专题：网络智能化与生成式人工智能

自智网络全栈部署技术研究与实践

薛飞[1]，陈彬[1]，刘静[2]，梁晓扬[2]，朱琳[2]，王凤[2]，李天[1]，张靓[1]，陈贞贞[1]，李潇[1]
（1. 中国移动通信集团广东有限公司，广东 广州 510632;
2. 中国移动通信有限公司研究院，北京 100053）

摘　要：自智网络通过构建智能化的网络基础设施，实现网络自主管理、自优化和自修复。自智网络分成能力建设和能力部署两个关键阶段，目前业界较少关注能力部署。首先系统研究了自智网络的能力部署阶段，随后介绍自智网络架构，然后提出了 3 项自智网络全栈部署核心技术，最后通过异常检测、智慧机房和设备巡检等案例验证核心技术的有效性。对自智网络能力部署进行了系统探讨，对运营商实施网络智能化转型具有重要的参考价值。

关键词：自智网络；全栈部署；训推一体；云边协同部署；AI 能力编织
中图分类号：TN915
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023181

Research and practice on technologies for full stack deployment of autonomous networks

XUE Fei[1], CHEN Bin[1], LIU Jing[2], LIANG Xiaoyang[2], ZHU Lin[2], WANG Feng[2],
LI Tian[1], ZHANG Liang[1], CHEN Zhenzhen[1], LI Xiao[1]
1. China Mobile Group Guangdong Co., Ltd, Guangzhou 510632, China
2. China Mobile Research Institute, Beijing 100053, China

Abstract: Autonomous networks achieve network self management, self optimization, and self repair by building intelligent network infrastructure. Autonomous network was divided into two key stages: AI model building and AI model deployment. However, the industry paid less attention to AI model deployment. The deployment phase of autonomous networks was systematically studied. Firstly, it elaborated on the independent deployment mode and full stack deployment mode of autonomous networks, and pointed out that full stack deployment was the main direction. Secondly, a detailed introduction was given to the full stack architecture with “five layers, dual domains, and four closed-loops”, which achieved full life cycle intelligence through a layered closed-loop design of resources and processes. Then, three core technologies for independent innovation were proposed: AI model training and inference integration to achieve rapid iterative updates of models, AI fabric technology to achieve customized application by rapid construction, and AI model cloud-edge collaborative deployment technology to achieve efficient application. Finally, the effectiveness of these three core technologies was verified through cases such as anomaly detection, smart telecommunication rooms, and equipment inspections. The deployment of autonomous networks was systematically

收稿日期：2023−07−28；修回日期：2023−09−06

explored, especially in terms of architecture design and core technology innovation, which had important reference value for telecommunication operators' network digital transformation.

Key words： autonomous network, full stack deployment, training and inference integration, cloud edge collaborative deployment, AI fabric

0 引言

随着5G、云计算、大数据、人工智能等新一代信息技术的快速发展，全球通信行业正在加速数字化转型的进程。为了应对日益复杂的网络环境和日渐激烈的市场竞争，适应用户对网络质量的更高要求，国际电信管理论坛（TMF）提出了“自智网络”的概念，并分为L0到L5等级来评估自智网络水平。自智网络的目标是通过自动化和智能化的网络基础设施，对内实现网络自服务、自发放、自修复的能力，减少人工运维，降低运营成本，对外实现通信网络服务的零等待、零接触、零故障[1]。

全球各大电信运营商、设备提供商和标准化组织都积极推动自智网络建设。第三代合作伙伴计划（3rd Generation Partnership Project，3GPP）在5G核心网提出了网络数据分析功能（NWDAF）网络人工智能网元[2-3]。美国AT&T提出了Network 3.0 Indigo战略，信息技术（information technology，IT）与通信技术（communication technology，CT）深度融合，网络与大数据密切结合[4]。国际电信联盟（ITU）成立了国际电信联盟电信标准部（ITU-T）自智网络焦点组，规划网络演进和技术用例[5]。中国移动持续开展自智网络项目落地，将自动化、智能化技术与通信网络的软/硬件、管理支撑系统以及运维流程和组织结构深度融合，提升网络运营和维护的效率、质量和体验，推动通信网络向更高水平、更智能化的自动驾驶网络演进[5]。中国联通构建智能化综合性数字信息基础设施，推动网络敏捷、高效、智能[6]。

