒性、特征各向同性、参数-梯度不确定性及梯度-层结构规整性四个方面量化架构的综合性能，并通过排名驱动的非线性几何机制将其融合为最终的评估指标，为搜索策略提供有效且可靠的搜索指引。四维指标相互协同共同搭建起层次化的评估体系：扰动鲁棒性体现模型的泛化能力，是架构适应真实场景的核心基础；特征各向同性量化模型的特征表达能力，决定模型表征复杂关系的质量；参数-梯度不确定性对模型的训练稳定性起到保障作用，是架构易训练的基础；梯度-层结构规整性细化结构适配性评估，为架构优化提供有效指引。四大维度层层递进、相互协同，构建起低成本、全方位的评估体系。

2.3.1 模型扰动鲁棒性评估

本维度聚焦于模型实际部署时的稳定性，其核心原理为：有着较强鲁棒性的架构，其深层特征的分布不会因输入的小幅度扰动而有明显的变化。根据该原理，本文将原始数据与扰动后的数据输入至模型当中，将二者深层特征分布的差异度进行分析，以此作为模型鲁棒性的量化指标，下面将详细阐述该指标的量化方法。

具体而言，本文通过对图像进行中心裁剪来模拟实际场景中的信息丢失问题，生成带有扰动的输入$\boldsymbol{x}_{\text{noisy}}$，如下式：

$$\boldsymbol{x}_{\text{noisy}}=\text{Resize}\left(\text{CenterCrop}(x,0.8*H,0.8*W),(H,W)\right)\tag{1}$$

其中，CenterCrop(·)为中心裁剪函数；Resize(·)为图像缩放函数；输入张量$\boldsymbol{x}\in\mathbf{R}^{B\times C\times H\times W}$表示输入的单批次图像，维度$B$、$C$、$H$、$W$分别表示数据的批次大小、通道数、图像高度与宽度。在模型性能评估过程中，每个训练批次均由从数据集中随机采样的128张图像组成，以此确保数据能够充分反映数据集整体的分布特征。

随后，为消除量纲差异且进一步增强对特征分布的敏感度，本文基于秩相关系数来衡量原始输入图像与$\boldsymbol{x}_{\text{noisy}}$的分布一致性，将每个样本的特征向量按照元素值排序后赋予秩系数，得到秩矩阵$\boldsymbol{R}_{\text{orig}}\in\mathbf{R}^{B\times D}$与$\boldsymbol{R}_{\text{noise}}\in\mathbf{R}^{B\times D}$。在此基础上，进一步计算一个批次的平均秩相关系数$\rho$，如下式：

$$\rho=\frac{1}{B}\sum_{b=1}^{B}\frac{\sum_{d=1}^{D}\left(\boldsymbol{R}_{\text{orig}}^{(b,d)}-\overline{\boldsymbol{R}}_{\text{orig}}^{(b)}\right)\cdot\left(\boldsymbol{R}_{\text{noisy}}^{(b,d)}-\overline{\boldsymbol{R}}_{\text{noisy}}^{(b)}\right)}{\sqrt{\sum_{d=1}^{D}\left(\boldsymbol{R}_{\text{orig}}^{(b,d)}-\overline{\boldsymbol{R}}_{\text{orig}}^{(b)}\right)^2\cdot\sum_{d=1}^{D}\left(\boldsymbol{R}_{\text{noisy}}^{(b,d)}-\overline{\boldsymbol{R}}_{\text{noisy}}^{(b)}\right)^2+10^{-8}}}\tag{2}$$

其中，$\overline{\boldsymbol{R}}_{\text{orig}}^{(b)}$和$\overline{\boldsymbol{R}}_{\text{noisy}}^{(b)}$分别代表第$b$个样本秩向量的均值。最后，通过指数映射函数将特征差异度转化至0到100的范围之内，作为扰动鲁棒性S_{robust}的最终的评估分数，其数值公式如下：

$$S_{\text{robust}}=100\cdot\text{e}^{-15\cdot(1-\rho)}\tag{3}$$

2.3.2 模型特征空间各向同性评估

本维度聚焦于模型的特征表达质量，其核心原理是基于特征空间的各向同性[15]：表达能力强的网络，其隐含层特征应该分散在不同的维度中，避免信息冗余。基于该原理，本文通过分析特征协方差矩阵的主

成分分布情况对其实现量化:若特征均匀分布且熵值偏高,表明特征空间的各向同性越明显;反之,则说明特征空间各向同性较弱。下文将详细说明量化指标的具体计算流程。

具体而言,给定单批次的输入图像 $\boldsymbol{x}\in\mathbf{R}^{B\times C\times H\times W}$(与第2.3.1节采样方式相同),依靠前向传播并注册钩子,采集网络中前 L 个主块的输出特征张量。针对第 l 个主块,将其输出特征 $\boldsymbol{f}_l\in\mathbf{R}^{B\times c_l\times h_l\times w_l}$($c_l$ 为通道数,h_l 与 w_l 为特征图的高和宽)重塑为特征矩阵 $\boldsymbol{F}_l\in\mathbf{R}^{c_l\times n_l}$($n_l=B\times h_l\times w_l$ 为特征样本总数);接着对特征矩阵 $\boldsymbol{F}_l$ 进行中心化处理,得到中心化特征矩阵 $\tilde{\boldsymbol{F}}_l$;进而通过添加正则化参数 $\delta=10^{-8}$ 计算正则化协方差矩阵,以避免因样本匮乏导致的矩阵奇异问题。在此基础上,对实对称协方差矩阵 $\boldsymbol{V}_l$ 执行特征值分解,得到特征值 $\lambda_l^{(1)},\lambda_l^{(2)},\cdots,\lambda_1^{(c_l)}$。随后设定阈值 $\varepsilon=10^{-10}$ 对特征值进行非负调整与L1归一化,得到主成分概率分布 $\tilde{\lambda}_l^{(i)}$,其数学表达式如下式:

$$\tilde{\lambda}_l^{(i)}=\frac{\max\left(\lambda_l^{(i)},\varepsilon\right)}{\sum_{j=1}^{c_l}\max\left(\lambda_l^{(j)},\varepsilon\right)},i=1,2,\cdots,c_l \tag{4}$$

