神经网络轻量化技术：从静态压缩到动态计算的演进与展望

王恩良[1,2]，阎庆昕[1,2]，达明添[1,2]，孙知信[1,2]
（1.南京邮电大学江苏省邮政大数据技术与应用工程研究中心，江苏 南京 210003；
2.南京邮电大学国家邮政局邮政行业技术研发中心（物联网技术），江苏 南京 210003）

摘　要：神经网络规模增长与边缘设备算力受限之间的矛盾推动了轻量技术的发展。基于此，梳理了从静态压缩、神经架构搜索到动态计算的三阶段演进：静态压缩通过量化、剪枝与蒸馏实现模型优化；神经架构搜索突破人工设计限制；动态计算实现按需推理。通过建立“参数–结构–知识”统一框架，解析了动态结构与条件计算等核心机制。轻量技术正从固定优化走向自适应计算、从孤立方法走向协同融合，为构建高效可扩展模型提供重要理论基础。

关键词：神经网络轻量化；模型压缩；神经架构搜索；动态计算；条件计算；知识蒸馏；自适应推理
中图分类号：TP183
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025157

Lightweight neural network techniques: evolution and prospect from static compression to dynamic computation

WANG Enliang[1,2], YAN Qingxin[1,2], DA Mingtian[1,2], SUN Zhixin[1,2]
1. Engineering Research Center of Post Big Data Technology and Application of Jiangsu Province, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. Research and Development Center of Post Industry Technology of the State Posts Bureau (Internet of Things Technology), Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: The conflict between neural network's growing scale and edge device's limited resources has accelerated lightweight neural network techniques. Their evolution was examined through three stages: static compression (quantization, pruning, knowledge distillation), neural architecture search, and dynamic computation. A unified "parameter-structure-knowledge" framework was proposed to reveal underlying connections, analyzing core mechanisms like dynamic structures and conditional computation. Lightweight technology is evolving from fixed to adaptive optimization and from isolated to integrated approaches, offering key insights for building efficient, scalable models.

Keywords: lightweight neural network, model compression, neural architecture search, dynamic computation, conditional computation, knowledge distillation, adaptive inference

0　引言

深度神经网络（DNN, deep neural network）作为变革性工具，已广泛应用于医疗、金融、农业、工业、互联网等领域，并在图像处理、语音识别、

收稿日期：2025–07–04；**修回日期：**2025–09–05
通信作者：孙知信，sunzx@niupt.edu.cn
基金项目：国家自然科学基金资助项目（No.61972208, No.62272239）；江苏省农业科技创新基金资助项目（No.CX(22)1007）；贵州省科技支撑基金资助项目（No.[2023]一般272）
Foundation Items: The National Natural Science Foundation of China (No.61972208, No.62272239), Jiangsu Agriculture Science and Technology Innovation Fund (No.CX(22)1007), Guizhou Provincial Key Technology Research and Development Program (No.[2023]一般 272)

自然语言理解和自动化系统中发挥关键作用。然而，DNN 规模和复杂度的激增导致计算开销呈指数级上升（如图 1 所示），远超摩尔定律增速[1]，其参数量从 AlexNet[2] 的 6 200 万剧增至 GPT-4 的 1.75 万亿，加剧了计算需求与资源间的矛盾。

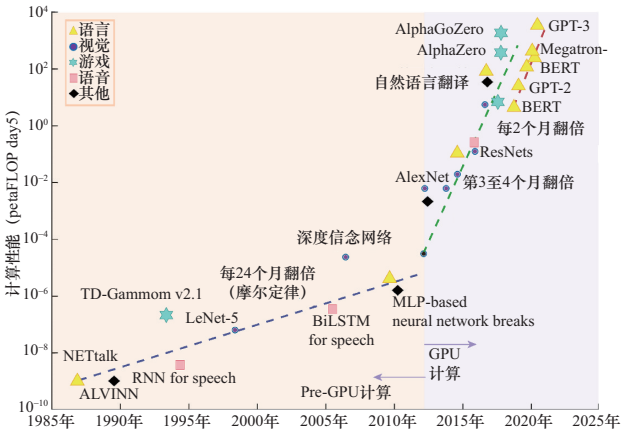

图 1　全球计算开销增速超过摩尔定律

应对这一瓶颈的两大技术路径是互补且协同的：云端部署[3]与轻量化模型设计[4]。云端部署减轻了设备负担，但引入了网络时延、实时性及安全问题。轻量化模型设计具备本地低时延与高隐私优势，其性能则受限于终端资源。面向不同需求，2 种技术路径可共同演进。鉴于大型 DNN 普遍存在参数冗余[5]，轻量化技术通过高效利用此冗余，成为资源受限场景下实现高效边缘部署的关键技术，也是下文探讨重点。

神经网络轻量化技术的发展经历了 3 个明显的演进阶段。第一阶段（2015—2018 年）以静态压缩技术为主导，包括参数量化[6]、网络剪枝[7]、知识蒸馏[8]和低秩分解[9]等方法，对预训练模型进行一次性复杂度优化。典型工作 Deep Compression[4] 融合剪枝/量化/编码技术，显著提升了 AlexNet/VGG-16 压缩效率，但面临目标单一、压缩率受限及强专家依赖等瓶颈。

第二阶段（2018—2021 年）以神经架构搜索（NAS, neural architecture search）[10]为代表，推动轻量化从“压缩”向“设计”跃迁。NASNet[11]首次验证自动架构超越人工设计，MnasNet[12]实现精度–时延多目标优化，但所得静态架构缺乏输入适应性。

第三阶段（2021 年至今）以动态计算[13]为框架，输入感知的动态深度[14]、通道宽度[15]及跨层路由[16]为特征，突破“单一模型适配全样本”局限。例如，AdaViT[17]在 ImageNet 上以 0.8% 精度损失换取 2.5 倍加速，彰显场景化效率优化潜力。

上述从静态压缩–自动设计–动态计算的演进轨迹，深刻反映了对模型效率与适应性的双重追求。当前核心挑战在于：理论框架缺失阻碍轻量化技术的内在关联阐释与协同机制构建；动态计算范式受制于训练/部署复杂度；硬件自适应优化与多技术融合方法论尚不完善。

本文旨在对神经网络轻量化技术进行系统性综述。核心贡献在于：1) 提出轻量化技术“静态压缩–神经架构搜索–动态计算”三阶段演进模型，刻画了 2015 年至今的技术发展脉络。该模型揭示了轻量化技术从后处理优化、设计时优化到运行时自适应的范式转变，反映了优化粒度从局部到全局、执行模式从静态到动态、目标函数从单一到多维的演进规律；2) 构建“参数–结构–知识”统一分析框架，从 3 个维度解构轻量化技术体系，阐明量化、剪枝、蒸馏等技术的互补关系及协同机制，为技术选型与融合提供理论依据；3) 系统剖析动态计算范式，从动态网络结构、条件参数生成和动态知识传递 3 个层面阐述其实现机制，通过条件计算框架分析参数–结构–知识的运行时协同；4) 分析轻量化技术在终身自适应、硬件–算法协同、隐私保护等方向的发展趋势，探讨从静态优化向环境感知计算演进的技术路径。

1　轻量化技术的统一理论框架

深度神经网络的轻量化本质是多目标优化问题，需要在模型性能、计算效率和资源约束之间寻求平衡。为深入理解不同轻量化技术的内在联系，本节提出了一个统一理论框架，从优化目标、约束条件和搜索空间 3 个维度对现有技术进行系统分析。

神经网络轻量化可以形式化为如式(1)和式(2)所示约束优化问题。

$$\min_{\theta'} \mathcal{L}(\theta', \mathcal{D}) + \lambda \mathcal{R}(\theta') \tag{1}$$

$$\text{s.t.} \mathcal{C}(\theta') \leqslant \mathcal{C}_{\text{budget}} \tag{2}$$

其中，θ'为轻量化后的模型参数，$\mathcal{L}$为任务损失函数，$\mathcal{D}$为数据集，$\mathcal{R}$为正则化项，$\mathcal{C}$为资源消耗度量（如参数量、每秒浮点操作数 FLOPS、时延等），$\mathcal{C}_{\text{budget}}$为资源预算。基于优化变量的不同，轻量化技术可分为 3 个基本范式。

1.1　技术演进的内在逻辑

轻量化技术的演进遵循从局部到全局、静态到动态、单目标到多目标的规律，其发展可划分为3个阶段。

静态压缩技术采用全局统一压缩策略（如8 bit量化[18]或高剪枝率[19]），实现2~50倍压缩率，但受限于一次性优化、专家依赖及层间特性忽略。神经架构搜索技术通过NAS和MnasNet[20]将轻量化嵌入训练过程（5~100倍压缩率），实现分层任务适应，但其搜索成本高昂且输出静态架构难以满足动态需求。动态计算技术基于动态网络和条件计算[21]实现样本自适应优化（2~10倍压缩率），在推理时按输入复杂度调整计算路径，显著提升灵活性，其演进特征如表1所示。

此演进映射研究认知的深化，从“如何减少冗余”到“如何设计高效结构”，进一步思考“如何自适应计算”。静态压缩阶段关注的是“如何减少冗余”，通过后处理方式优化预训练模型。神经架构搜索阶段转向“如何设计高效结构”，将轻量化纳入模型设计过程。动态计算阶段则聚焦于“如何自适应计算”，实现了真正的按需计算。

从技术实现角度看，参数、结构与知识三要素的协同优化构成统一设计空间，如图2所示。参数优化依托量化技术[22]、低秩分解[23]与稀疏化表示降低存储与计算开销；结构优化借助网络剪枝、轻量化模块设计[24]与神经架构搜索[25]实现高效拓扑；知识优化通过蒸馏机制[26]（包括知识蒸馏、自蒸馏与在线蒸馏[27]）传递隐式特征表征，以维持模型性能。

三要素通过六类双向路径形成协同机制。参数与结构间：量化感知剪枝以压缩后权重分布指导剪枝，结构调整亦反向制约量化策略——稀疏网络要求差异化比特分配[28]。结构与知识间：蒸馏约束NAS的搜索空间，结构先验为蒸馏提供归纳偏置[29]。参数与知识间：量化知识恢复通过蒸馏补偿信息损失，特征分布可进一步引导参数精度分配。协同优化呈现层次化效益：单一优化可实现局部提升（参数压缩4~32倍、结构加速2~10倍、知识恢复精度2%~5%），而协同策略可突破单维边界，达成50~100倍综合压缩、20~50倍端到端加速与不足1%的精度损失，如Deep Compression、CLIP-Q、QKD与AutoCompress等一系列工作验证了该框架的有效性[30]。

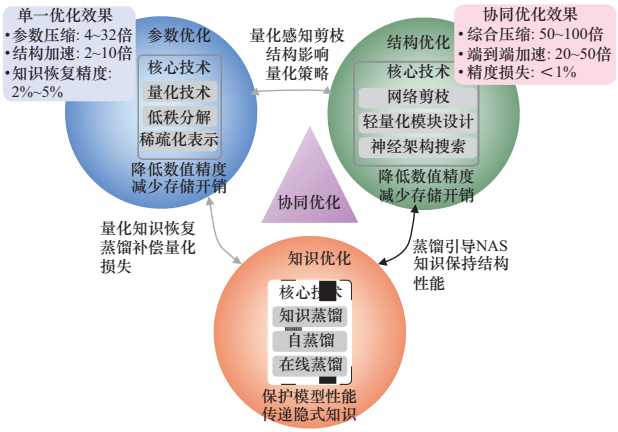

图2　参数–结构–知识三要素的协同优化关系

中心三角区集中体现三要素的协同交互，参数信息损失通过结构与知识补偿，结构容量下降由参数与知识弥补，知识传递开销受参数–结构联合调控。该相互增强机制在维持性能的前提下支持极限压缩，持续拓展轻量化技术边界。

1.2　评估指标体系

轻量化技术的评估需要综合考虑多个维度的指标，传统的单一指标已无法全面反映轻量化效果，如表2所示。

在实际应用中，不同场景对各项指标的重视程度不同。轻量化技术的效能评估需构建场景驱动的差异化框架，移动端部署因严苛的电池容量与存储空间限制，对模型体积及推理时延（需满足<100 ms人机交互阈值）高度敏感。云端场景则优先保障吞吐量与精度，其高并发处理需求可兼容单次推理时延。这种评估维度差异源于资源约束–

表1　神经网络轻量化技术的演进特征

技术代别	优化时机	优化粒度	适应性	代表方法	压缩率范围	主要局限
第一代：静态压缩	训练后	固定全局	无	量化、剪枝、蒸馏	2~50倍	一次性优化，经验依赖
第二代：神经架构搜索	训练中	分层差异化	任务级	NAS、MnasNet	5~100倍	搜索开销大，静态结果
第三代：动态计算	推理时	样本自适应	输入级	动态网络、条件计算	2~10倍	训练复杂，部署困难

服务目标的根本性对立——前者受物理条件刚性制约，后者由服务质量（QoS, quality of service）需求主导。

表 2　轻量化技术的多维度评估指标

指标类别	具体指标	计算方法	典型范围
模型复杂度	参数量压缩率	$\frac{\|\theta\|}{\|\theta'\|}$	2~100 倍
	FLOPS 减少率	$1-\frac{\text{FLOPS}'}{\text{FLOPS}}$	50%~95%
	内存占用	模型大小/MB	1~50
计算效率	推理时延	GPU/CPU/边缘设备时延/ms	1~100
	吞吐量	样本数/s	100~10 000
	能耗	焦耳/推理/J	0.01~1
模型性能	Top-1 精度损失	Acc − Acc′	0~5%
	任务特定指标	mAP、BLEU 等	任务相关
部署友好性	硬件兼容性	支持的硬件平台数	1~10
	量化友好度	INT8 精度损失	0~2%

参数量与 FLOPS 的互补衰减、精度与效率的帕累托竞争边界、硬件稀疏支持度对时延优化的调制效应，构成指标间的 3 类拓扑关系。因此，基于指标的约束分层原则明确为嵌入式场景以内存/能耗为硬约束；交互式应用将时延列为首要目标；离线系统则可妥协计算开销换取精度最大化。

1.3　技术选择与协同策略

轻量化技术的选择与协同需综合考量硬件平台特性、应用场景约束及开发资源限制，以如图 3 所示的决策流程为此提供系统指引。

实践表明：量化与剪枝组合[31]可同步压缩存储与计算开销，但需规避误差累积；蒸馏与神经架构搜索协同[32]实现学生架构的自动发现，达成知识与结构的联合优化；动态计算融合静态压缩[33]在维持自适应能力的同时降低基础复杂度。