自智网络能力包括自动化能力和网络AI能力。为了统一表述，下文自智网络能力特指网络AI能力。自智网络分成能力建设和能力部署两个关键阶段。在能力建设阶段，研究人员致力于将网络AI技术应用于通信网络，开发出支持网络感知、诊断、预测和控制的AI模型和规则逻辑，输出了大量研究成果[7-10]。在能力部署阶段，自智网络建设侧重能力实际应用和落地效果，以及具体应用性能和影响评估。目前，业界较少关注自智网络的能力部署环节。其原因主要在于：一方面自智网络仍处发展阶段，研究重心仍集中在能力建设；另一方面能力部署需要长期的工程实践经验积累。本文专门研究自智网络的能力部署阶段，提出了自智网络全栈部署的核心技术要点，并通过相关典型案例进行探讨总结。

自智网络能力部署面临以下挑战：（1）通信网络数据复杂多变，现有网络AI能力泛化性不足，无法适配网络环境的频繁变化；（2）网络AI能力以单点方式部署，容易导致AI能力形成“模型孤岛”，不同能力之间缺乏协同联动。同时，单个AI能力感知范围有限，难以实现端到端网络状态的全局感知，制约了网络问题的准确定位和全局优化能力；（3）网络AI能力主要通过RESTful接口集中进行能力开放，这种交互模式无法满足涉敏数据、大数据量和低时延等特定应用场景要求。

1 自智网络体系架构

1.1 自智网络能力部署模式

自智网络能力有两种部署模式，分别是独立部署模式和全栈部署模式。

独立部署模式，也称为“烟囱式”部署。在这种模式下，每个AI应用需要各自收集样本数据、准备算法框架，提供部署推理所需要的算力资源。独立部署模式定制化程度高，可以针对具

体业务需求进行优化，但是整个过程独立封闭，可能导致公共部件重复建设，数据和算力资源浪费，模型研发和部署周期较长。

全栈部署模式，通过构建统一的算力资源池、数据中台、AI 模型训练和推理平台，使不同的 AI 应用可以共享基础算力资源、特征数据和 AI 能力。全栈部署模式提高资源利用效率，降低 AI 应用的开发和部署成本，缩短上线时间，且便于统一管理和协同。综合来看，全栈部署模式是自智网络能力部署的主要方向。

1.2　自智网络全栈部署

TMF 组织从用户视角出发，定义了自智网络的“3 层 4 闭环”结构模型[11]。这种结构模型过于概念化和抽象，无法有效指导运营商的自智网络具体建设。为解决此问题，本文借鉴网络分层思想，提出了自智网络全栈部署模式。自智网络全栈部署模式通过汇聚算力层、数据层、算法层、能力层和应用层 5 层资源，分别从逻辑域和系统域，面向场景进行网络 AI 应用部署，并且能力层对通信网络业务层、承载层和传输层进行跨层跨域联动闭环。自智网络全栈部署架构概括为“5 层双域 4 闭环”。

自智网络的 5 层架构由算力层、数据层、算法层、能力层和应用层构成，如图 1 所示。每一层都具备明确的资源构成和功能角色。算力层和数据层提供基础设施资源，算法层负责核心算法能力，能力层对网络进行智能化控制，应用层通过编排构建服务。自智网络采用逻辑域和系统域的双域架构。逻辑域从业务流程角度进行网络 AI 资源的逻辑抽象，系统域则是逻辑域在具体网管系统中的映射实现。自智网络通过 4 类 AI 能力闭环，构建自顶向下的业务层到传输层的闭环管理体系。

2　自智网络全栈部署的关键技术

自智网络全栈部署主要有 3 个关键技术，分别是：（1）AI 模型训推一体，该技术打通 AI 模型从训练到部署的全流程，通过动态增量学习与实时推理的无缝对接，保证模型的及时升级和快速迭代；（2）AI 能力编织，该技术将各类 AI 能力的 API 组合编排，以支撑复杂的网络智能化应用的快速构建；（3）云边协同的 AI 能力部署，该技术通过云端集中训练和管理 AI 模型，边缘节点部署实时推理，实现云端全局优化和边缘低时延