进一步地,基于该分布计算第 l 个主块的特征表达熵 s_l^e,以量化该主块特征分布的分散度,其计算方式如式(5)所示。最后,将全部 L 个主块的特征表达熵进行累加,得到架构的整体各向同性评分 S_{isotropy},其计算方式如式(6)所示。

$$s_l^e=-\sum_{i=1}^{c_l}\tilde{\lambda}_l^{(i)}\cdot\log\left(\tilde{\lambda}_l^{(i)}+\varepsilon\right) \tag{5}$$

$$S_{\text{isotropy}}=\sum_{l=1}^{L}s_l^e \tag{6}$$

2.3.3　模型参数-梯度联合不确定性评估

本维度聚焦于模型的参数复杂度和效能,其核心原理为:对于性能优良的架构,其参数空间应具备更高的信息密度,且不同位置参数的重要性不同。基于该原理,本文利用奇异值分解量化每层权重的有效值占比,以此来衡量信息密度,并且以自定义加权的方法来量化不同参数的重要性。该维度和第2.3.4节相互呼应,二者共同搭建起对参数初始效能的多维评估,具体评估的流程如下。

具体而言,本文提出的多因素加权策略以各层参数展平后的二维矩阵 $\boldsymbol{W}$ 为计算基础,通过融合四个维度的关键权重因子,量化得到各层的综合重要性权重。其中,基础规模权重 $I_{\text{base}}=\sqrt{\text{numel}(\boldsymbol{W})}$($\text{numel}(\boldsymbol{W})$ 为参数矩阵 $\boldsymbol{W}$ 的元素总数),通过参数总量表征模型的计算复杂度,是重要性评估的基础维度;网络深度系数 $\alpha_{\text{depth}}=1+\frac{b}{5\cdot\max(1,B)}$(其中 b 为层所在主块的索引,B 为主块的总数量),用于建模层在梯度传播路径中的位置,网络层级越深,对模型输出的贡献越显著;模块堆叠系数 $\alpha_{\text{stack}}=1+0.05\cdot\max(0,t_{\text{stack}}-1)$($t_{\text{stack}}$ 为正常单元堆叠次数),用以表征模块堆叠时的特征累积效应,堆叠次数越多,权重越高;功能角色权重 α_{role} 依据不同层的类型对网络功能的贡献度进行赋值:降采样层负责判别性信息的保留,对最终模型的性能与效率具有重要影响,赋予最高权重1.60;特征提取层以提取特征为主要任务,分配权重1.35,高于基准值;预处理层用于确保输入数据的稳定性,分配权重1.10,略高于基准;分类层很大程度上依赖于前序特征输出,权重相对较低,设定为0.90;全局池化层和其他层仅为信息交换的桥梁,分配基准权重1.00。最终,通过式(7)对上述四类权重因子进行加权融合,进而获得各层的综合重要性权重。

$$I=I_{\text{base}}\cdot\alpha_{\text{role}}\cdot\alpha_{\text{depth}}\cdot\alpha_{\text{stack}} \tag{7}$$

针对不同尺度的权值矩阵,本文首先对其进行奇异值分解,得到奇异值序列 $\{\sigma_i\}$。随后,本文设定累计能量阈值 $\tau=0.95$,有效秩 R_{eff} 定义为满足 $\frac{\sum_{i=1}^{R_{\text{eff}}}\sigma_i}{\sum_{i=1}^{R_{\text{full}}}\sigma_i}\geqslant\tau$ 的最小正整数(其中,$R_{\text{full}}=\min(n_{\text{rows}},n_{\text{cols}})$),若 $\|\boldsymbol{W}\|<10^{-10}$,该层初始化阶段信息贡献较弱,设定该层有效秩比为 $\rho=0$,否则,$\rho=\frac{R_{\text{eff}}}{R_{\text{full}}}$。最后,本文以各层综合重要性权重 I 为加权系数,将有效层的秩比 ρ 进行加权平均,进而得到该模型参数梯度的联合不确定性的评分 $S_{\text{uncertainty}}$,具体数学表达如式(8)所示。

$$S_{\text{uncertainty}}=100\times\frac{\sum_{\text{layers}}I\cdot\rho}{\sum_{\text{layers}}I} \tag{8}$$

2.3.4　模型梯度-层结构规整性评估

本维度聚焦于随机输入下各层梯度的统计分布规律,该维度与第2.3.3节的不确定性互为补充,共同对模型的复杂度进行量化表征。其具体原理为:有着良好协调性的架构,其各层梯度分布应保持相对平稳。根据该原理,本文量化了模型梯度更新方向与参数初始化符号的契合程度,随后对各层级参数的不确定性分布进行了拟合,最后结合层级结构的相关属性为熵值分配权重,实现了对候选架构复杂性的综合评估,具体数学建模过程如下。

具体而言,为更加稳定地获取模型的统计特征,本文生成了 N($N=5$)组符合标准正态分布的输入样

本$\boldsymbol{X}$以及对应的目标标签$\boldsymbol{T}$,随后将生成的数据输入到待评估模型中进行反向传播,以此来获取候选架构中可训练参数的梯度$\nabla_{\boldsymbol{W}}L$(L为损失函数的损失值)。对于不同参数$\boldsymbol{W}$及其梯度$g=\nabla_{\boldsymbol{W}}L$,首先需要计算其方向感知不确定性$U_{\text{sign}}$,具体计算的表达式如式(9)所示。其中,$\eta$表示学习率,$\tau=10^{-9}$表示平滑因子。

$$U_{\text{sign}}=\frac{|g|\cdot\eta}{|\boldsymbol{W}|+\tau}\cdot\text{sign}(g)\cdot\text{sign}(\boldsymbol{W}) \tag{9}$$