从技术实施角度看，有效的技术组合应遵循以下一般原则：1) 优先进行结构优化，确定高效的基础架构，为后续优化奠定良好基础；2) 应用参

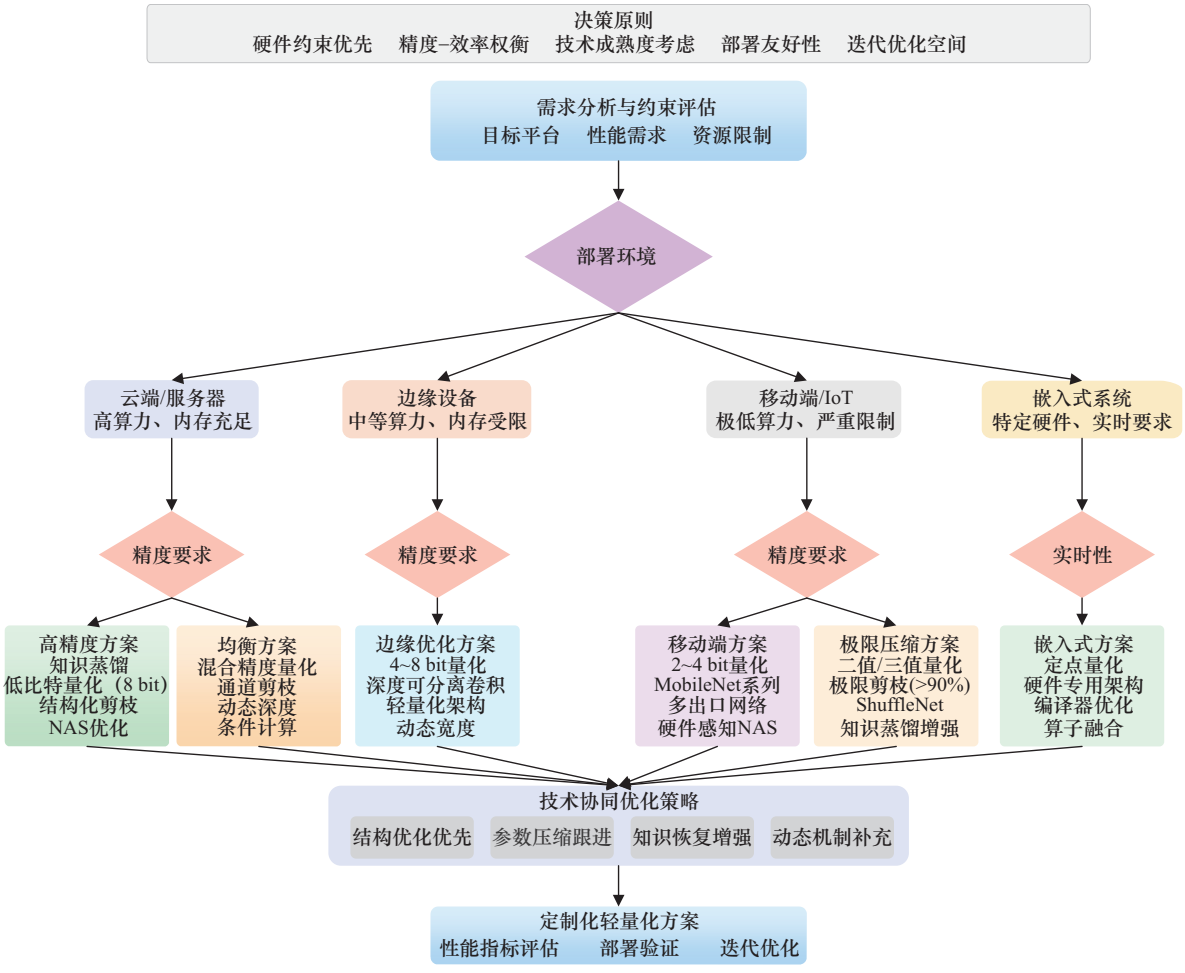

图 3　轻量化技术选择的决策流程

数压缩技术，进一步降低模型复杂度和存储需求；3) 使用知识优化方法，有效恢复压缩过程中造成的性能损失；4) 考虑引入动态计算机制，实现运行时的自适应优化。

2　静态压缩——轻量化技术的奠基时代

静态压缩技术作为神经网络轻量化的第一代方法，在此期间奠定了该领域的理论和实践基础。这一时期的技术特征是对预训练模型进行一次性优化，通过固定的压缩策略在保持模型性能的同时降低计算复杂度。本节从参数、结构和知识3个维度系统分析静态压缩技术的原理、方法和局限性。

2.1　参数空间的优化策略

2.1.1　量化技术及其演进

参数量化通过将权重和激活值从高精度数值转换为低精度数值，实现存储空间和计算复杂度的显著降低。量化过程可以视为一个优化问题。

$$W_q^* = \arg\min_{W_q} L_{\text{quant}}(W, W_q) + \gamma L_{\text{model}}(f_{W_q}(x), y) \quad (3)$$

其中，L_{quant}是量化损失，γ是正则化参数，$f_{W_q}(x)$是使用量化权重的模型预测。

量化技术历经从粗放到精细的演进：BinaryConnect[22]率先实现二值化量化（+1/–1），获得极高压缩率但伴随显著精度损失；XNOR-Net[34]和BinaryNet[35]优化训练策略后精度仍不足实用。三值化量化（±1/0）通过TWN[36]和TTQ[37]利用稀疏性提升精度–压缩平衡。低比特量化（2~8位）成为主流方案，DoReFa-Net[38]建立权重/激活/梯度统一框架，PACT[39]则通过参数化激活裁剪提升训练稳定性。训练策略分化为2条路径：量化感知训练（QAT, quantization-aware training）在训练中模拟量化效应（如LSQ[28]可学习步长机制）；训练后量化（PTQ, post-training quantization）直接压缩预训练模型（如ACIQ[40]解析裁剪、AdaRound[41]自适应舍入）。

基于上述发展，非对称量化的演变成为突破性能瓶颈的关键技术。针对对称量化偏移分布缺陷，PWLQ[42]采用分段位宽策略（4 bit MobileNet-v2提升27.42%精度），贝叶斯非对称量化[43]通过变分推理优化分区，HAQ[44]则结合强化学习实现硬件感知的混合精度分配。如表3所示，量化技术始终在压缩率与精度保持间寻求明显的权衡关系。

2.1.2　基于低秩假设的矩阵分解

低秩分解从线性代数角度实现模型压缩，其核心假设是神经网络的权重矩阵存在大量冗余，可以用低秩矩阵近似表示。对于权重矩阵$\boldsymbol{W} \in \mathbb{R}^{m \times n}$，低秩分解寻找

$$\boldsymbol{W} \approx \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}, \boldsymbol{U} \in \mathbb{R}^{m \times r}, \boldsymbol{V} \in \mathbb{R}^{n \times r}, r \ll \min(m,n) \quad (4)$$

Denton等[23]首次将奇异值分解（SVD, singular value decomposition）应用于CNN加速，通过对卷积层进行SVD实现2~3倍加速。Zhang等[45]提出了端到端的SVD压缩框架，不仅压缩了全连接层，而且还对卷积层进行了压缩。Tucker分解将高阶张量分解为核心张量和因子矩阵的乘积，Kim等[46]将其应用于CNN压缩，通过控制每个模式的秩实现灵活的压缩率。CANDECOMP/PARAFAC（CP）分解将一个高阶张量分解为若干个秩–1张量的和，Lebedev等[47]通过非线性最小二乘法计算4D卷积核的CP分解，在第2个卷积层实现4倍性能提升。

近年来低秩分解技术呈现3个发展趋势。自适应秩确定成为研究热点，Wen等[48]提出了结构化稀疏学习方法，在训练中自动确定合适的秩。与注意力机制的结合展现巨大潜力，LoRA[49]通过在注意力层引入低秩更新实现高效微调，已成为大模型压缩的标准方法。混合分解策略通过组合多种技术获得更优效果，Liu等[50]结合Tucker分解和块稀疏

表3　主要量化技术的性能对比

方法类别	代表方法	量化位宽/bit	ImageNet Top-1精度损失	压缩率	推理加速比	硬件友好性
二值化	XNOR-Net[34]	1	11.0%	32倍	58倍	极高
三值化	TWN[36]	2	3.9%	16倍	16倍	高
低比特（uniform）	DoReFa-Net[38]	4	1.8%	8倍	4倍	高
低比特（learned）	LSQ[28]	4	0.8%	8倍	4倍	中
混合精度	HAQ[44]	2~8	0.5%	6倍	3.5倍	中
非对称量化	PWLQ[42]	4	0.9%	8倍	4.2倍	中

化，在视觉任务中取得显著压缩效果。

2.2 网络结构的稀疏化方法

2.2.1 网络剪枝演进

网络剪枝通过移除冗余连接对网络结构进行简化，其理论基础源于彩票假设[51]，密集网络中存在稀疏子网络，能够达到与原网络相当的性能。剪枝方法根据粒度可分为非结构化和结构化剪枝两类。

非结构化剪枝在权重级别进行稀疏化，Han等[4]的深度压缩通过迭代剪枝–重训练实现了AlexNet 9倍压缩。然而，非结构化剪枝在标准硬件上难以加速，因此结构化剪枝成为实用化的主流方向。Li等[19]提出了基于滤波器范数的通道剪枝，通过移除整个卷积核实现规则稀疏。ThiNet[52]通过最小化特征图重建误差选择保留通道，在VGG-16上实现16.63倍加速且精度损失仅0.52%。

动态剪枝策略突破了静态方法的局限。DepGraph[53]将图论引入剪枝领域，通过显式建模层间依赖关系指导剪枝过程，避免了结构破坏导致的性能劣化。权重卷积核混合剪枝[54]剪除对卷积神经网络整体精度贡献较小的卷积核。此外，对剪枝过的模型再进行权重剪枝实现进一步的模型压缩，并引入误剪恢复机制提升可靠性。ManiDP[55]从流形学习角度切入，将实例级别的流形信息引入剪枝决策，在ResNet-34上实现55.3%计算量压缩，仅损失0.57%精度。图4展示了不同剪枝策略的对比，说明了从局部独立到全局协同的技术演进。

2.2.2 轻量化架构设计范式

与剪枝的减法思路不同，轻量化架构设计从根本上重新思考卷积操作的实现方式。深度可分离卷积成为移动端网络的标准模块，MobileNet-V1[24]将标准卷积分解为深度卷积和逐点卷积，计算量降低10%~12%。MobileNet-V2[56]引入倒残差结构和线性瓶颈层，解决了低维特征的信息损失问题。MobileNet-V3[57]结合硬件感知的神经架构搜索和轻量级注意力模块，在精度和效率上达到新的平衡。

分组卷积通过限制通道间连接降低计算复杂度。ShuffleNet-V1[58]创新性地引入通道重排操作，解决了分组卷积信息流通受限的问题。ShuffleNet-V2[59]基于内存访问代价（MAC, memory access cost）和并行度分析，提出了高效网络设计的4条准则，进一步提升了实际推理速度。

轻量化bottleneck设计通过“压缩–处理–扩展”模式优化计算效率。倒残差块先扩展通道数再压缩，更好地保留了特征信息。Squeeze-and-Excitation（SE）模块[60]将通道注意力引入bottleneck，通过建模通道依赖性实现自适应特征重标定。

图5对比了轻量化架构核心模块设计策略：深度可分离卷积解耦空间/通道维度降低计算量，分组卷积限制通道连接减少参数，倒残差结构通过“扩展–处理–压缩”模式平衡表达力与复杂度。ConvNeXt[61]采用大核卷积替代小卷积堆叠，在现代化训练策略下超越Transformer性能。设计特性

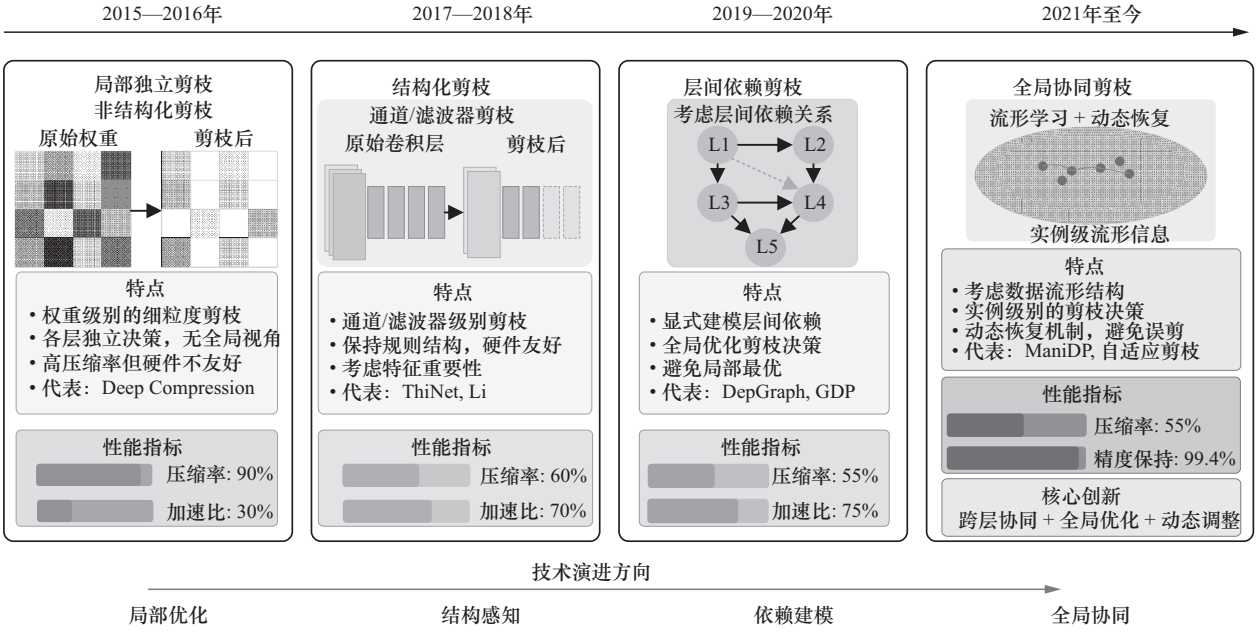

图4 剪枝策略的技术演进从局部独立到全局协同

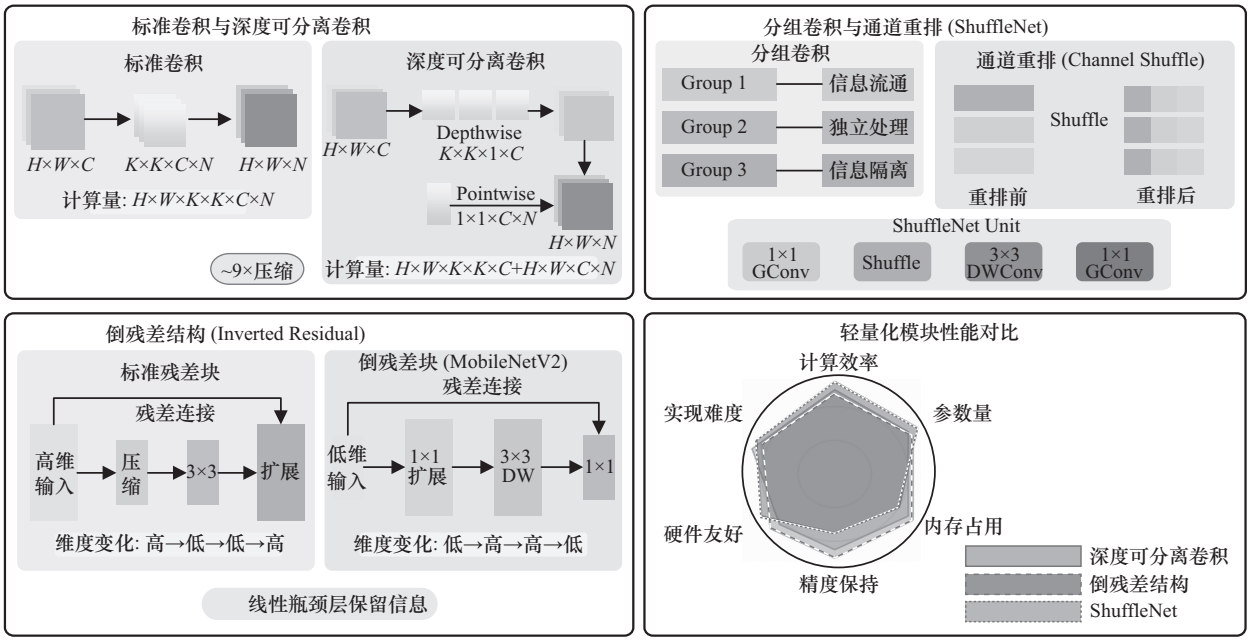

图5　轻量化架构核心模块对比

与性能总结如表4所示。

2.3　基于知识传递的模型压缩

知识传递作为模型压缩的重要范式，通过信息的有效迁移实现大模型向小模型的转化，借助软目标分布学习实现性能–效率平衡[26]。如图6所示，依据师生交互模式知识蒸馏可分为3类范式，由于在线蒸馏和自蒸馏突破了对预训练模型的依赖，实现协同演化和自我进化，本节暂不做介绍。