在此基础上,将N次采样所得到的有效数据U_{sign}合并为向量$\boldsymbol{V}_{\text{valid}}$(去掉绝对值小于$10^{-9}$的近似零值)后,计算其绝对值序列的均值$\mu$及标准差$\sigma$。若$\sigma>10^{-9}$,采用高斯分布对每个元素绝对值$|v_i|$的概率分布进行拟合;若$\sigma\leqslant10^{-9}$,表示分布高度集中,采用均匀分布进行拟合。随后,按照拟合后的概率分布$\{p_i\}$,计算该层的梯度熵H_{layer},具体如下:

$$H_{\text{layer}}=-\sum_i p_i\cdot\log(p_i+\tau) \tag{10}$$

此外,为进一步量化不同类型层对模型训练过程中的差异化影响,本文引入了角色倍率ω_{role}(数值同2.3.3节)、深度倍率$\omega_{\text{depth}}=1.0+\frac{b}{5B}$(其中$b$是指层所在主块的索引,而$B$为主块的总数)以及堆叠倍率$\omega_{\text{stack}}=1.0+0.05\cdot\max(0,s-1)$(其中$s$为Normal块的堆叠数量)对模型当中的不同层$H_{\text{layer}}$进行加权,得到经加权后的层级熵值$H_{\text{weighted}}$,其数学表达式如式(11)所示。

$$H_{\text{weighted}}=H_{\text{layer}}\cdot\omega_{\text{role}}\cdot\omega_{\text{depth}}\cdot\omega_{\text{stack}} \tag{11}$$

最后,对所有有效层($\boldsymbol{V}_{\text{valid}}$长度$\geqslant10$)的加权熵进行求和得到总加权熵$H_{\text{total}}=\sum H_{\text{weighted}}$,并进一步计算出平均加权熵$\hat{H}=\frac{H_{\text{total}}}{M}$(其中$M$为有效层的数量)。若$M$的值为0,则直接返回评分$S_{\text{uniformity}}=0$;否则,则根据式(12)计算中间得分$S$,并将该中间得分归一化至区间[0, 100],作为最终的模型规整度分数$S_{\text{uniformity}}$。

$$S=\ln(1+\hat{H}) \tag{12}$$

2.3.5 排名驱动的多指标几何融合方法

在候选架构性能评定的过程中,对多指标进行科学融合是筛选优秀架构的关键。然而,多数现有的融合方式均运用线性融合模式,这易受量纲差异性的影响,难以对模型性能开展精确且全面的量化。为此,本文提出排名驱动的多指标几何融合方法。该方法首先将各维度分数映射至具有统一尺度的相对排名,以此消除量纲异质性所造成的负面作用,随后借助几何平均完成多维指标的融合。该方法可为候选架构的筛选提供更科学、准确且可靠的综合性能评估依据,具体的量化过程如下。

具体而言,首先,针对m个待评估候选架构和k个代理评估指标(本文$k=4$),将各架构在各指标下的原始评分整合为$m\times k$维得分矩阵$\boldsymbol{S}\in\mathbf{R}^{m\times k}$。随后,针对第$j$个指标的得分向量$\boldsymbol{s}_j=[s_{1,j},s_{2,j},\cdots,s_{m,j}]^{\text{T}}$,采用稠密排名法按照升序方式对其进行排列,并为其中各个元素赋予对应的升序秩$r_{i,j}$,再通过归一化处理得到秩比$p_{i,j}$,形成$m\times k$维的秩比矩阵$\boldsymbol{P}\in\mathbf{R}^{m\times k}$。最后,对于第$i$个候选架构的秩比行向量,通过计算其各元素乘积的$k$次方根的方式得到最终融合得分$S(i)$,其数学表达式如式(13)所示。

$$S(i)=100\times\left(\prod_{j=1}^{k}P(i,j)\right)^{\frac{1}{k}} \tag{13}$$

3 实验验证与结果分析

本节进行了全面的实验,以验证所提协同优化效率与可靠性的递进式神经架构搜索方法的有效性。首先,本文验证了所提多维低成本模型性能评估策略的预测准确性。随后,在自建焊缝缺陷检测、公开APTOS-2019[16]两个跨领域数据集上,与主流手动设计模型及代表性的无训练NAS方法进行全方位性能及效率对比,验证了本文方法在实际应用场景下的优越性与跨领域泛化性。此外,本文所有实验均在统一的软硬件环境下完成,以确保实验过程的公平性与实验结果的可复现性,硬件环境配置为14th Gen Intel(R) Core(TM) i5-14600KF @ 3.50 GHz处理器、32 GB内存、单张NVIDIA GeForce RTX 4070 Ti SUPER独立显卡,软件环境配置为Python 3.7、PyTorch 1.13.0深度学习框架、CUDA 11.6。

3.1 低成本模型性能评估策略性能验证

为验证所提多维低成本模型性能评估策略的有效性、优越性与鲁棒性,本节在NAS基准搜索空间上,从预测相关性、采样鲁棒性、指标贡献度及应用效果四个维度,开展了系统性的实验与分析。

3.1.1 与先进方法的定量对比

为验证所提低成本评估策略的先进性,本文在采样数量为1 000个的基础上,将所提策略与现有先进的方法进行跨数据集横向对比,具体对比结果如表1所示。值得注意的是,本文描述的相关性系数,均表征所提策略的预测分数与模型真实性能间的相关性。实验结果表明,本文所提策略在NAS-Bench-201[17]基准搜索空间中,肯德尔相关性系数τ与斯皮尔曼相关性系数ρ均全面超越所有对比方法。在CIFAR-10上,COER-NAS的τ值较次优方法NASWOT提升0.119 3(提升20.1%),ρ值提升0.113;在CIFAR-100上,τ值与

ρ 值分别较 NASWOT 高出 0.068 和 0.065;在更具挑战性的 ImageNet16-120 上,τ 值与 ρ 值仍较 NASWOT 分别提升 0.086 4 和 0.074。上述结果证实,本文设计的非线性评估策略能有效捕捉架构性能关联特征,从而在低成本模型性能评估中展现出显著的先进性与有效性。

上述定量对比结果证明了所提评估策略的竞争力。此外,为进一步验证本文方法的稳定性,本文开展了采样鲁棒性实验(结果见图 3)。结果表明,在 10、100、500、1 000、2 000、3 000 这六种不同的采样规模的情况下,本文策略都能使肯德尔相关性系数维持较高且平稳的水平,进一步验证了其在资源受限等复杂 NAS 场景中的应用价值。