离线蒸馏采用预训练教师指导的经典范式，其技术演进由知识表示形式驱动。响应式知识直接匹

表4　　**轻量化架构设计方法对比**

架构系列	核心创新	参数量	FLOPS	ImageNet Top-1	目标场景
MobileNet-V1[24]	深度可分离卷积	4.2×10^6	569×10^6	70.6%	移动端
MobileNet-V2[56]	倒残差+线性瓶颈	3.4×10^6	300×10^6	72%	移动端
MobileNet-V3[57]	NAS+SE+h-swish	5.4×10^6	219×10^6	75.2%	移动端
ShuffleNet-V1[58]	分组卷积+通道重排	1.8×10^6	140×10^6	67.4%	极低算力
ShuffleNet-V2[59]	MAC 感知设计	2.3×10^6	146×10^6	69.4%	极低算力
EfficientNet-B0[62]	复合缩放	5.3×10^6	390×10^6	77.3%	通用

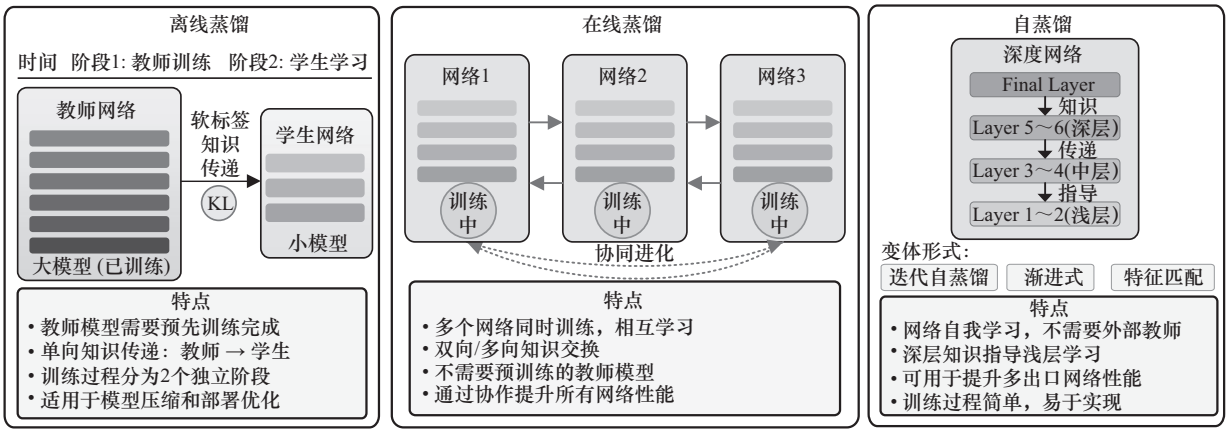

图6　知识蒸馏的3种范式:离线、在线与自蒸馏

配输出概率分布：Kim 等[63]通过序列级蒸馏实现机器翻译 4 倍加速，钟锐等[64]则提出了分层解耦目标类与非目标类知识；对于特征式知识对齐中间层表示，FitNets[65]直接匹配特征，而 AT[66]通过空间注意力转换规避维度约束；对于关系式知识保持样本间结构化关系的处理，RKD[67]保留几何关系，CRD[68]最大化互信息，TADML[69]更扩展至多模型对抗互学习。

多教师蒸馏利用知识多样性增强泛化能力，LFME[70]针对长尾问题设计了分工式专家网络，在 ImageNet-LT 上将 ResNet-50 精度提升 4.3%；AEKD[71]实现样本难度感知的权重分配；教师助理[72]通过中介模型分解师生容量鸿沟。知识合并技术解决多任务冗余问题，Shen 等[73]通过特征对齐与紧凑表示使单模型掌握多教师专长，在 Office-Home 四域任务中达到集成模型 95.3% 性能提升且存储需求降至 25%。

2.4 多技术协同优化策略

静态压缩技术的协同应用展现出超越单一方法的优化潜力。量化与剪枝联合面临的核心挑战是误差协同效应，CLIP-Q[74]提出了并行优化策略实现互补：剪枝识别冗余结构，量化压缩保留参数。其关键在于理解技术交互机制——剪枝改变权重分布特性，影响量化策略选择。基于此，APQ[75]设计了自适应框架，依层级重要性动态调整压缩强度：关键层保留高精度/多通道，冗余层采用激进压缩。在此框架中，知识蒸馏发挥性能恢复作用：QKD[76]统一量化感知训练与蒸馏损失，Joint-KD[77]则验证中间状态的迁移价值。

此类协同机制的自然延伸催生复合优化范式，Deep Compression[4]通过序列化策略实现技术级联——剪枝集中权重分布，量化利用集中特性，编码技术挖掘稀疏潜力。低秩分解的融合进一步拓展维度：Hierarchy BPE[78]针对 Transformer 注意力矩阵设计层次化压缩，以低秩约束捕获冗余并压缩嵌入长度。自适应组合代表工作 AutoCompress[79]将技术选择、参数配置与执行顺序建模为马尔可夫过程，通过强化学习发现非直观高效方案，突破单方法性能极限。

2.5 静态压缩方法的局限分析

静态压缩的根本局限源于其“一次优化、永久固定”范式与现实动态需求的本质冲突：统一压缩策略忽视样本复杂度差异（如简单图像与复杂场景的资源需求鸿沟），而 DynaBERT[80]证实了按查询复杂度调整容量的巨大潜力；优化目标的单一性也难以平衡精度/时延/能耗等多维约束，Mao 等[81]的多目标尝试仍受制于帕累托前沿复杂性。

其局限体现为四重矛盾：压缩率极限受任务复杂度制约，Lottery Ticket Hypothesis[82]虽证实稀疏子网络存在性但搜索代价高昂；专家依赖困境在 NAS-BERT[83]中凸显——自动搜索持续超越手工设计；DynamicViT[84]证明架构兼容障碍要求特定优化；CompressedLottery[85]发现可解释性缺失使压缩保留关键激活的机制不明。

上述局限正驱动轻量化技术的三重范式跃迁：从固定压缩到自适应计算、从手工设计到自动搜索以及从离线优化到在线动态调整。神经架构搜索实现全流程自动化，动态网络提供运行时灵活性，共同突破静态压缩方法的固有边界，开启了轻量化技术的自动化时代。

3 自动架构搜索——轻量化设计的智能化变革

神经架构搜索标志着深度学习从手工设计向自动化设计的重要转变。在轻量化领域，NAS 通过将资源约束纳入搜索目标，实现了精度与效率的自动平衡，开创了模型设计的新范式。与静态压缩的分维度优化不同，NAS 将结构设计与参数配置视为统一的搜索空间，通过自动化方法同时探索网络拓扑和数值配置的最优组合。本节重点阐述这种参数–结构联合优化的方法论基础、技术实现和关键突破，知识维度则主要体现在搜索效率提升的辅助机制中。

3.1 神经架构搜索的方法论基础

如图 7 所示，NAS 理论框架包含 3 个核心组件，搜索空间定义了可能的网络架构集合，搜索策略决定了如何高效地探索这个空间，而性能评估则提供了架构搜索优劣的判断标准，共同构成了自动化架构设计的完整流程。

3.1.1 搜索空间的构建与表示

搜索空间定义了 NAS 可探索的架构集合。NASNet[11]提出了 cell-based 设计，通过搜索基本单元结构并堆叠形成完整网络，将搜索空间从 10^{20} 量级压缩到 10^{10} 量级。FBNetV2[86]采用更激进的策

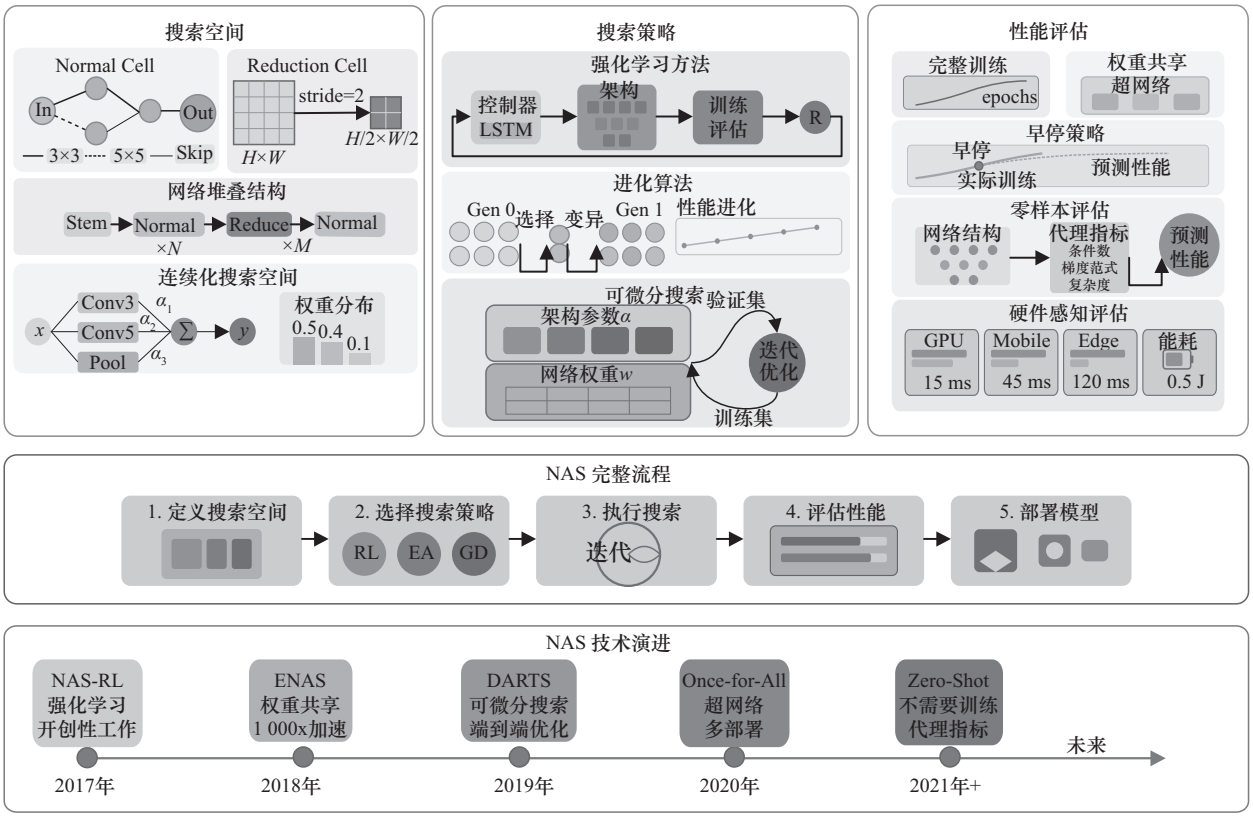

图7　神经架构搜索理论框架

略，通过维度搜索允许每层独立选择输入输出维度、扩展比例等超参数，极大地扩展了设计自由度。MixNet[87]的创新体现在操作粒度上，在同一层组合3×3、5×5、7×7等不同尺寸卷积核，打破了传统单一卷积核的假设。

搜索空间的表示方法从离散走向连续，离散表示将架构编码为操作选择序列，适用于强化学习和进化算法，但面临组合爆炸和梯度无法回传的问题。为解决这一问题，DARTS[25]通过Softmax实现操作的连续松弛

$$\bar{o}^{(i,j)}=\sum_{o\in O}\frac{\exp\left(\alpha_o^{(i,j)}\right)}{\sum_{o'\in O}\exp\left(\alpha_{o'}^{(i,j)}\right)}o(x) \tag{5}$$

使架构参数α可通过梯度下降优化。与此不同，Once-for-All[88]提出了另一种思路：训练包含所有可能架构的超网络，通过渐进式收缩提取满足不同资源约束的子网络，实现“一次训练、多次部署”的目标。

3.1.2　搜索策略的算法演进

NAS的搜索策略经历了从黑箱优化到梯度优化的演进。强化学习方法将架构构建过程建模为马尔可夫决策过程，NASNet使用长短期记忆（LSTM, long short-term memory）网络控制器生成架构描述，通过子模型的验证精度作为奖励信号，使用近端策略优化（PPO, proximal policy optimization）算法更新控制器参数，MnasNet引入多目标奖励函数实现多维度优化。

可微分搜索将离散选择转化为连续优化，DARTS形式化为双层优化问题，内层优化网络权重，外层优化架构参数。PC-DARTS[89]通过部分通道采样缓解操作选择偏差，SNAS[90]使用Gumbel-Softmax重参数化技巧实现单路径前向传播。

进化算法在多目标优化中表现优异，AmoebaNet[91]通过锦标赛选择和正则化进化在ImageNet上超越人工设计，EZNAS[92]使用遗传编程发现零成本架构评估代理指标，如雅可比矩阵的条件数、梯度范数等零成本指标，可在网络未训练时预测其潜在性能。

3.2　高效神经架构搜索关键技术

3.2.1　超网络与权重共享

超网络训练标志着NAS效率革命的起点。ENAS[93]的核心贡献在于论证参数共享机制的有效

性——通过训练包含所有候选架构的单一超网络，能够大幅降低传统独立训练架构的计算成本。如图 8 所示，该机制中不同子架构通过激活超网络内的特定路径实现功能，训练阶段随机采样子架构共享权重避免重复计算，评估阶段则直接提取冻结参数进行性能评测。

权重共享的有效性依赖于相对性能排序保持一致的假设，参数量较大的架构因获得更多梯度更新机会而形成不公平竞争。解决方案包括均匀采样平衡训练机会、操作级别 dropout 抑制共适应效应和多阶段训练逐步固化架构决策。

扩展超网络规模需精细化设计，BigNAS[94]训练包含 1 021 个架构的超网络实现三大突破：弹性分辨率输入（128~1 024 像素）适应多尺度需求，整层 dropout 支持可变深度，混合精度训练优化内存开销；其渐进收缩方法从完整网络逐步裁剪至目标架构。训练策略采用“三明治规则”，每个 batch 同步优化最大、最小及随机采样的架构以保证极端性能边界。

3.2.2　硬件感知与资源约束优化

现代硬件的复杂性导致 FLOPS 指标失去可靠性，相同 FLOPS 的网络实际性能可能相差数倍，这种差异源于内存访问模式、并行度限制及特殊指令支持等底层因素。ProxylessNAS[29]开创性地引入硬件时延作为直接优化目标，构建操作时延查找表实现快速评估。HADE[95]则基于处理器建模技术发展资源感知时延公式，完成跨平台运行时延预测。

硬件资源约束的引入从根本上重新定义了神经架构搜索的问题范畴，面对极端资源受限的应用场景，MCUNet[96]专门针对内存极度受限的微控制器（<256 KB SRAM）设计搜索策略，通过内存感知的架构设计在 ImageNet 子集上达到 70.7% 准确率。多约束优化需平衡内存/时延/能耗目标，其中内存作为硬约束通过架构剪枝保障，时延由操作选择优化，能耗则依赖动态调节机制适应。在此背景下，软硬件协同设计理念的兴起打破了传统计算系统中算法与硬件之间的抽象层次界限。DANCE[97]将硬

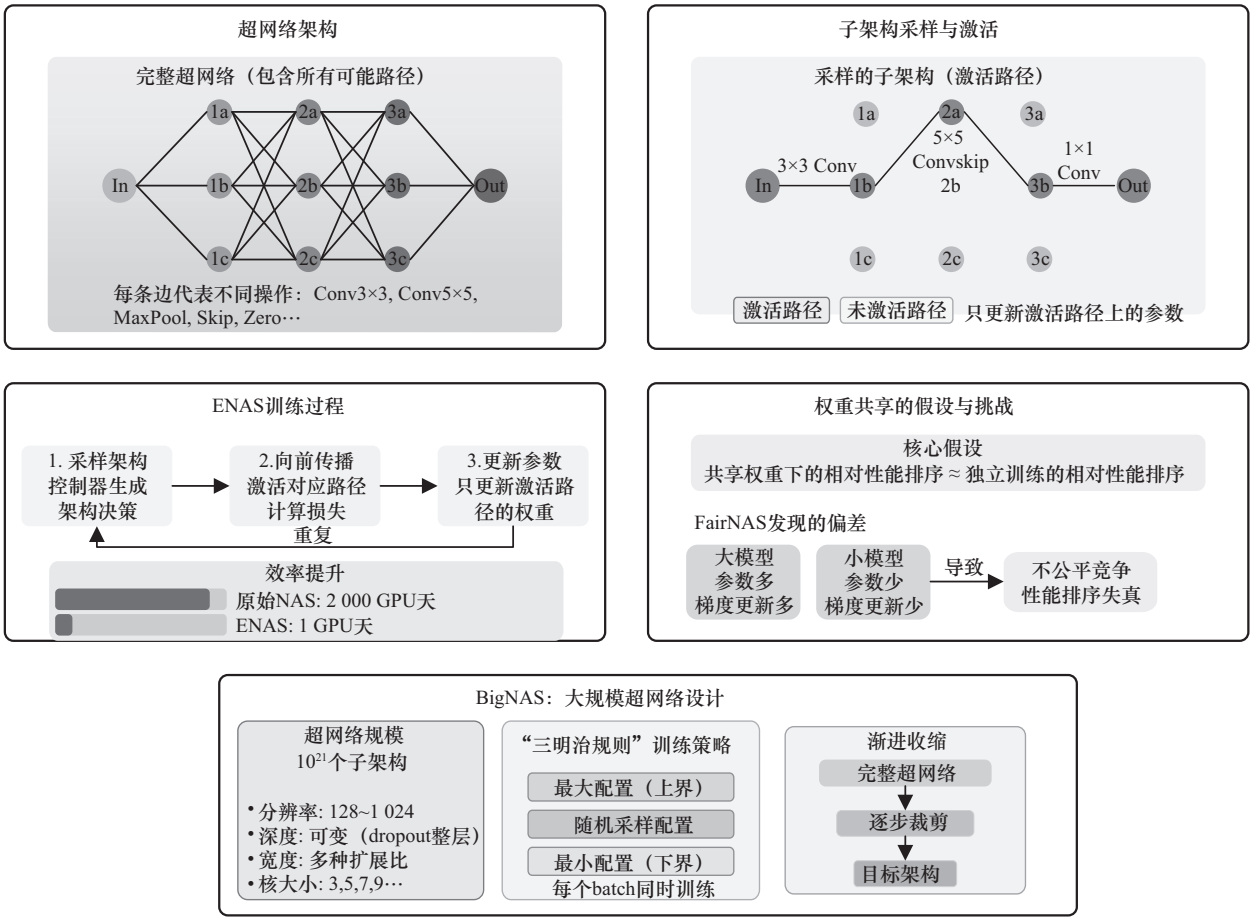

图 8　超网络与权重共享机制

件架构参数纳入搜索空间，实现算法与硬件的联合优化，获得2~3倍能效提升。BitFusion[98]设计了支持灵活比特宽度的硬件和相应的混合精度网络，通过bit-level数据流设计实现不同精度操作的硬件复用，揭示了算法–硬件co-design的巨大潜力。

3.2.3　迁移学习加速搜索

搜索知识的复用是提高效率的重要途径。通过维护跨任务架构种群，新任务可基于历史优秀架构信息进行初始化，大幅减少搜索时间。其关键在于识别可迁移的架构模式——虽然深度、宽度等全局属性具有任务特定性，但残差连接、深度可分离等局部模式往往跨任务有效，其对比如表5所示。

跨域迁移面临更大挑战。T-NAS[99]通过特征对齐实现机器视觉（CV, computer vision）到自然语言处理（NLP, natural language processing）的架构迁移，定义领域无关的架构特征，学习特征变换矩阵对齐不同域的特征空间，在目标域微调迁移的架构。实验表明，跨域迁移可减少60%的搜索时间。

元学习提供了更原则性的迁移框架。Meta-NAS[100]将架构搜索形式化为元学习问题，内循环在具体任务上搜索，外循环学习跨任务搜索策略。通过MAML框架学习搜索空间先验、评分函数初始化等元参数，新任务上只需5~10个架构评估即可找到近优解，将搜索成本降至小时级别。

3.3　自动搜索的局限性分析

尽管NAS的搜索成本有所下降，但对资源受限的研究者仍过高，完整搜索需数十天且迭代成本昂贵。其核心问题在于搜索空间设计悖论：优秀空间需领域知识，但NAS目标恰是减少对此依赖，当前空间实为人类知识的编码，真正创新仍靠人工。理论缺失严重制约发展：搜索空间持续产出优化架构的机制、权重共享收敛性、架构泛化能力评估等核心问题仍不明晰，虽有部分收敛性分析（如DARTS），但普适性理论保证与可解释性仍缺乏，阻碍洞察泛化。在实际部署中，NAS所得结构常过于复杂且不规则，传统编译器难以优化，硬件建模误差导致搜索与真实性能差距显著，框架差异加剧了不确定性。NAS的发展需在理论框架（理解设计原理）、高效搜索（实现交互）、增强人机协作（透明工具）、标准化部署流程等方向突破，以解决根本挑战，提升适应能力。尽管NAS在自动化设计方面取得了突破，但其产出的静态架构仍无法适应输入的动态变化。这促使研究者探索更加灵活的计算范式——让网络能够根据输入特性实时调整参数、结构和知识流动，实现真正的按需计算。

4　动态计算——轻量化技术的自适应范式

动态计算作为第三代轻量化范式，标志着从静态优化向运行时自适应的根本转变。区别于静态压缩的“一次优化”与神经架构搜索的“设计时优化”，其核心价值在于推理阶段按输入特性实时调整计算策略，突破静态压缩方法处理异构输入的效率瓶颈，并为资源受限场景部署开辟新路径。

该范式植根于计算资源需求的输入异质性：简单样本的完整前向传播存在显著冗余，而复杂样本需充分调用网络表达能力。这种本质差异驱动研究者探索条件优化框架下的灵活执行模式。从理论角度看，动态计算可以形式化为条件优化问题。

$$f(x;\theta,\phi)=\sum_{i=1}^{L} g_i(x;\phi)h_i(x;\theta_i) \tag{6}$$

其中，$g_i(x;\phi)$为门控函数，决定第i个计算单元的激活状态，$h_i(x;\theta_i)$为具体的计算操作，ϕ为门控参数。本节系统阐述动态计算技术如何实现参数、结构和知识3个维度的运行时协同——动态网络结构实现拓扑的实时调整，条件参数生成完成数值的输入自适应，动态知识传递支撑信息的灵活流动，最终在条件计算框架中达成三者的深度融合。

4.1　动态网络结构的自适应机制

动态网络结构通过在推理时调整网络的拓扑结构实现计算的灵活性。如图9所示，动态网络

表5　高效搜索技术对比

技术类别	代表方法	加速比	精度损失	适用场景	关键限制
权重共享	ENAS[93]	1 000倍	1%~2%	快速原型	训练偏差
硬件感知	ProxylessNAS[29]	20倍	<1%	目标部署	平台依赖
资源约束	MCUNet[96]	50倍	2%~3%	边缘设备	搜索空间
迁移学习	MetaNAS[100]	10~50倍	<1%	多任务场景	域差异

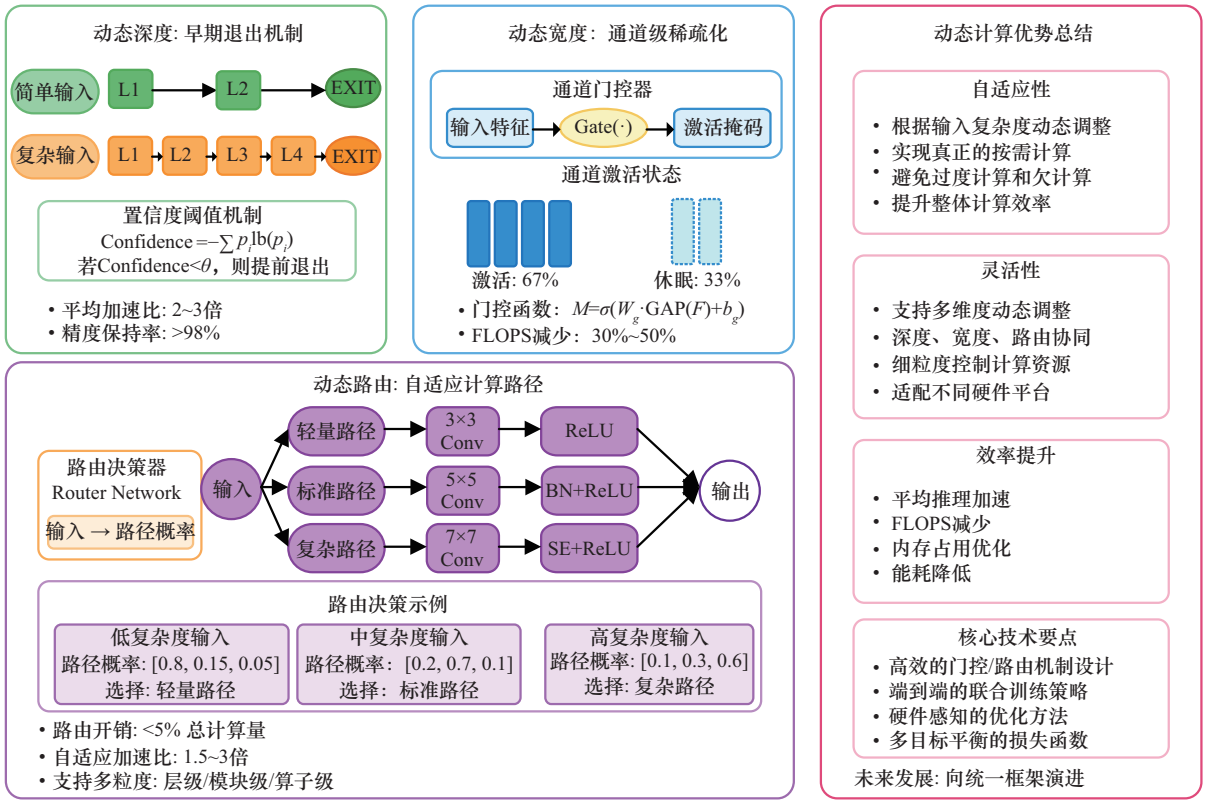

图 9　动态网络结构的自适应机制

结构的自适应机制主要包含 3 个维度：深度维度通过早期退出机制让简单样本在浅层完成推理，避免深层的冗余计算；宽度维度根据特征重要性动态调整通道数量，实现计算资源的精准分配；路由维度则为不同输入选择个性化的计算路径，最大化网络效率。3 种机制既可以独立应用，也可以协同工作。

4.1.1　动态深度调整机制

动态深度计算颠覆了传统神经网络固定层数的设计范式。早期的多出口网络研究为这一技术奠定了基础。BranchyNet[101]率先在网络的不同深度设置多个分类器，允许简单样本在浅层退出。这种设计虽然直观，但面临着训练不稳定和精度损失的挑战。

多尺度密集网络（MSDNet, multi-scale dense network）[20]提出了更加精细的解决方案。通过构建多个不同尺度的特征提取路径，MSDNet 能够在保持高精度的同时实现灵活的计算深度调整。其核心创新在于密集连接的引入，使浅层特征能够直接传递到深层，缓解了早期退出导致的信息损失问题。在 CIFAR-100 数据集上，MSDNet 在减少 50% 计算量的情况下实现了仅损失 0.5% 精度的优异表现。

随着 Transformer 架构在 NLP 领域的成功，动态深度技术也迎来了新的发展机遇。DeeBERT[102]将早期退出机制引入 BERT 模型，通过在每个 Transformer 层的输出位置添加分类头，实现了根据输入文本复杂度进行自适应推理的能力。该方法的关键技术创新在于引入了基于信息熵的置信度评估机制。

$$C_l = -\sum_{i=1}^{K} p_i^{(l)} \operatorname{lb} p_i^{(l)} \tag{7}$$

当第 l 层的预测熵 C_l 低于预设阈值 τ 时，模型即可提前退出。阈值 τ 的选择通常基于验证集的精度–效率权衡曲线确定，典型取值范围为[0.1, 0.5]，或采用可学习参数与模型联合优化。该机制在 GLUE 基准测试中实现了 2~3 倍的推理加速，同时保持了 98% 以上的精度。