表 1　不同低成本评估方法的对比结果表

Table 1　Comparison results of different low-cost evaluation methods

Method	Year	CIFAR-10		CIFAR-100		ImageNet16-120	
		τ	ρ	τ	ρ	τ	ρ
#Params	—	0.545 9	0.720 9	0.566 9	0.737 8	0.536 3	0.704 1
#Flops	—	0.505 8	0.695 2	0.529 3	0.716 3	0.500 5	0.684 0
Snip[18]	2018	0.437 4	0.599 5	0.494 4	0.657 1	0.435 2	0.577 7
Grasp[19]	2020	0.216 6	0.318 2	0.199 0	0.287 4	0.270 1	0.395 3
SynFlow[20]	2020	0.546 5	0.740 2	0.578 3	0.771 2	0.569 1	0.755 5
GradNorm[21]	2021	0.348 7	0.466 3	0.382 3	0.517 5	0.324 2	0.437 2
NASWOT[22]	2021	0.592 7	0.778 7	0.637 0	0.820 3	0.611 7	0.792 3
GradSign[23]	2021	0.138 2	0.209 6	0.140 2	0.205 9	0.098 9	0.137 7
ZiCo[24]	2023	0.550 2	0.743 3	0.589 7	0.780 9	0.592 3	0.783 0
LCMNAS[7]	2023	0.544 4	0.705 0	0.532 5	0.692 7	0.496 2	0.645 5
ePADS(σ=0.5)[25]	2025	0.526 9	0.713 9	0.564 3	0.753 7	0.547 6	0.737 0
COER-NAS	2026	0.712 0	0.891 2	0.705 2	0.885 5	0.698 1	0.866 3

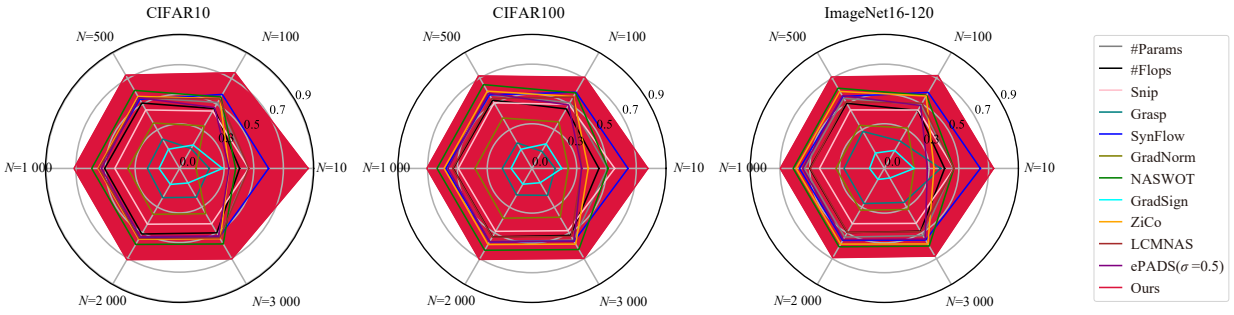

图 3　不同数据集及采样数下评估方法相关性雷达图

Figure 3　Radar chart of correlation among evaluation metrics across datasets and sampling sizes

3.1.2　多指标贡献度消融研究

为直观展示不同的评估指标对模型性能评估的实际贡献大小,本节在统一采样 1 000 个架构的条件下,进行了系统性的消融实验,结果如表 2 所示。表中融合方式(Fusion Method,FM)列表征不同代理指标的聚合类型,其中 L 代表线性融合(Linear),NL 代表非线性融合(Non-Linear);其余列依次为是否含有对应的代理指标与三个数据集当中的肯德尔相关性系数 τ。实验结果表明,单一代理指标仅能提供有限的架构性能评估信息,在 CIFAR-10 数据集上对应的 τ 值最高仅为 0.577 6,难以全面刻画网络架构的真实性能。随着代理指标数量逐步增加,评估性能呈现阶梯式提升,两指标与三指标组合在 CIFAR-10 上的 τ 值可分别达到 0.617 8 与 0.687 0 的最优值,充分验证了不同代理指标间的互补特性。在采用全部四个代理指标时,本文所提非线性几何融合机制取得最优效果,在 CIFAR-10、CIFAR-100、ImageNet16-120 数据集上的 τ 值分别达到 0.712 0、0.705 2、0.698 1,相较于线性融合方式在 CIFAR-10 数据集上提升 0.013 7,评估性能显著更优。综上可知,本文设计的多维度互补代理指标与非线性几何融合方法能够有效整合评估信息,提升架构性能预测精度,是实现高相关性性能预测的关键所在。

表2 低成本评估指标消融实验

Table 2 Ablation study of low-cost evaluation metrics

Index	S_{robust}	$S_{isotropy}$	$S_{uncertainty}$	$S_{uniformity}$	FM	CIFAR-10	CIFAR-100	ImageNet16-120
1	✔				—	0.489 7	0.502 9	0.345 1
2		✔			—	0.419 2	0.368 3	0.399 3
3			✔		—	0.555 3	0.579 8	0.548 5
4				✔	—	0.577 6	0.605 5	0.584 2
5	✔	✔			NL	0.599 6	0.569 5	0.547 7
6	✔		✔		NL	0.608 8	0.625 9	0.546 6
7	✔			✔	NL	0.617 8	0.636 8	0.556 0
8			✔	✔	NL	0.590 7	0.618 7	0.588 4
9	✔	✔	✔		NL	0.687 0	0.671 5	0.659 7
10		✔	✔	✔	NL	0.682 6	0.675 9	0.674 9
11	✔		✔	✔	NL	0.630 3	0.652 1	0.596 5
12	✔	✔	✔	✔	L	0.698 3	0.670 1	0.591 9
13	✔	✔	✔	✔	NL	0.712 0	0.705 2	0.698 1

3.1.3 最终搜索性能验证

为进一步验证所提方法在实际应用中的效能，本文以NAS-Bench-201作为基准搜索空间，在固定采样规模为1 000个的条件下，选取现有先进的无训练评估方法，分别对CIFAR-10、CIFAR-100及ImageNet16-120这三个数据集的表现展开对比。不同的方法均独立运行3次，每轮搜索均采用所得架构的最优模型性能作为最终结果，实验结果采用性能的平均值±标准差的形式进行呈现。其中Optimal表示待选模型当中最优模型的准确率，作为本实验的参照标准。最终的实验结果如表3所示。