4.1.2　自适应通道宽度优化

动态通道宽度技术从网络的横向维度实现计算效率优化，其发展经历了从静态到动态、从粗粒度到细粒度的演进过程[52]。通过引入通道级稀疏性

约束，实现了网络宽度的灵活调整，虽然这一早期工作主要关注静态剪枝，但其提出的通道重要性评估方法为后续动态宽度技术奠定了重要基础。

SlimmableNets[103]实现了真正意义上的动态宽度调整，通过训练支持多种宽度配置的单一网络，能够在推理时根据资源约束选择合适的执行宽度。其核心技术创新包括：1) 可切换批归一化（SBN, switchable batch normalization）为不同宽度维护独立的统计量；2) 渐进式训练策略，从最宽配置开始逐步训练窄配置；3) 知识蒸馏增强，利用宽模型指导窄模型训练。在 ImageNet 上，SlimmableNets 实现了从 0.5~2.0 倍的灵活宽度调整，精度损失控制在 2% 以内。通道门控网络（CGN, channel gating network）[104]引入了基于输入的动态通道选择机制，通过设计轻量级的门控网络，为每个输入样本生成通道选择掩码，使网络根据输入特征复杂度自适应地激活不同数量的通道。

动态双门控网络（DGNet, dynamic dual gating network）[105]进一步提升了通道选择的精细度，通过引入空间和通道门控的双重机制，实现了更加灵活的特征选择。空间门控关注特征图的空间重要性分布，通道门控则评估不同通道的贡献度，2 种门控的协同作用使网络能够同时在空间和通道维度上进行动态优化。最新的研究趋势关注硬件感知的动态宽度优化，基于 PCDARTS 改进的搜索算法[106]通过神经架构搜索自动确定每层的最优宽度配置，Once-for-All Networks[88]则提供了支持多种硬件平台的统一训练框架，将动态宽度技术从算法优化推向了实际部署，显著提升了其应用价值。

4.1.3　跨层动态路由机制

跨层动态路由代表了网络结构自适应的最高形式，通过为每个输入定制计算路径实现极致的效率优化。CapsNet[107]的动态路由算法虽然最初设计用于胶囊间的信息传递，但其迭代式路由思想为后续研究提供了重要启发。

BlockDrop[108]利用强化学习训练策略网络，动态跳过 ResNet 中不必要的残差块，在保持精度的同时减少 25% 的计算量。SkipNet[109]采用更加轻量的门控设计，通过在每个层后添加二值门控单元，SkipNet 能够动态决定是否执行当前层的计算。门控决策基于循环神经网络（RNN, recurrent neural network）的隐状态，能够捕捉层间的依赖关系。

$$g_t = \sigma(W_g h_{t-1} + U_g x_t + b_g) \tag{8}$$

其中，h_{t-1} 为 RNN 的隐状态，x_t 为当前层的输入特征。GaterNet[110]提出了全局门控的创新设计。与逐层决策不同，GaterNet 通过专门的门控网络一次性生成整个网络的路由配置。因此，全局视角使路由决策能够考虑网络的整体结构，避免了局部最优的问题。实验表明，全局门控相比逐层决策能够获得更好的精度–效率权衡。

动态卷积[111]将路由思想扩展到算子级别。通过为每个卷积层维护多个卷积核，并根据输入动态组合，实现了更细粒度的计算自适应。

$$y = \sum_{i=1}^{K} \pi_i(x)\,\mathrm{Conv}_i(x) \tag{9}$$

其中，π_i 为注意力权重，Conv_i 为第 i 个卷积核。Resolution Adaptive Networks[112]探索了多分辨率的动态路由。通过构建处理不同分辨率输入的并行路径，并根据输入复杂度选择合适的分辨率，实现了计算量与输入难度的自适应匹配，在视频理解等任务中展现出了巨大潜力。Transformer 架构的兴起为动态路由带来了新的机遇，DynamicViT[84]通过 Token 级别的动态稀疏化实现了高效的视觉 Transformer。基于 Token 重要性评分，DynamicViT 能够在保留关键信息的同时丢弃冗余 Token，性能对比如表 6 所示。

4.2　动态参数生成与动态网络配置

条件参数生成技术通过根据输入特征动态生成或调整网络参数，实现了更加灵活的模型适应能力。这种方法突破了传统网络固定参数的限制，为每个输入样本定制专属的计算参数。

4.2.1　基于元学习的参数生成

条件参数生成的理论基础可以追溯到元学习领域。HyperNetworks[113]首次提出了使用一个网络（超网络）生成另一个网络（主网络）的参数。使网络参数能够根据任务或输入动态调整。

$$\theta = h_\phi(z) \tag{10}$$

其中，h_ϕ 为超网络，z 为条件输入（如任务描述、输入特征等），θ 为生成的主网络参数。

Dynamic Filter Networks[114]将这一思想应用到视觉任务中。通过为每个空间位置生成专属的卷积核，实现了空间自适应的特征提取。

$$F_{\mathrm{out}}(i,j) = \sum_{m,n} K_{ij}(m,n)\,F_{in}(i+m, j+n) \tag{11}$$

表6 主要动态网络结构方法的性能对比

方法类别	代表方法	动态维度	平均加速比	精度保持率	额外参数开销
动态深度	BranchyNet[101]	深度	2.1倍	94.8%	11.3%
	MSDNet[20]	深度+尺度	2.7倍	99.2%	13.6%
	DeeBERT[102]	深度	3.2倍	97.6%	4.7%
动态宽度	SlimmableNets[103]	宽度	1.8~3.7倍	98.3%	—
	CGN[104]	通道	2.4倍	97.1%	2.8%
	DGNet[105]	空间+通道	2.6倍	98.5%	6.4%
动态路由	BlockDrop[108]	层级跳跃	1.3倍	99.7%	1.9%
	SkipNet[109]	层级跳跃	1.6倍	99.3%	1.2%
	DynamicViT[84]	Token路由	2.2倍	99.1%	1.2%

其中，K_{ij}为位置(i,j)处动态生成的卷积核，显著提升了网络对空间变化的建模能力，在视频预测、立体匹配等任务上取得了优异效果。

CondConv[115]提出了更加高效的条件卷积设计，通过学习少量基础卷积核的线性组合系数，CondConv在保持参数效率的同时实现了动态参数生成。

$$W=\sum_{i=1}^{n}\alpha_i(x)W_i \tag{12}$$

其中，α_i为输入相关的组合系数，W_i为第i个基础卷积核。在MobileNet和ResNet上的实验表明，CondConv能够以极小的额外开销（<10%参数增量）带来显著的性能提升（1%~2%精度提升）。

4.2.2 动态架构参数化

近期研究将条件参数生成技术扩展到了网络架构参数的动态调整领域。动态卷积[111]通过注意力机制动态组合多个卷积核参数，实现了算子级别的条件计算能力。在保持模型整体容量不变的前提下，显著提升了网络的特征表达能力和建模性能。WeightNet[116]提出了更加通用的权重生成框架体系，通过设计专门的权重预测网络模块，能够根据当前输入特征的特性动态生成整层的权重参数，从而实现了参数级别的自适应调整机制。

4.3 动态知识传递机制

动态知识传递机制将蒸馏范式从静态师生框架扩展至运行时自适应领域，实现知识在网络执行中的灵活流动。区别于第3节的离线蒸馏，该机制强调知识的实时生成、选择与应用，为动态网络提供性能保障。如图10所示，其形成4类核心机制：在线协同学习实现多网络实时知识交换；自蒸馏增强利用网络内部层次/时序知识；运行时知识路由按输入特征动态选择知识源；跨模态动态融合整合异构模态知识。

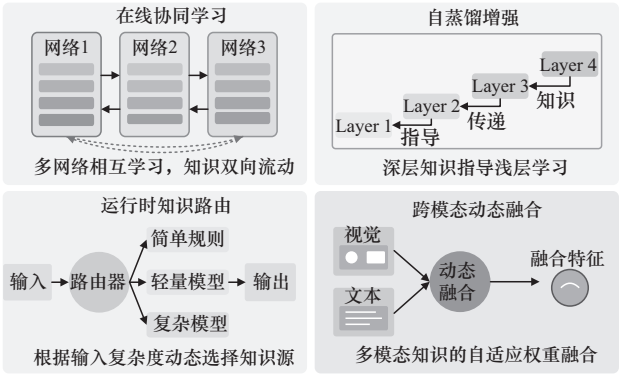

图10 动态知识传递的4种机制

4.3.1 在线蒸馏与自蒸馏

在线蒸馏与自蒸馏构成动态知识传递的双引擎。在线蒸馏颠覆传统串行训练模式，建立教师–学生协同进化机制。DML[117]使多个网络互为师生，在CIFAR-100数据集上将ResNet-32精度从69.94%提升至71.62%，其增益源于正则化效应与解空间多样性探索。OKDDip[118]扩展该框架，通过正交约束使4个ResNet-32集成精度达74.89%，超越ResNet-110教师模型（73.73%）。分层协同学习优化大规模网络训练效率，KDCL[119]则引入对比学习增强特征判别性，不仅最大化网络间的一致性，还通过负样本增强了特征的判别能力。

自蒸馏将知识生产与消费统一于单网络，Be Your Own Teacher[27]利用深层特征指导浅层学习，使ResNet在CIFAR-100数据集上精度提升1.49%并赋予自适应推理能力。数据失真引导的自蒸馏[120]

通过扰动强度分级实现知识传递，将 WideResNet 错误率降低 1.17%。在多代迭代策略中，Born-Again Networks[121]三代蒸馏使语言模型困惑度降低 5.6，PS-KD[122]动态调整目标难度，ImageNet 上 ResNet-18 精度提升 2.22% 且无额外耗时。

新兴方向持续拓展边界：数据自由蒸馏解决隐私与存储限制，DAFL[123]生成伪数据训练无损学生模型，DFQN[124]实现无数据极低比特量化；跨模态蒸馏[125]打通 RGB-深度等模态壁垒；动态蒸馏[31]自适应调整温度参数与损失权重，消除超参数调优负担。

4.3.2　动态知识路由

动态知识路由技术通过智能化的知识选择和传递机制，实现了根据任务需求和环境变化自适应调整知识流向的能力。局部模块组构（LMC, local module composition）[126]提出了模块化的局部路由方法，通过模块特定的路由机制实现自动任务推断和知识整合，支持独立训练模块的无缝合并，在持续学习场景下有效缓解了灾难性遗忘问题。动态知识路由网络（DRKN, dynamic knowledge routing network）[127]将知识路由应用于对话系统，通过考虑候选关键词之间的语义知识关系进行精确的话题预测，相比基线方法在成功率上提升超过 20%，平滑度得分提高 0.6 以上。

在网络优化领域，消息传递深度强化学习（MPDRL, message passing deep reinforcement learning）[128]创新性地将图神经网络引入深度强化学习框架，通过拓扑结构中链路间的消息传递过程提取可利用知识，在 3 个互联网服务提供商（ISP, Internet service provider）网络拓扑上实现了更优的负载均衡性能。Network Tomography with ML[129]在部分网络拓扑未知和路由动态变化的条件下，通过机器学习方法准确估计网络性能指标，相比传统方法显著提升了估计精度。研究表明，动态知识路由技术通过灵活的知识选择、传递和组合机制，能够有效应对复杂多变的应用场景。

4.3.3　跨模态动态知识融合

跨模态动态知识融合技术通过整合不同模态间的互补信息，实现了更丰富的知识表示和更准确的推理能力。跨模态动态蒸馏（CMDD, cross-modal dynamic distillation）[130]首次实现了跨模态的动态知识传递，动态对齐机制根据输入质量和模态可用性自适应调整跨模态知识权重，在视觉–语言任务上实现了 3.2% 的性能提升。VeXKD[131]提出了一个通用框架，将跨模态融合与知识蒸馏相结合，通过在特定特征层应用知识蒸馏使单模态学生网络获得跨模态洞察而不增加推理开销。C2KD[132]针对跨模态知识蒸馏中的模态差距问题，设计了定制化的双向蒸馏机制和动态样本选择策略，通过代理网络渐进式传递跨模态知识，有效缓解了模态失衡和软标签错位问题。

在复杂认知任务中，跨模态融合展现出了独特优势：OracleSage[133]将层次化视觉理解与图神经网络语义推理相结合，通过动态消息传递机制捕获视觉组件与语义概念间的深层关系；FusionRM[134]利用文本语义的表达能力弥合视觉和语言模态间的知识鸿沟，通过融合隐式视觉知识和显式文本知识创建统一的语义嵌入空间；CMC[135]采用两阶段训练策略，先通过跨模态知识蒸馏捕获模态间相关性，再进行特征融合优化，充分挖掘多模态信息的互补价值。

在面向实际应用的跨模态融合系统中，CMF-CNN[136]针对工业场景设计了模态特定和跨模态知识共享的并行架构，通过在线软标签训练增强了系统在复杂环境下的稳定性。IC-MKD[137]创新性地处理了不完整模态问题，通过信息解耦和相互知识蒸馏机制，使缺失部分模态的场景也能受益于完整的多模态知识。

4.4　条件计算

条件计算作为动态计算技术的集大成者，通过构建统一的理论框架体系整合了动态网络结构、条件参数生成和动态知识传递等各种核心机制，不仅实现了单一维度的性能优化，更重要的是通过多维度协同优化达到了前所未有的计算灵活性和执行效率。

4.4.1　条件计算的演进脉络

条件计算的概念最早由 Bengio 等[21]提出，其核心思想是让网络根据输入选择性地激活部分计算单元。这一开创性工作奠定了条件计算的理论基础，但早期的技术实现面临着梯度估计、训练不稳定等挑战。

随机门控机制是早期条件计算的主要实现方法，通过引入可学习的门控函数和随机噪声，研究者尝试在保持可微性的同时实现离散决策。然而，这种方法存在高方差问题，会导致训练效率低下。

后续研究通过改进梯度估计方法[137]和引入结构化稀疏性[138]，逐步提升了条件计算的实用性。

条件计算的真正突破来自专家混合（MoE, mixture of experts）系统的现代化改造，传统 MoE 存在负载不均衡、训练不稳定等问题，限制了其在深度学习中的应用。Sparsely-Gated MoE[139]通过引入稀疏门控和 Top-K 路由机制，成功将 MoE 扩展到了深度网络。这一工作不仅解决了计算效率问题，还展示了条件计算在模型容量扩展上的巨大潜力。

4.4.2　大规模条件计算系统

GShard[140]将条件计算推向了新的规模。通过精心设计的分布式架构和自动分片策略，GShard 成功训练了包含 6 000 亿参数的语言模型。其核心创新如下。

专家并行实现了 MoE 层的高效分布式训练。每个设备负责一部分专家，通过 All-to-All 通信完成 Token 的路由和聚合。这种设计将通信开销限制在可接受范围内，使超大规模模型训练成为可能。

容量因子机制解决了负载均衡问题。通过为每个专家设置容量上限，并引入辅助损失促进均匀分配，GShard 确保了训练的稳定性。

图 11 展示了专家混合系统负载均衡机制：输入 Token 首先通过门控网络计算每个专家的路由概率，然后基于容量限制进行分配；当某个专家接近容量上限时，系统会通过辅助损失函数引导后续 Token 选择其他专家，从而实现全局负载均衡；对于超出容量的 Token，系统采用溢出处理策略，既保证了计算效率，又维持了模型性能。当某个专家达到容量上限时，多余的 Token 会被丢弃，但通过随机性和辅助损失函数，这种丢弃对系统最终性能的影响被降到最低。

Switch Transformers[141]进一步简化和优化了大规模条件计算。通过将 Top-K 路由简化为 Top-1（每个 Token 只路由到一个专家），Switch Transformers 显著降低了通信开销和计算复杂度。同时，通过引入专家 dropout 和改进的初始化策略，提升了训练的稳定性。在相同的计算预算下，Switch Transformers 相比密集模型实现了 7 倍的预训练速度提升。

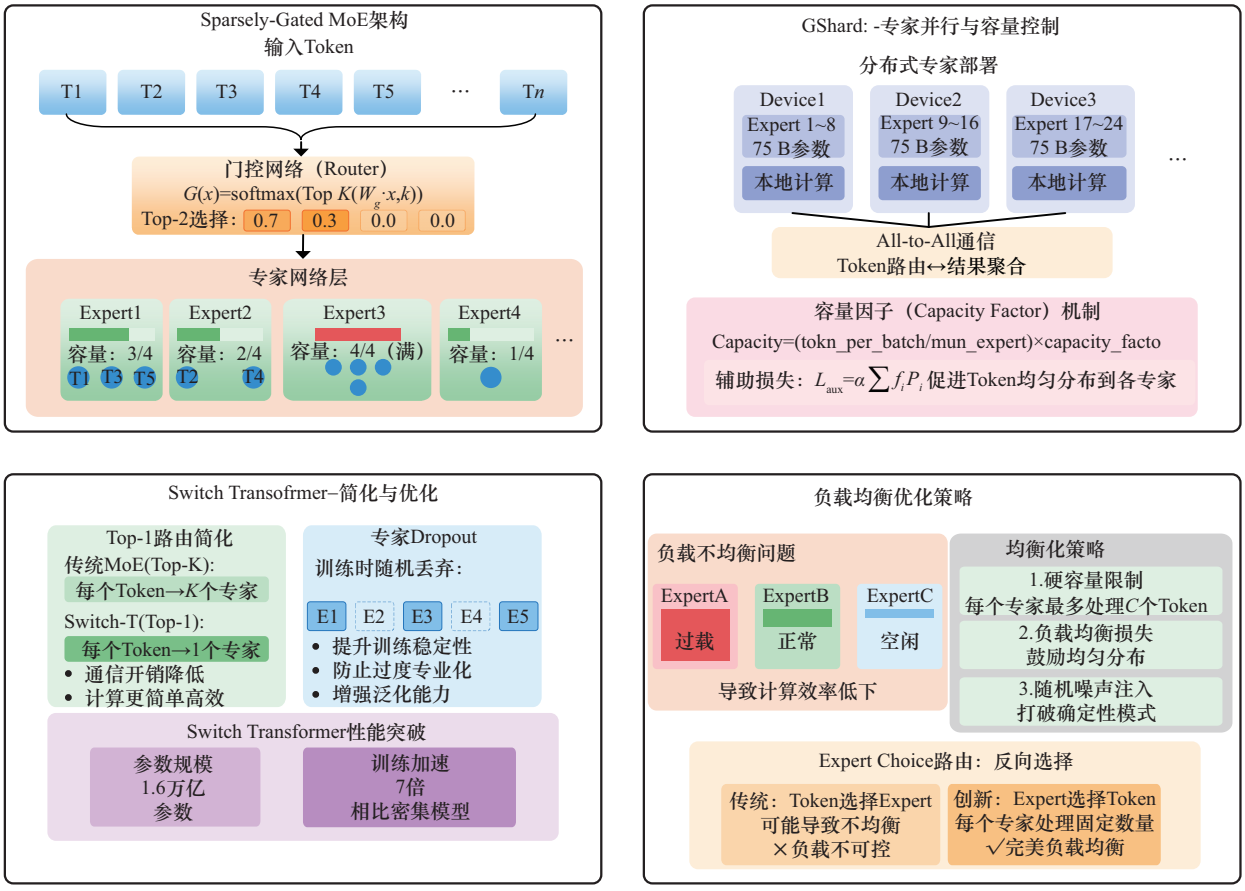

图 11　专家混合系统负载均衡机制

4.4.3　条件计算的新型架构

近期研究在条件计算架构设计方面取得了重要突破，主要体现在 3 个创新方向。MoD（mixture of depths）[142]将条件计算扩展到序列维度，允许不同位置的 Token 经过不同深度的处理，简单 Token 跳过某些层，而复杂 Token 经过完整计算。CoLT5[143]提出了条件计算长文本 Transformer，通过在前馈层和注意力层对重要 Token 分配更多计算资源，实现高效的长序列处理。Expert Choice Routing[144]颠覆传统路由方式，改为专家选择 Token，从根本上解决负载均衡问题。

传统 MoE 仅在模型维度实现稀疏化，而新一代架构设计实现了多维度的条件化：序列维度的深度选择、组件维度的功能调节和路由维度的策略优化，因此模型能够根据输入特性进行细粒度的计算调整，实现更精确的资源分配和更高的计算效率。图 12 展示了条件计算的统一框架，从输入分析到动态执行的完整流程。

4.4.4　条件计算的发展趋势

条件计算正在成为下一代神经网络的核心特征，其发展呈现出理论深化和应用拓展并行的态势。随着模型规模的持续增长和应用场景的多样化，固定计算模式的局限性日益凸显，条件计算提供了一条可持续的发展道路。从技术发展轨迹看，条件计算正从经验驱动的架构设计向理论指导的系统优化转变。

可解释条件计算通过可视化路由模式、分析专家专业化程度、追踪 Token 轨迹等方法，研究者开始揭示条件计算的内在工作原理并致力于理解和解释条件决策的机制。可解释性的研究成果有利于改进现有方法，并指导后续研究设计新型和创新的条件架构。自适应条件计算探索了条件策略的在线学习和适应，通过元学习框架，模型能够快速适应新的任务或数据分布，动态调整条件计算策略。在持续学习、少样本学习等场景中展现出巨大潜力，各框架代表工作对比如表 7 所示。

未来发展趋势表明，条件计算将朝着更加智能化和自动化的方向演进。第一，硬件感知的条件计算将成为重要发展方向，通过深度理解不同硬件平台的特性，实现硬件与软件协同优化。第二，多模态条件计算将扩展条件决策的应用范围，根据不同模态数据的特性动态调整计算策略。第三，终身学习条件计算将实现模型在持续学习过程中的动态适应，为人工智能系统的长期演化提供技术支撑。

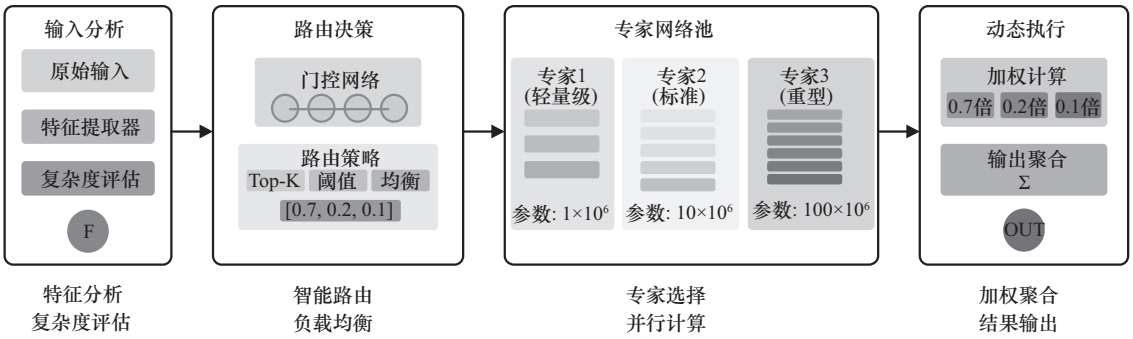

图 12　条件计算的统一框架:从输入分析到动态执行的完整流程

表 7　　主要条件计算框架的技术特征与应用分析

框架类别	代表工作	核心创新	稀疏度	扩展性	主要应用	技术挑战
混合专家系统	Sparsely-Gated MoE[139]	稀疏门控机制	87.50%	千亿级	语言建模	专家利用率
分布式 MoE	GShard[140]	自动并行分片	99.80%	6 000 亿	机器翻译	跨节点通信
简化路由	Switch Transformers[141]	单专家路由	99.75%	1.6 万亿	预训练加速	容量溢出
深度条件化	MoD[142]	层级跳跃机制	72%	百亿级	自适应推理	深度决策
长文本优化	CoLT5[143]	重要性感知计算	68%	百亿级	文档理解	Token 重要性评估
双向选择	Expert Choice Routing[144]	专家主动选择	91%	千亿级	均衡负载	路由复杂度
模块化计算	ACM	渐进式细化	76%	十亿级	多模态任务	门控优化

5 实践现状、关键挑战与未来展望

5.1 多场景的轻量化落地现状

如第1节所述，轻量化技术的落地效能本质上取决于“参数–结构–知识”三维设计空间与具体场景资源约束的匹配程度。不同部署层级因其迥异的计算、内存与功耗预算，驱动了技术选型的差异化，同时也充分印证了硬件–算法协同优化的必要性。

在终端设备层的部署中，极端资源约束下的实时感知成为核心诉求，其主要矛盾体现在参数存储与计算访存的极限压缩。微控制器（MCU, microcontroller unit）领域是这一矛盾的集中体现：通过硬件感知的神经架构搜索（结构优化）与编译–运行时协同优化，MCUNetV2[145]在32 KB SRAM内实现了ImageNet 71.8% Top-1精度；其所采用的patch-based inference本质上是一种动态结构调整策略，使激活内存压缩了4~8倍。此类进展与TinyML生态研究[146]共同确立了超低功耗场景以结构搜索为主导、参数压缩为支撑的技术栈。对于边缘计算节点（如Raspberry Pi、Jetson Nano），其能力可支持更复杂的视觉任务，技术范式也进一步延伸至知识优化与动态计算的协同。实践表明，边缘场景的轻量化本质上是参数（模型体积）、结构（网络拓扑）、知识（任务性能）与硬件资源之间的联合博弈。

云端大模型部署将核心矛盾从单一模型精度转向了吞吐量与成本优化，此举推动了知识调度和动态结构等动态计算范式成为主导。其优化体现在算子级参数压缩与专用硬件的深度协同：NVIDIA TensorRT-LLM集成了KV-Cache Early Reuse与推测解码技术[147-148]，使单卡Llama-2推理吞吐量提升至基准的3.6倍；通过定制化参数量化策略，AWS Inferentia-2在EC2 Inf1实例上实现了2.3倍吞吐增益与70%的成本削减[149]。针对稀疏大模型，动态调度机制（如HarMoEny）通过异步专家预取与负载自适应调度，使MoE（动态结构）推理吞吐提升了37%~70%[150]；MoE-Infinity则利用显存动态裁剪策略优化了单机时延稳定性[151]。可以看出，条件计算是实现云端极致效率的必由之路。

端–边–云协同架构的成熟性依赖于轻量化技术跨层级适配能力，其核心挑战在于将统一的“参数–结构–知识”框架分解并映射至异构算力资源。这就要求算法本身具备内在的动态性与可分解性。F2SC算法即为典型解决方案，通过动态分配分块推理任务与通信链路，端到端时延得以降低28%，这正是结构优化与计算流程动态调整在系统层面的体现。与此同时，匹配的专用硬件生态（如Inf1、Jetson Orin Nano等）为“云训练–边适配–端执行”的分层决策体系提供支撑，最终推动轻量化技术从孤立优化迈向覆盖硬件感知设计、编译优化、运行时调度及跨域资源管理的全栈协同演进，从而实现模型效率、系统吞吐与场景需求之间的深度契合。

5.2 近程技术演进与应用扩散

轻量化技术的近程演进清晰地体现了从“静态优化”向“动态计算”范式的全面迁移。其创新均可在“参数–结构–知识”框架内找到对应坐标，并呈现出多维度协同增强的趋势，持续拓展计算–时延–精度约束下的Pareto前沿。

在参数维度，动态稀疏训练（DST, dynamic sparse training）实现了从离线剪枝到在线演化的范式跃迁。Structured RigL方法通过约束N:M拓扑稀疏，在90%剪枝率下维持ImageNet基准精度，并使GPU推理时延降低41.2%[152]。ICLR-2025研究进一步验证DST在对抗腐蚀场景中可超越等规模稠密模型的鲁棒性，且不增加计算开销[153]。该技术在百万级标签空间实现单卡训练与推理，显存占用压缩至传统方案的37%[154]，标志着稀疏化成为端到端部署管线的基础设施。结构维度的突破以稀疏专家混合模型的动态路由机制为代表，Meta-FAIR通过解耦门控失衡、专家装载与通信放大三大瓶颈，提出负载自适应的动态路由框架，在语言模型任务中实现6~11倍吞吐增益与40%显存压缩[155]，为训练–推理全流程稀疏化提供工程范式。知识维度则呈现显式解耦趋势，TensorRT-LLM的推测解码采用草稿–验证双级流水线架构，在Llama-70B模型上达成2.8~3.6倍吞吐提升且保持文本保真度[148]，奠定了知识路由轻重分离的理论基础。

与此同时，跨模态动态计算取得重要进展，基于Nyströmformer的NiCTRAM方法以线性时间复杂度重构注意力机制，在VIS-IR跨模态重识别任务中使Rank-1准确率提升4.2个百分点，参数量削减50%[156]，实证了动态计算在夜视终端等严苛边缘场景的部署可行性。技术扩散进一步重塑协同推

理系统架构：Edge-Cloud 分层推理从静态切分演进至自适应层切分与通信联合优化，在智能驾驶感知系统中实现 40 ms 端侧时延与 40% 带宽削减[157]；工程工具链创新加速技术落地，SplitTracr 构建分割–效能量化评估体系[158]，SplitEE 通过层级早退机制将自然语言推理计算深度压缩 38% 而精度损失控制在 0.4%[159]，MBRL 驱动的 Transformer 动态调度算法在无线边缘场景维持 QoS 前提下降低 27% 时延[160]。

综上，动态稀疏化（参数维度）、专家路由优化（结构维度）与知识裁剪（计算维度）共同构成轻量化技术的近程内核，在计算–时延–精度约束下持续拓展 Pareto 前沿，为硬件协同设计提供理论锚点。

5.3　远景突破方向与关键挑战

“参数–结构–知识”框架统摄了现有轻量化技术，然而面向终身学习、隐私保护等远景需求，该框架面临扩展性挑战。当前研究正经历从静态压缩向终身自适应智能体的范式跃迁，这一进程面临三重相互耦合的基础性挑战，亟须突破现有理论边界。

终身自适应学习要求模型在开放环境中持续演化，这对现有框架提出了时序维度的新要求。联邦学习中虽已实现端侧低精度训练在非独立同分布（IID, interface identifier）数据下的全精度收敛，动态 PTQ/QAT 切换策略也带来 2.4 倍训练加速，但开放环境中的概念漂移缺乏可微数学表征，导致现有量化–蒸馏框架难以适配设备异构性。当“分群蒸馏+按需精度”策略部署于终端设备时，量化误差累积引发跨设备诊断结论差异高达 15% 的模型分裂风险，揭示出环境感知稳定性理论的缺失——亟须建立漂移量化模型与分布式共识机制以约束演化失控。