实验结果表明，本文所提COER-NAS在三个数据集上均展现出接近最优的评估性能，验证了其架构筛选的鲁棒性。在CIFAR-10数据集当中，本文策略的测试准确率达93.41%，比次优的ZiCo提高了0.19个百分点，与理论最优值仅相差0.79个百分点，差距显著小于其他对比方法；在更复杂的ImageNet16-120数据集上，其准确率为44.33%，分别较ZiCo、NASWOT提高了0.30和0.21个百分点，与理论最优值仅相差1.24个百分点；在CIFAR-100数据集当中，测试准确率为70.23%，虽并未达到最优表现，但显著优于ZiCo、LCMNAS等多数无训练的评估方法，与次优的NASWOT，仅仅相差0.45个百分点。值得注意的是，COER-NAS在CIFAR-100上的标准差处于最低水平，远低于NASWOT以及SynFlow；在CIFAR-10和ImageNet16-120上，整体也呈现出良好的稳定性。综上所述，本文所提策略在三个数据集上的性能与最优值最为接近，证明了多维评估体系可有效捕获模型性能的核心决定要素，在几乎无训练的条件下实现了高效可靠的架构搜索，为资源受限场景中的自动化模型架构设计提供了良好的解决思路。

3.2 实际应用场景下所提方法的性能评估

为评估COER-NAS在实际场景中的实用性，本文选取了工业焊缝缺陷检测以及医疗视网膜病变分级这两类不同场景的数据集，旨在证明本文提出方法在实际应用场景中的综合优势和跨领域的泛化性。此外，为保证HD-DUAAS超参数设定的合理性，本文依照不同的超参数设计策略搭建了5组具有代表性的超参数组合，并通过对比实验，挑选出性能和效率达成最优平衡的组合，并将其作为后续实验所采用的超参数配置。

3.2.1 数据集介绍

本文使用了面向工业场景的焊缝缺陷检测数据集[26]，该数据集包含未焊透、气孔、夹渣、裂纹、未熔合五类缺陷类型的样本。使用Tektronix DPO 2024B示波器、KARL DEUTSCH（ECHOGRAPH）探伤仪与2.5P 9×9 K2.5斜探头构成的采集系统，对缺陷试样进行数据采集，累计获取到544条超声A扫一维时序信号。借助马尔可夫转移场变换，将这些时域信号生成二维视觉图像以适配模型输入。此外，为缓解样本数量匮乏的问题，本文利用随机垂直翻转、随机平移、颜色抖动、随机擦擦与随机旋转5种增强策略将样本数量扩充至2 915张。

选用面向医疗场景的APTOS-2019数据集，用于糖尿病视网膜病变0~4级的分级任务。该数据集共收集了3 662张眼底图像，其主要挑战在于其类别的严重失衡以及病变特征细微这两个方面：0级（无病变）样本多达1 805张，而3级（重度病变）样本仅有193张。此外，早期病变样本视觉上的特征十分微弱，这对模型识别细微判别特征、增强对少数类样本

表3　不同低成本评估策略性能对比

Table 3　Performance comparison of different low-cost evaluation strategies

策略类型	策略	搜索时间/s	CIFAR-10		CIFAR-100		ImageNet16-120	
			测试/%	验证/%	测试/%	验证/%	测试/%	验证/%
无训练方法	Random	N/A	83.89±12.15	80.54±12.42	57.53±14.85	57.29±14.93	29.13±14.02	29.47±13.32
	Snip	409.35	90.58±1.78	87.39±1.52	63.48±2.02	62.62±1.92	32.49±5.69	32.30±5.05
	Grasp	930.21	90.59±1.79	87.27±1.43	59.78±4.13	59.81±4.86	31.58±5.16	31.04±4.33
	SynFlow	403.44	93.14±0.73	89.60±0.79	71.04±0.80	71.43±0.36	39.90±6.97	39.80±7.12
	GradNorm	413.30	90.64±1.83	87.59±1.72	63.48±2.02	62.62±1.92	32.29±5.50	32.40±5.14
	NASWOT	347.17	93.12±0.19	89.87±0.30	70.68±0.94	70.31±0.55	44.12±1.18	43.46±0.78
	ZiCo	505.91	93.22±0.35	89.67±0.16	69.68±0.84	69.37±1.22	44.03±1.18	43.50±0.78
	LCMNAS	791.36	89.70±0.88	86.86±1.03	67.93±2.86	67.67±2.62	38.53±0.40	38.30±1.80
	ePADS	524.46	93.11±0.17	89.89±0.32	70.44±0.53	70.15±0.38	39.81±6.29	39.71±5.91
	COER-NAS	652.61	93.41±0.31	89.68±0.25	70.23±0.41	69.80±0.58	44.33±1.62	43.46±1.78
最优值	Optimal	N/A	94.20±0.12	91.27±0.15	72.83±0.14	72.04±0.78	45.57±0.97	46.36±0.54

的敏感程度有了更高要求。

为保证模型输入的一致性与稳定性,本文对这两类数据集均进行了统一的预处理工作。将所有图像均统一缩放至224 × 224,并对输入数据进行标准正则化的相关处理,以降低不同来源、不同尺度数据对模型训练的干扰。在此基础上,按照4∶1的比例将数据集划分为训练集和测试集,用于后续的模型训练及评估。

3.2.2　搜索策略最优超参数组合的实验遴选

为确定NAS中搜索策略的最优超参数组合,本文围绕基线对照、全局搜索、快速收敛、低计算成本、精度强化五大核心需求,预设了5组有典型倾向性的超参数组合,分别对应经典基准、高探索性、高利用性、高效性与深度优化的导向,以探索最适配本文的超参数组合。实验选取APTOS-2019数据集作为评估数据集,以控制变量的方式开展对比实验,保证超参数组合为唯一变量。为降低随机因素对结果所产生的干扰,各超参数组合都独立且重复运行3次,并且以性能(最优架构的评估分数)和效率(搜索平均耗时)作为双维度的评估考量,实现对超参数组合综合性能的全面度量。各组合具体超参数的配置情况如表4所示,对应的性能与效率对比结果如表5所示。

实验结果表明,经典基准组合(序号1)有着63.86%的分类性能,平均耗时为4 085.02 s,性能的稳定性及效率都处于基础水平;高探索性组合(序号2)将性能提升到了65.49%,但平均耗时激增到6 584.97 s,计算成本大幅增加;高利用性组合(序号3)的性能为64.88%,平均耗时为3 123.45 s,其在效率方面有一定程度的优化,但性能的波动幅度较大,鲁棒性不足;高效性组合(序号4)取得了65.13%的优良性能,且平均耗时仅为661.05 s,其性能稳定性和计算效率达成平衡;深度优化组合(序号5)性能指标为65.36%,但平均耗时高达4 693.08 s。综合考量模型性能表现与计算效率的工程化需求,序号4对应的轻量化超参数组合被确定为后续实验的固定配置。

表4　HD-DUAAS超参数组合配置表

Table 4　HD-DUAAS hyperparameter configuration table

序号	P	MaxIter	T	交叉概率	变异概率
1	50	50	3	0.8/0.7/0.6	0.05/0.03/0.03
2	80	100	2	0.9/0.8/0.8	0.08/0.05/0.05
3	40	40	4	0.7/0.6/0.5	0.03/0.02/0.02
4	30	10	3	0.8/0.7/0.6	0.05/0.03/0.03
5	60	70	3	0.85/0.75/0.7	0.06/0.04/0.04

表5　不同超参数组合的性能与效率对比表

Table 5　Performance and efficiency comparison of different hyperparameter combinations

组合序号	三次运行的均值±标准差/%	平均耗时/s
1	63.86 ± 2.58	4 085.02
2	65.49 ± 0.35	6 584.97
3	64.88 ± 3.41	3 123.45
4	65.13 ± 0.58	661.05
5	65.36 ± 1.02	4 693.08

3.2.3　所提方法在焊缝数据集上的性能验证

本节在自建的焊缝缺陷检测数据集上,验证所提方法的实用价值。训练阶段采用Adam优化器进行参数更新,初始学习率设为1×10^{-4},批次大小为16,最大迭代次数epoch为150轮。学习率采用余弦退火策略进行动态调整,同时引入早停机制(patience = 20)以防止过拟合,损失函数选用交叉熵损失。评估维度同时兼顾性能与复杂度,性能方面将准确率、F1分数、召回率作为核心检测指标,复杂度层面则统计参数量与推理时间,并选取传统手工设计的模型与NAS

自动设计模型进行比较，以此综合验证所提方法的有效性与优越性。表6呈现了实验的详细结果。

实验结果表明，所提方法在该任务中表现出较优的综合性能。从分类性能来看，所提方法搜索得到的最优架构在测试集上实现97.22%的分类准确率，F1分数与召回率分别达97.21%与97.22%。对比无训练NAS方法，本文架构准确率较MSTF-NAS提升1.39个百分点，较PO-NAS大幅提升8.03个百分点，即便与手工构建的最优视觉模型MAViT-S相比，仍实现了2.43个百分点的精度跃升。这一显著的精度优势可从图4最优架构的结构设计得到明确解释，该架构采用三阶段递进式层级设计，通过特征提取分支与降采样操作的解耦，实现了不同阶段特征信息的高效挖掘与传递。在模型效率层面，本文最优架构参数量为10.962×10^6，单图推理时间仅0.41 ms，完全满足工业场景实时性要求。本文方法的搜索时间为227.97 s，相较于MSTF-NAS的306.94 s耗时更短，且性能更优，这种效率与性能的双重优化，核心得益于HD-DUAAS的定向优化能力，以及所提多维低成本评估策略对架构性能的精准预测与高效筛选。

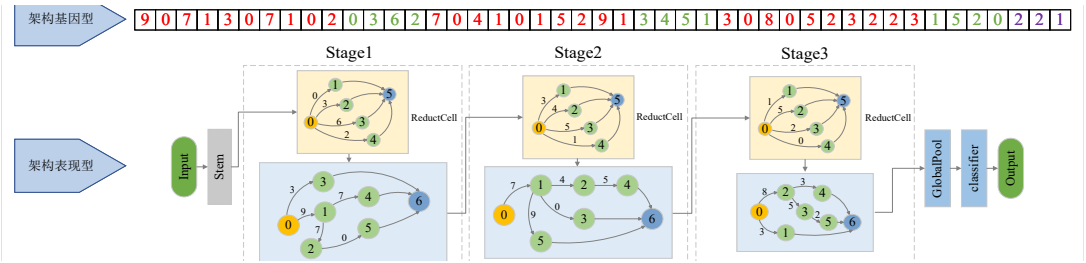

图4 自建焊缝数据集上的最优架构示意图

Figure 4 Schematic of the optimal architecture on the custom weld dataset

表6 不同模型在自建焊缝缺陷检测数据集上的性能与效率对比

Table 6 Comparison of performance and efficiency of different models on a self-built weld defect detection dataset

Type	Years	Methods	Param/M	Inference time/ms	Accuracy/%	F1 Score/%	Recall/%	Searching time/s
手动构建	2021	PVT-Small[27]	23.586	0.31	76.04	76.10	76.04	—
	2021	CoAtNet-0[28]	26.579	0.24	93.06	93.07	93.06	—
	2022	EfficientFormer_l3[29]	31.389	0.31	68.58	68.71	68.58	—
	2022	Mobilevit-V2_200[30]	18.402	0.45	89.24	89.16	89.24	—
	2023	ConvNeXt V2-Tiny[31]	27.801	0.20	91.84	91.81	91.84	—
	2023	PVTv2-B2[32]	24.852	0.30	93.75	93.77	93.75	—
	2024	InceptionNeXt-T[33]	25.756	0.43	80.73	80.46	80.73	—
	2025	EfficientViM_M4[34]	18.021	0.79	90.28	90.25	90.28	—
	2025	MAViT-S[35]	25.093	2.14	94.79	94.79	94.79	—
无训练NAS	2023	EBNAS[36]	8.190	1.41	60.76	58.68	60.76	7 872.00
	2024	MSTF-NAS[37]	9.659	0.55	95.83	95.83	95.83	306.94
	2025	PO-NAS[38]	4.020	1.06	89.19	89.21	89.22	10 921.64
	2026	COER-NAS	10.962	0.41	97.22	97.21	97.22	227.97

3.2.4 所提方法在APTOS-2019数据集上的性能验证

为进一步验证所提方法的跨领域泛化能力，本文在APTOS-2019糖尿病视网膜病变分级数据集上进一步开展了评估。实验配置与评估指标均与第3.2.3节保持一致，仅将初始学习率调整为5×10^{-5}。实验结果如表7所示。