硬件–算法原子融合面临物理瓶颈的多维复杂性挑战。存算一体器件通过 RRAM-CIM 架构达成 41 TOPS/W 能效里程碑，PUF 技术实现计算–加密原生集成，但稀疏模式、量化误差与电路映射的协同优化已被证明属于 NP-Hard 问题，现有工具链因跨层抽象断层无法逼近全局最优解。更严峻的是，模拟计算单元虽实现 10^5 级能效跃升，其固有随机噪声对注意力机制的信噪比（SNR, signal-to-noise ratio）劣化（$\Delta\mathrm{SNR} > 6$ dB）直接挑战推理可信度，迫使误差–能效联合优化理论必须与器件革新同步突破。

碳足迹与隐私的量化困局要求将轻量化框架纳入新的优化维度。通信感知量化也能隐藏客户端样本规模，但在大模型时代背景下，绿色 AI 因缺乏碳排放因子的动态计量模型而陷入节碳理论不足和自证陷阱。同时，PIM 架构的百倍能效增益与潜在差分攻击面形成本质矛盾，传统同态加密的能效瓶颈要求开发基于零知识证明的可验证隐私框架，以破解能效–隐私的零和博弈。

范式跃迁路径由此明晰，轻量化技术需升维至具备环境感知能力的有机体范式，通过流形几何稳定性理论（防范概念漂移）、NP-Hard 协同优化代数（贯通硬件–算法断层）、碳–隐私联合度量模型（锚定监管标准）重构技术内核。唯有如此，方能在可持续发展约束下实现终身自适应部署，推动人机协作范式从工具性辅助迈向自主决策协同。

5.4　未来展望

展望未来，神经网络轻量化技术正经历从经验优化向理论–系统协同的范式跃迁，其演进将呈现多维融合与自主重构的双螺旋特征。研究重心将转向构建参数–结构–知识统一优化流形，通过任务感知机制与硬件微分几何的耦合，实现跨维度技术组合的智能涌现——如在资源动态波动场景中，自主激活知识蒸馏通路与结构化稀疏的拓扑纠缠，使模型具备环境自适应的计算重构能力。这一进程将驱动全栈自动化闭环的成熟，编译器优化、运行时调度等离散决策被抽象为连续可微空间，形成具备反事实推理能力的自主决策引擎，最终达成从硬件感知建模到部署调优的零人工干预，彻底消除技术落地门槛。

在理论研究上，亟须突破现有认知边界，基于信息瓶颈理论与流形学习几何建立轻量化的性能边界方程，严格量化知识蒸馏的语义保真衰减函数与稀疏化–硬件噪声的联合误差传播机制。此类理论不仅解释压缩极限的本质，更为神经形态计算、量子计算等新兴范式提供跨硬件拓扑映射法则——将卷积操作编码为脉冲序列的时空流形，或将注意力机制重构为量子纠缠的泡利算子优化，在新型计算基座上重建轻量化技术栈。

因此，研究者需以微分几何视野贯通理论–工程断层，以拓扑兼容思维融合经典与新兴硬件范式。方能构建具备环境意识、自主进化与可信认证能力的轻量化有机体，最终实现无处不在却隐于无

形的可持续智能，为工业物联网、普适医疗、智慧农业、智慧城市等领域提供符合碳–隐私边界的决策基座。

6 结束语

本文系统梳理了神经网络轻量化技术从静态压缩、神经架构搜索到动态计算的演进脉络。基于“参数–结构–知识”统一框架，揭示了不同技术的内在协同机制：静态压缩通过量化、剪枝与蒸馏实现离线优化，但受限于固定策略；神经架构搜索突破手工设计瓶颈，实现自动化架构生成；动态计算则依据输入特性实时调整策略，达成按需计算。实践表明，轻量化技术正向多维协同与动态自适应演进。未来需构建环境感知的统一框架，深化理论、硬件协同及可解释性研究，推动技术向智能化跃迁，为边缘智能与普惠AI提供关键支撑。

参考文献：

[1] MEHONIC A, KENYON A J. Brain-inspired computing needs a master plan[J]. Nature, 2022, 604(7905): 255-260.

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

[3] ZHANG Q C, YANG L T, CHEN Z K, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42: 146-157.

[4] HAN S, MAO H Z, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J]. arXiv Preprint, arXiv: 1510.00149, 2015.

[5] CHEN J A, NIU W, REN B, et al. Survey: exploiting data redundancy for optimization of deep learning[J]. ACM Computing Surveys, 2023, 55(10): 1-38.

[6] WU D L, YANG W H, ZOU X Y, et al. Smart-DNN+: a memory-efficient neural networks compression framework for the model inference[J]. ACM Transactions on Architecture and Code Optimization, 2023, 20(4): 1-24.

[7] GONG Z Y, ZHANG H F, YANG H, et al. A review of neural network lightweighting techniques[J]. Innovation & Technology Advances, 2023, 1(2): 1-16.

[8] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[9] LIU B C, WANG D W, LV Q, et al. Towards super compressed neural networks for object identification: quantized low-rank tensor decomposition with self-attention[J]. Electronics, 2024, 13(7): 1330.

[10] DONG X Y, KEDZIORA D J, MUSIAL K, et al. Automated deep learning: neural architecture search is not the end[J]. Foundations and Trends® in Machine Learning, 2024, 17(5): 767-920.

[11] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8697-8710.

[12] TAN M X, CHEN B, PANG R M, et al. MnasNet: platform-aware neural architecture search for mobile[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 2815-2823.

[13] HAN Y Z, HUANG G, SONG S J, et al. Dynamic neural networks: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7436-7456.

[14] SUN Y, LI J, XU X. Meta-GF: training dynamic-depth neural networks harmoniously[M]//Computer Vision-ECCV 2022. Berlin: Springer, 2022: 691-708.

[15] 乔俊飞, 韩红桂. RBF 神经网络的结构动态优化设计[J]. 自动化学报, 2010, 36(6): 865-872.
QIAO J F, HAN H G. Optimal structure design for RBFNN structure[J]. Acta Automatica Sinica, 2010, 36(6): 865-872.

[16] CHEN R, SHEN H, ZHAO Z Q, et al. Global routing between capsules[J]. Pattern Recognition, 2024, 148: 110142.

[17] MENG L C, LI H D, CHEN B C, et al. AdaViT: adaptive vision transformers for efficient image recognition[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 12299-12308.

[18] JACOB B, KLIGYS S, CHEN B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2704-2713.

[19] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[J]. arXiv Preprint, arXiv: 1608.08710, 2016.

[20] HUANG G, CHEN D L, LI T H, et al. Multi-scale dense networks for resource efficient image classification[J]. arXiv Preprint, arXiv: 1703.09844, 2017.

[21] BENGIO Y, LÉONARD N, COURVILLE A. Estimating or propagating gradients through stochastic neurons for conditional computation[J]. arXiv Preprint, arXiv: 1308.3432, 2013.

[22] COURBARIAUX M, BENGIO Y, DAVID J P. BinaryConnect: training deep neural networks with binary weights during propagations[J]. Advances in Neural Information Processing Systems, 2015, 28: 3123-3131.

[23] DENTON E L, ZAREMBA W, BRUNA J, et al. Exploiting linear structure within convolutional networks for efficient evaluation[C]//Proceedings of the Neural Information Processing Systems. Massachusetts: MIT Press, 2014: 1269-1277.

[24] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv Preprint, arXiv: 1704.04861, 2017.

[25] LIU H X, SIMONYAN K, YANG Y M. DARTS: differentiable architecture search[J]. arXiv Preprint, arXiv: 1806.09055, 2018.

[26] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv Preprint, arXiv: 1503.02531, 2015.

[27] ZHANG L F, SONG J B, GAO A N, et al. Be your own teacher: im-

prove the performance of convolutional neural networks via self distillation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 3712-3721.

[28] ESSER S K, MCKINSTRY J L, BABLANI D, et al. Learned step size quantization[J]. arXiv Preprint, arXiv: 1902.08153, 2019.

[29] CAI H, ZHU L G, HAN S. ProxylessNAS: direct neural architecture search on target task and hardware[J]. arXiv Preprint, arXiv: 1812.00332, 2018.

[30] LIU S W, YIN L, MOCANU D C, et al. Do we actually need dense over-parameterization? in-time over-parameterization in sparse training[C]//Proceedings of the International Conference on Machine Learning, 2001.

[31] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究综述[J]. 计算机学报, 2022, 45(8): 1638-1673.
SHAO R R, LIU Y A, ZHANG W, et al. A survey of knowledge distillation in deep learning[J]. Chinese Journal of Computers, 2022, 45(8): 1638-1673.

[32] LIU Y, JIA X H, TAN M X, et al. Search to distill: pearls are everywhere but not the eyes[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 7536-7545.

[33] 黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624-653.
HUANG Z H, YANG S Z, LIN W, et al. Knowledge distillation: a survey[J]. Chinese Journal of Computers, 2022, 45(3): 624-653.

[34] RASTEGARI M, ORDONEZ V, REDMON J, et al. XNOR-net: imageNet classification using binary convolutional neural networks[C]//European Conference on Computer Vision. Berlin: Springer, 2016: 525-542.

[35] COURBARIAUX M, HUBARA I, SOUDRY D, et al. Binarized neural networks: training deep neural networks with weights and activations constrained to +1 or −1[J]. arXiv Preprint, arXiv: 1602.02830, 2016.

[36] LI F F, LIU B, WANG X X, et al. Ternary weight networks[J]. arXiv Preprint, arXiv: 1605.04711, 2016.

[37] ZHU C Z, HAN S, MAO H Z, et al. Trained ternary quantization[J]. arXiv Preprint, arXiv: 1612.01064, 2016.

[38] ZHOU S C, WU Y X, NI Z K, et al. DoReFa-net: training low bitwidth convolutional neural networks with low bitwidth gradients[J]. arXiv Preprint, arXiv: 1606.06160, 2016.

[39] CHOI J, WANG Z, VENKATARAMANI S, et al. PACT: parameterized clipping activation for quantized neural networks[J]. arXiv Preprint, arXiv: 1805.06085, 2018.

[40] BANNER R, NAHSHAN Y, HOFFER E, et al. Post-training 4-bit quantization of convolution networks for rapid-deployment[J]. arXiv Preprint, arXiv: 1810.05723, 2018.

[41] NAGEL M, AMJAD R A, BAALEN M V, et al. Up or down? Adaptive rounding for post-training quantization[J]. arXiv Preprint, arXiv: 2004.10568, 2020.

[42] FANG J, SHAFIEE A, ABDEL-AZIZ H, et al. Post-training piecewise linear quantization for deep neural networks[C]//Computer Vision-ECCV 2020. Berlin: Springer International Publishing, 2020: 69-86.

[43] CHIEN J T, CHANG S T. Bayesian asymmetric quantized neural networks[J]. Pattern Recognition, 2023, 139: 109463.

[44] WANG K, LIU Z J, LIN Y J, et al. HAQ: hardware-aware automated quantization with mixed precision[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 8604-8612.

[45] ZHANG X Y, ZOU J H, HE K M, et al. Accelerating very deep convolutional networks for classification and detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 1943-1955.

[46] KIM Y D, PARK E, YOO S, et al. Compression of deep convolutional neural networks for fast and low power mobile applications[J]. arXiv Preprint, arXiv: 1511.06530, 2015.

[47] LEBEDEV V, GANIN Y, RAKHUBA M, et al. Speeding-up convolutional neural networks using fine-tuned CP-decomposition[J]. arXiv Preprint, arXiv: 1412.6553, 2014.

[48] WEN W, WU C P, WANG Y D, et al. Learning structured sparsity in deep neural networks[C]//Advances in Neural Information Processing Systems. Barcelona: NIPS, 2016: 2074-2082.

[49] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[J]. arXiv Preprint, arXiv: 2106.09685, 2021.

[50] LIU Y, NG M K. Deep neural network compression by Tucker decomposition with nonlinear response[J]. Knowledge-Based Systems, 2022, 241: 108171.

[51] FRANKLE J, CARBIN M. The lottery ticket hypothesis: finding sparse, trainable neural networks[J]. arXiv Preprint, arXiv: 1803.03635, 2018.

[52] LUO J H, WU J X, LIN W Y. ThiNet: a filter level pruning method for deep neural network compression[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 5068-5076.

[53] FANG G F, MA X Y, SONG M L, et al. DepGraph: towards any structural pruning[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 16091-16101.

[54] 靳丽蕾, 杨文柱, 王思乐, 等. 一种用于卷积神经网络压缩的混合剪枝方法[J]. 小型微型计算机系统, 2018, 39(12): 2596-2601.
JIN L L, YANG W Z, WANG S L, et al. Mixed pruning method for convolutional neural network compression[J]. Journal of Chinese Computer Systems, 2018, 39(12): 2596-2601.

[55] TANG Y H, WANG Y H, XU Y X, et al. Manifold regularized dynamic network pruning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 5016-5026.

[56] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4510-4520.

[57] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 1314-1324.

[58] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.

[59] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 122-138.

[60] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.

[61] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 11966-11976.

[62] TAN M X, LE Q V. EfficientNet: rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. Long Beach: PMLR, 2019: 6105-6114.

[63] KIM Y, RUSH A M. Sequence-level knowledge distillation[C]//Proceedings of the 2016 Conference on Empirical Methods in NaturalLanguage Processing. Stroudsburg, PA, USA: ACL, 2016: 1317-1327.

[64] 钟锐, 宋亚锋, 周晓康. 分层蒸馏解耦网络的低分辨率人脸识别算法[J]. 计算机应用研究, 2025, 42(6): 1900-1908.
ZHONG R, SONG Y F, ZHOU X K. Hierarchical knowledge distillation decoupling network for low-resolution face recognition algorithm[J]. Application Research of Computers, 2025, 42(6): 1900-1908.

[65] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[J]. arXiv Preprint, arXiv: 1412.6550, 2014.

[66] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer[J]. arXiv Preprint, arXiv: 1612.03928, 2016.

[67] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 3962-3971.

[68] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive representation distillation[J]. arXiv Preprint, arXiv: 1910.10699, 2019.

[69] 赖轩, 曲延云, 谢源, 等. 基于拓扑一致性对抗互学习的知识蒸馏[J]. 自动化学报, 2023, 49(1): 102-110.
LAI X, QU Y Y, XIE Y, et al. Topology-guided adversarial deep mutual learning for knowledge distillation[J]. Acta Automatica Sinica, 2023, 49(1): 102-110.

[70] XIANG L Y, DING G G, HAN J G. Learning from multiple experts: self-paced knowledge distillation for long-tailed classification[J]. arXiv Preprint, arXiv: 2001.01536, 2020.

[71] DU S, YOU S, LI X, et al. Agree to disagree: adaptive ensemble knowledge distillation in gradient space[J]. Advances in Neural Information Processing Systems, 2020, 33: 12345-12355.

[72] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5191-5198.

[73] SHEN C C, WANG X C, SONG J, et al. Amalgamating knowledge towards comprehensive classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019, 33(1): 3068-3075.

[74] TUNG F, MORI G. CLIP-Q: deep network compression learning by inparallel pruning-quantization[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7873-7882.

[75] LIANG T L, GLOSSNER J, WANG L, et al. Pruning and quantization for deep neural network acceleration: a survey[J]. Neurocomputing, 2021, 461: 370-403.

[76] KIM J, BHALGAT Y, LEE J, et al. QKD: quantization-aware knowledge distillation[J]. arXiv Preprint, arXiv: 1911.12491, 2019.