实验结果表明，本文所提方法搜索得到的最优架构，在类别高度不平衡、特征难以区分的医疗测试集上取得了81.94%的准确率，F1分数与召回率分别达到81.32%和81.94%。在无训练NAS范畴内，其准确率较MSTF-NAS提升1.11个百分点，较PO-NAS和EBNAS分别提升3.77和9.30个百分点，F1分数与召回率也实现了显著提升，同时优于手工设计的先进模型，较CoAtNet-0和MAViT-S分别提升1.68和2.68个百分点。究其根本，这一性能优势可归因于本文提出的多维评估策略。本文认为，“梯度-层结构规整性”有效缓解了类别不平衡数据带来的训练波动，“特征空间各向同性”评估则增强了模型对少数类别样本的判别能力，从而实现了更优的泛化性能。从效率层面看，最终架构参数量为13.825×10^6，单张图片推理时

间仅0.46 ms,满足实时分析需求,完整的神经架构搜索耗时为559.34 s,显著低于EBNAS和PO-NAS,仅略高于MSTF-NAS,充分体现了无训练NAS在效率与性能上的平衡优势。

表7　不同模型在APTOS-2019数据集上的性能与效率对比

Table 7　Performance and efficiency comparison of different models on the APTOS-2019 dataset

Type	Years	Methods	Param/M	Inference time/ms	Accuracy/%	F1 Score/%	Recall/%	Searching time/s
手动构建	2021	PVT-Small	23.586	0.28	72.73	66.34	72.73	—
	2021	CoAtNet-0	26.579	0.24	80.26	79.81	80.26	—
	2022	EfficientFormer_l3	31.389	0.30	73.86	68.63	73.86	—
	2022	Mobilevit V2_200	18.402	0.22	78.84	76.94	78.84	—
	2023	ConvNeXt V2-Tiny	27.801	0.40	74.29	69.16	74.29	—
	2023	PVTv2-B2	24.852	0.32	78.55	76.68	78.55	—
	2024	InceptionNeXt-T	25.756	0.20	73.44	66.45	73.44	—
	2025	EfficientViM_M4	18.021	0.35	75.85	72.29	75.85	—
	2025	MAViT-S	25.093	2.47	79.26	78.56	79.26	—
无训练NAS	2023	EBNAS	7.956	2.13	72.64	65.00	72.64	9 996.078
	2024	MSTF-NAS	7.775	0.24	80.83	79.51	80.83	595.960
	2025	PO-NAS	4.443	2.46	78.17	59.64	58.11	27 649.860
	—	COER-NAS	13.825	0.46	81.94	81.32	81.94	559.340

4　结束语

本文针对NAS在实际应用中面临的搜索空间功能耦合、高维空间探索效率低下及低成本性能评估维度单一等系统性问题,提出一种效率与可靠性协同的递进式NAS方法。核心工作包括:设计了递进式表征降维可分离的架构搜索空间,通过解耦特征提取与降采样操作,实现了架构空间高效降维的同时保留深层判别性信息;提出了高维解耦式单元自适应架构搜索策略,结合编码分块解耦与单元级定向遗传操作,有效改善高维编码空间搜索效率;构建了多维低成本模型性能评估策略,通过互补的四维评估体系与非线性融合策略,保障了架构筛选的稳定性及评估结果的可靠性。实验结果表明,所提方法在模型综合性能以及跨领域泛化能力上优势显著。在NAS-Bench-201基准搜索空间的CIFAR-10数据集上,肯德尔相关性系数达0.712 0。在工业焊缝缺陷检测和医疗视网膜病变分级的实际任务中,所得最优架构在保证毫秒级推理延迟的基础上,分类精度分别达到了97.22%和81.94%,且搜索效率显著高于现有的无训练方法,充分体现其实际应用价值。综上,本文所提方法为NAS技术从实验室走向工程化应用提供了系统性技术支撑。

本文未来工作将围绕以下三方面进行开展:(1)进一步优化模型性能评估策略,扩展评估维度并改善融合机制,进一步提高模型性能的预测精度;(2)对搜索空间的设计进行优化,融入具有更强表达能力的网络结构,提升模型性能上限;(3)拓展所提框架至语义分割、目标检测等下游任务,系统验证其泛化能力。

参考文献

[1] 章晋睿, 龙婷婷, 张德宇, 等. 端智能推理加速技术综述[J]. 电子学报, 2025, 53(4): 1063-1102.
Zhang Jinrui, Long Tingting, Zhang Deyu, et al. On-device intelligence acceleration technologies: A survey[J]. Acta Electronica Sinica, 2025, 53(4): 1063-1102. (in Chinese)

[2] Li Yangyang, Liu Guanlong, Shang Ronghua, et al. Meta knowledge assisted evolutionary neural architecture search[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(10): 10225-10237.

[3] Liang Jing, Liu Genyue, Bi Ying, et al. Evolutionary neural architecture search for remote sensing image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(10): 17886-17900.

[4] Zou Juan, Tong Jinghui, Xia Yizhang, et al. TS-ENAS: Two-stage evolution for cell-based network architecture search[J]. Expert Systems with Applications, 2026, 296: 129150.

[5] 蒋鹏程, 薛羽. 基于排序得分预测的演化神经架构搜索方法[J]. 计算机学报, 2024, 47(11): 2522-2535.
Jiang Pengcheng, Xue Yu. Evolutionary neural architecture search with predictor of ranking-based score[J]. Chinese Journal of Computers, 2024, 47(11): 2522-2535. (in Chinese)

[6] Yang Jiechao, Liu Yong, Wang Wei, et al. PATNAS: A path-based training-free neural architecture search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(3): 1484-1500.

[7] Lopes V, Alexandre L A. Toward less constrained macro-neural architecture search[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(2): 2854-2868.

[8] Onzo B M, Xue Yu, Neri F. Surrogate-assisted evolutionary neural architecture search based on smart-block discovery[J]. Expert Systems with Applications, 2025, 277: 127237.

[9] 张睿, 张鹏云, 孙超利. 基于多域融合及神经架构搜索的语音增强方法[J]. 通信学报, 2024, 45(2): 225-239.
Zhang Rui, Zhang Pengyun, Sun Chaoli. Speech enhancement method based on multi-domain fusion and neural architecture search[J]. Journal on Communications, 2024, 45(2): 225-239. (in Chinese)

[10] Wu Mengting, Lin H I, Tsai C W. A training-free neural architecture search algorithm based on search economics[J]. IEEE Transactions on Evolutionary Computation, 2024, 28(2): 445-459.

[11] Yamasaki T, Wang Zhehui, Luo T, et al. RBFleX-NAS: Training-free neural architecture search using radial basis function kernel and hyperparameter detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 10057-10071.