[77] ZEESHAN M, RAJ R, ANAND A, et al. Knowledge distillation-based AIoT framework for efficient wireless gesture sensing in B5G/6G networks[J]. IEEE Network, 2025, PP(99): 1.

[78] ZHANG K C, LI J, LI Z, et al. Transformer-based code model with compressed hierarchy representation[J]. Empirical Software Engineering, 2025, 30(2): 60.

[79] LIU N, MA X L, XU Z Y, et al. AutoCompress: an automatic DNN structured pruning framework for ultra-high compression rates[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 4876-4883.

[80] HOU L, HUANG Z Q, SHANG L F, et al. DynaBERT: dynamic BERT with adaptive width and depth[J]. Advances in Neural Information Processing Systems, 2020, 33: 9782-9793.

[81] MAO J F, WANG X T, AIZAWA K. The lottery ticket hypothesis in denoising: towards semantic-driven initialization[C]//European Conference on Computer Vision. Berlin: Springer Nature Switzerland, 2024: 93-109.

[82] FAN C X, GUO D, WANG Z Q, et al. Multi-objective convex quantization for efficient model compression[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2313-2329.

[83] XU J, TAN X, LUO R Q, et al. NAS-BERT: task-agnostic and adaptive-size BERT compression with neural architecture search[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2021: 1933-1943.

[84] RAO Y M, ZHAO W L, LIU B L, et al. DynamicViT: efficient vision transformers with dynamic token sparsification[J]. Advances in Neural Information Processing Systems, 2021, 34: 13937-13949.

[85] WANG C Q, ZHANG G D, GROSSE R. Picking winning tickets before training by preserving gradient flow[J]. arXiv Preprint, arXiv: 2002.07376, 2020.

[86] WAN A, DAI X L, ZHANG P Z, et al. FBNetV2: differentiable neural architecture search for spatial and channel dimensions[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12962-12971.

[87] TAN M X, LE Q V. MixConv: mixed depthwise convolutional kernels[J]. arXiv Preprint, arXiv: 1907.09595, 2019.

[88] CAI H, GAN C, WANG T Z, et al. Once-for-all: train one network and specialize it for efficient deployment[J]. arXiv Preprint, arXiv: 1908.09791, 2019.

[89] XU Y H, XIE L X, ZHANG X P, et al. PC-DARTS: partial channel

connections for memory-efficient architecture search[J]. arXiv Preprint, arXiv: 1907.05737, 2019.

[90] XIE S R, ZHENG H H, LIU C X, et al. SNAS: stochastic neural architecture search[J]. arXiv Preprint, arXiv: 1812.09926, 2018.

[91] REAL E, AGGARWAL A, HUANG Y P, et al. Regularized evolution for image classifier architecture search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 4780-4789.

[92] AKHAURI Y, MUNOZ J P, JAIN N, et al. EZNAS: evolving zero-cost proxies for neural architecture scoring[J]. Advances in Neural Information Processing Systems, 2022, 35: 30459-30470.

[93] PHAM H, GUAN M Y, ZOPH B, et al. Efficient neural architecture search via parameter sharing[J]. arXiv Preprint, arXiv: 1802.03268, 2018.

[94] YU J H, JIN P C, LIU H X, et al. BigNAS: scaling up neural architecture search with big single-stage models[C]//European Conference on Computer Vision. Berlin: Springer, 2020: 702-717.

[95] ZHANG L L, HAN S H, WEI J Y, et al. Nn-Meter: towards accurate latency prediction of deep-learning model inference on diverse edge devices[C]//Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. New York: ACM Press, 2021: 81-93.

[96] LIN J, CHEN W M, LIN Y J, et al. MCUNet: tiny deep learning on IoT devices[J]. arXiv Preprint, arXiv: 2007.10319, 2020.

[97] CHOI K, HONG D, YOON H, et al. DANCE: differentiable accelerator/network co-exploration[C]//Proceedings of the 2021 58th ACM/IEEE Design Automation Conference (DAC). Piscataway: IEEE Press, 2021: 337-342.

[98] SHARMA H, PARK J, SUDA N, et al. Bit fusion: bit-level dynamically composable architecture for accelerating deep neural network[C]//Proceedings of the 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA). Piscataway: IEEE Press, 2018: 764-775.

[99] PENG H W, DU H, YU H Y, et al. Cream of the crop: distilling prioritized paths for one-shot neural architecture search[J]. Advances in Neural Information Processing Systems, 2020, 33: 17955-17964.

[100] ELSKEN T, STAFFLER B, METZEN J H, et al. Meta-learning of neural architectures for few-shot learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12362-12372.

[101] TEERAPITTAYANON S, MCDANEL B, KUNG H T. BranchyNet: fast inference via early exiting from deep neural networks[C]//Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2016: 2464-2469.

[102] XIN J, TANG R, LEE J, et al. DeeBERT: dynamic early exiting for accelerating BERT inference[J]. arXiv Preprint, arXiv: 2004.12993, 2020.

[103] YU J H, YANG L J, XU N, et al. Slimmable neural networks[J]. arXiv Preprint, arXiv: 1812.08928, 2018.

[104] HUA W Z, ZHOU Y, SA C D, et al. Channel gating neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32: 1-11.

[105] LI F R, LI G, HE X Y, et al. Dynamic dual gating neural networks[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 5310-5319.

[106] 赵加坤, 刘云辉, 孙俊. 基于PCDARTS改进的神经网络架构搜索算法[J]. 计算机与数字工程, 2022, 50(4): 691-696, 720.
ZHAO J K, LIU Y H, SUN J. An improved neural architecture search algorithm based on PCDARTS[J]. Computer & Digital Engineering, 2022, 50(4): 691-696, 720.

[107] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[J]. Advances in Neural Information Processing Systems, 2017, 30: 3856-3866.

[108] WU Z X, NAGARAJAN T, KUMAR A, et al. BlockDrop: dynamic inference paths in residual networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8817-8826.

[109] WANG X, YU F, DOU Z Y, et al. SkipNet: learning dynamic routing in convolutional networks[C]//Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 420-436.

[110] CHEN Z R, LI Y, BENGIO S, et al. You look twice: GaterNet for dynamic filter selection in CNNs[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 9164-9172.

[111] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic convolution: attention over convolution kernels[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 11027-11036.

[112] YANG L, HAN Y Z, CHEN X, et al. Resolution adaptive networks for efficient inference[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 2366-2375.

[113] HA D, DAI A, LE Q V. HyperNetworks[J]. arXiv Preprint, arXiv: 1609.09106, 2016.

[114] BRABANDERE B D, XU J, TUYTELAARS T, et al. Dynamic filter networks[J]. Advances in Neural Information Processing Systems, 2016, 29: 667-675.

[115] YANG B, BENDER G, LE Q V, et al. CondConv: conditionally parameterized convolutions for efficient inference[J]. Advances in Neural Information Processing Systems, 2019, 32: 1-12.

[116] MA N N, ZHANG X Y, HUANG J W, et al. WeightNet: revisiting the design space of weight networks[C]//European conference on Computer Vision. Berlin: Springer, 2020: 776-792.

[117] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4320-4328.

[118] CHEN D F, MEI J P, WANG C, et al. Online knowledge distillation with diverse peers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020, 34(4): 3430-3437.

[119] CHEN P G, LIU S, ZHAO H S, et al. Distilling knowledge via knowledge review[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 5006-5015.

[120] XU T B, LIU C L. Data-distortion guided self-distillation for deep neural networks[C]//Proceedings of the AAAI Conference on Artifi-

cial Intelligence. Honolulu: AAAI Press, 2019: 5565-5572.

[121] FURLANELLO T, LIPTON Z C, TSCHANNEN M, et al. Born again neural networks[C]//Proceedings of the 35th International Conference on Machine Learning, 2018.

[122] KIM K, JI B, YOON D, et al. Self-knowledge distillation with progressive refinement of targets[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 6547-6556.

[123] CHEN H T, WANG Y H, XU C, et al. Data-free learning of student networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 3513-3521.

[124] CHOI Y, CHOI J, EL-KHAMY M, et al. Data-free network quantization with adversarial knowledge distillation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 3047-3057.

[125] THOKER F M, GALL J. Cross-modal knowledge distillation for action recognition[C]//Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2019: 6-10.

[126] OSTAPENKO O. Towards maintainable machine learning development through continual and modular learning[D]. Montreal: University of Montreal, 2024

[127] QIN J H, YE Z, TANG J H, et al. Dynamic knowledge routing network for target-guided open-domain conversation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8657-8664.

[128] HE Q, WANG Y, WANG X W, et al. Routing optimization with deep reinforcement learning in knowledge defined networking[J]. IEEE Transactions on Mobile Computing, 2024, 23(2): 1444-1455.

[129] SARTZETAKIS I, VARVARIGOS E. Network tomography with partial topology knowledge and dynamic routing[J]. Journal of Network and Systems Management, 2023, 31(4): 73.

[130] ANDONIAN A, CHEN S X, HAMID R. Robust cross-modal representation learning with progressive self-distillation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 16409-16420.

[131] JI Y Z, CHEN Y J, YANG L Q, et al. VeXKD: the versatile integration of cross-modal fusion and knowledge distillation for 3D perception[J]. Advances in Neural Information Processing Systems, 2024, 37: 125608-125634.

[132] HUO F S, XU W C, GUO J C, et al. C2KD: bridging the modality gap for cross-modal knowledge distillation[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 16006-16015.

[133] JIANG H Q, PAN Y, CHEN J H, et al. Oraclesage: towards unified visual-linguistic understanding of oracle bone scripts through cross-modal knowledge fusion[J]. arXiv Preprint, arXiv: 2411.17837, 2024.

[134] ZHU L G, ZHOU F, WANG S P, et al. A language-guided cross-modal semantic fusion retrieval method[J]. Signal Processing, 2025, 234: 109993.

[135] WEI Z M, PAN H Y, QIAO L B, et al. Cross-modal knowledge distillation in multi-modal fake news detection[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 4733-4737.

[136] XU Y D, FENG K, YAN X A, et al. Cross-modal fusion convolutional neural networks with online soft-label training strategy for mechanical fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2024, 20(1): 73-84.

[137] KWAK M G, MAO L C, ZHENG Z Y, et al. A cross-modal mutual knowledge distillation framework for Alzheimer's disease diagnosis: addressing incomplete modalities[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 14218-14233.

[138] LOUIZOS C, WELLING M, KINGMA D P. Learning sparse neural networks through L_0 regularization[J]. arXiv Preprint, arXiv: 1712.01312, 2017.

[139] SHAZEER N, MIRHOSEINI A, MAZIARZ K, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer[J]. arXiv Preprint, arXiv: 1701.06538, 2017.

[140] LEPIKHIN D, LEE H, XU Y Z, et al. GShard: scaling giant models with conditional computation and automatic sharding[J]. arXiv Preprint, arXiv: 2006.16668, 2020.

[141] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv Preprint, arXiv: 2101.03961, 2021.

[142] RAPOSO D, RITTER S, RICHARDS B, et al. Mixture-of-depths: dynamically allocating compute in transformer-based language models[J]. arXiv Preprint, arXiv: 2404.02258, 2024.

[143] WÓJCIK B, DEVOTO A, PUSTELNIK K, et al. Adaptive computation modules: granular conditional computation for efficient inference[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(20): 21510-21518.

[144] ZHOU Y Q, LEI T, LIU H X, et al. Mixture-of-experts with expert choice routing[J]. arXiv Preprint, arXiv: 2202.09368, 2022.

[145] LIN J, CHEN W M, CAI H, et al. Mcunetv2: memory-efficient patch-based inference for tiny deep learning[J]. arXiv Preprint, arXiv: 2110.15352, 2021.

[146] LIN J, ZHU L G, CHEN W M, et al. Tiny machine learning: progress and futures[J]. IEEE Circuits and Systems Magazine, 2023, 23(3): 8-34.

[147] VAIDYA N, OH F, COMLY N. Optimizing inference on large language models with NVIDIA TensorRT-LLM[R]. NVIDIA Technical Blog, 2023.

[148] NVIDIA Developer Blog. TensorRT-LLM speculative decoding boosts inference throughput by up to 3.6×[R]. 2024.

[149] Amazon Web Services. Amazon EC2 Inf1 instances: high-performance, low-cost ML inference[R]. 2025.

[150] DOUCET Z, SHARMA R, VOS M D, et al. HarMoEny: efficient multi-GPU inference of MoE models[J]. arXiv Preprint, arXiv: 2506.12417, 2025.

[151] XUE L Y, FU Y, LU Z, et al. MoE-infinity: efficient MoE inference on personal machines with sparsity-aware expert cache[J]. arXiv Pre-

print, arXiv: 2401.14361, 2025.

[152] LASBY M, GOLUBEVA A, EVCI U, et al. Dynamic sparse training with structured sparsity[J]. arXiv Preprint, arXiv: 2305.02299, 2023.

[153] WU B Q, XIAO Q, WANG S X, et al. Dynamic sparse training versus dense training: the unexpected winner in image corruption robustness[J]. arXiv Preprint, arXiv: 2410.03030, 2024.

[154] ULLAH N, SCHULTHEIS E, LASBY M, et al. Navigating extremes: dynamic sparsity in large output spaces[J]. Advances in Neural Information Processing Systems, 2024, 37: 117210-117236.

[155] HUANG H Y, ARDALANI N, SUN A N, et al. Toward efficient inference for mixture of experts[C]//Proceedings of the 38th International Conference on Neural Information Processing Systems. New York: ACM Press, 2024: 84033-84059.

[156] MISHRA R K, MONDAL A, MATHEW J. Nystromformer based cross-modality transformer for visible-infrared person re-identification[J]. Scientific Reports, 2025, 15: 16224.

[157] LIU J, DU Y, YANG K, et al. Edge-cloud collaborative computing on distributed intelligence and model optimization: a survey[J]. arXiv Preprint, arXiv: 2505.01821, 2025.

[158] BOVEE N, ALI I, BITLA S, et al. SplitTracr: a flexible performance evaluation tool for cooperative inference and split computing[C]// Proceedings of the Companion of the 16th ACM/SPEC International Conference on Performance Engineering. New York: ACM Press, 2025: 6-10.

[159] BAJPAI D J, TRIVEDI V K, YADAV S L, et al. SplitEE: early exit in deep neural networks with split computing[C]//Proceedings of the Third International Conference on Artificial Intelligence and Machine Learning Systems. New York: ACM Press, 2023: 1-9.

[160] CHEN Y X, LI R P, YU X X, et al. Adaptive layer splitting for wireless LLM inference in edge computing: a model-based reinforcement learning approach[J]. arXiv Preprint, arXiv: 2406.02616, 2024.

[作者简介]

王恩良（1998– ），男，江苏南京人，南京邮电大学博士生，主要研究方向为神经架构、深度学习。

阎庆昕（2000– ），女，山西太原人，南京邮电大学博士生，主要研究方向为神经架构、强化学习。

达明添（2001– ），男，江苏南京人，南京邮电大学硕士生，主要研究方向为神经网络设计、机器学习、大模型应用。

孙知信（1964– ），男，江苏南京人，博士，南京邮电大学教授、博士生导师，主要研究方向为计算机理论与技术、计算机网络及安全。