[12] 杨乐, 马永杰, 平镐羽, 等. 角度修正和分级多种群的动态多目标进化算法[J]. 电子学报, 2024, 52(9): 3278-3290.
Yang Le, Ma Yongjie, Ping Gaoyu, et al. Dynamic multi-objective evolutionary algorithm based on angle correction and hierarchical multi-population[J]. Acta Electronica Sinica, 2024, 52(9): 3278-3290. (in Chinese)

[13] He Guang, Lu Xiaoli. Good point set and double attractors based-QPSO and application in portfolio with transaction fee and financing cost[J]. Expert Systems with Applications, 2022, 209: 118339.

[14] Huang Ting, Tang Xiaoan, Zhao Shuangyao, et al. Linguistic information-based granular computing based on a tournament selection operator-guided PSO for supporting multi-attribute group decision-making with distributed linguistic preference relations[J]. Information Sciences, 2022, 610: 488-507.

[15] Cai Xingyu, Huang Jiaji, Bian Yuchen, et al. Isotropy in the contextual embedding space: Clusters and manifolds[C/OL]//Proceedings of the 9th International Conference on Learning Representations, 2021. https://openreview.net/forum id=xYGNO86OWDH.

[16] Karthik, Maggie, Dane S. APTOS 2019 blindness detection[EB/OL]. [2026-03-13]. https://kaggle. com/competitions/aptos2019-blindness-detection.

[17] Dong Xuanyi, Yang Yi. NAS-bench-201: Extending the scope of reproducible neural architecture search[C/OL]//Proceedings of the 8th International Conference on Learning Representations, 2020. https://openreview. net/forum id=HJxyZkBKDr.

[18] Lee N, Ajanthan T, Torr P H. Snip: Single-shot network pruning based on connection sensitivity[C/OL]//Proceedings of the 7th International Conference on Learning Representations, 2019. https://openreview. net/forum id=B1VZqjAcYX.

[19] Wang Chaoqi, Zhang Guodong, Grosse R. Picking winning tickets before training by preserving gradient flow[C/OL]//Proceedings of the 8th International Conference on Learning Representations, 2020. https://openreview. net/forum id=SkgsACVKPH.

[20] Tanaka H, Kunin D, Yamins D L K, et al. Pruning neural networks without any data by iteratively conserving synaptic flow[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2020: 535.

[21] Abdelfattah M S, Mehrotra A, Dudziak Ł, et al. Zero-cost proxies for lightweight NAS[C/OL]//Proceedings of the 9th International Conference on Learning Representations, 2021. https://openreview. net/forum id=0cmMMy8J5q.

[22] Mellor J, Turner J, Storkey A, et al. Neural architecture search without training[C]//Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 7588-7598.

[23] Zhang Zhihao, Jia Zhihao. GradSign: Model performance inference with theoretical insights[C/OL]//Proceedings of the Tenth International Conference on Learning Representations, 2022. https://openreview.net/forum id=HObMhrCeAAF.

[24] Li Guihong, Yang Yuedong, Bhardwaj K, et al. ZiCo: Zero-shot NAS via inverse coefficient of variation on gradients[C/OL]//Proceedings of the Eleventh International Conference on Learning Representations, 2023. https://openreview.net/forum id=rwo-ls5GqGn.

[25] Huang Junhao, Xue Bing, Sun Yanan, et al. Efficient perturbation-aware distinguishing score for zero-shot neural

architecture search[J]. Applied Soft Computing, 2025, 182: 113447.

[26] Zhang Rui, Gao Meirong, Zhang Pengyun, et al. Research on an ultrasonic detection method for weld defects based on neural network architecture search[J]. Measurement, 2023, 221: 113483.

[27] Wang Wenhai, Xie Enze, Li Xiang, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 548-558.

[28] Dai Zihang, Liu Hanxiao, Le Q V, et al. CoAtNet: Marrying convolution and attention for all data sizes[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021: 303.

[29] Li Yanyu, Yuan Geng, Wen Yang, et al. EfficientFormer: Vision transformers at MobileNet speed[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2022: 940.

[30] Mehta S, Rastegari M. Separable self-attention for mobile vision transformers[J]. Transactions on Machine Learning Research, 2023, 2023.

[31] Woo S, Debnath S, Hu Ronghang, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 16133-16142.

[32] Wang Wenhai, Xie Enze, Li Xiang, et al. PVT v2: Improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.

[33] Yu Weihao, Zhou Pan, Yan Shuicheng, et al. InceptionNeXt: When inception meets ConvNeXt[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 5672-5683.

[34] Lee S, Choi J, Kim H J. EfficientViM: Efficient vision mamba with hidden state mixer based state space duality[C]//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 14923-14933.

[35] Fan Qihang, Huang Huaibo, Ai Yuaang, et al. Rectifying magnitude neglect in linear attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2025: 21505-21514.

[36] Shi Chaokun, Hao Yuexing, Li Gongyan, et al. EBNAS: Efficient binary network design for image classification via neural architecture search[J]. Engineering Applications of Artificial Intelligence, 2023, 120: 105845.

[37] Ali M J, Moalic L, Essaid M, et al. Evolutionary neural architecture search for 2D and 3D medical image classification[C]//Proceedings of the 24th International Conference on Computational Science. Cham: Springer, 2024: 131-146.

[38] Lin Mingzhuo, Luo Jianping. Per-architecture training-free metric optimization for neural architecture search[C]// The Thirty-Ninth Annual Conference on Neural Information Processing Systems, 2025. https://openreview.net/forum id=mVh0lIsdUl.

作者简介

张　睿　男,1987年2月出生于山西省太原市。现为太原科技大学计算机科学与技术学院教授、博士生导师。主要研究方向为神经架构搜索与自动机器学习。

E-mail: zhangrui@tyust.edu.cn

魏晓楠　男,2001年12月出生于山西省太原市。现为太原科技大学计算机科学与技术学院硕士研究生。主要研究方向为神经架构搜索与自动机器学习。

E-mail: s202420211020@stu.tyust.edu.cn

孙超利　女,1978年2月出生于浙江省诸暨市。现为太原科技大学计算机科学与技术学院教授、博士生导师。主要研究方向为代理模型辅助的进化优化与自动机器学习。

E-mail: chaoli.sun@tyust.edu.cn