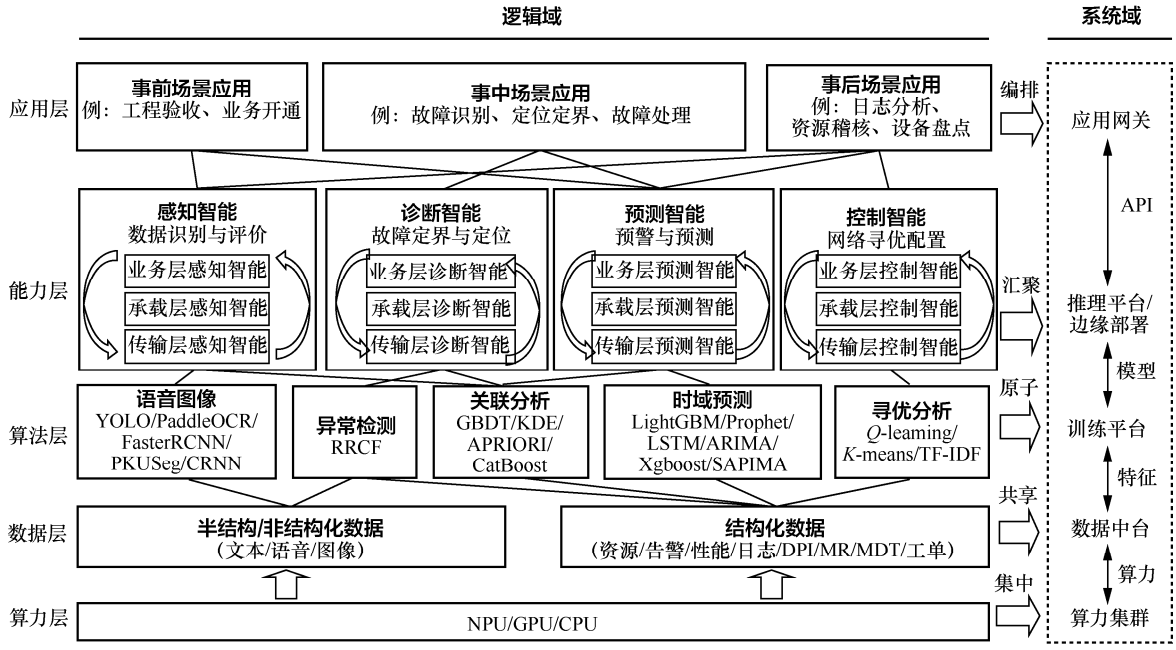

图 1　自智网络全栈部署体系

的结合，赋能网络的灵活智能。

2.1 关键技术 1：AI 模型训推一体技术

为提升网络 AI 模型的可泛化性，本文在自智网络 5 层双域结构基础上，参考 MLOps 思想[12-14]，提出了 3 种 AI 训推一体技术，分别是模型运行态训推一体同步模式、模型运行态训推一体异步模式和模型训练态训推一体模式。其中，模型运行态指网络 AI 模型训练完成后，被部署使用进行预测或决策的状态；模型训练态指网络 AI 模型在训练过程中的状态。

这 3 种技术将 AI 能力从手工更新模式转变为持续集成（continuous integration，CI）、持续交付（continuous deploy，CD）和持续训练（continuous training，CT）的流水线自动化模式，促进了 AI 能力的工程化落地。其区别在于，模型运行态训推一体的同步模式和异步模式直接在能力层完成全部流程，而模型训练态训推一体需要能力层、算法层和数据层等共同配合完成。

2.1.1 模型运行态训推一体同步模式

对于模型运行态训推一体同步模式，AI 能力既包含训练模块又包含推理模块。每次能力调用时，应用网关把所有训练数据和推理数据同时发送给 AI 能力。训练数据和推理数据同时发送，称为同步模式。训练模块进行训练，实时生成模型后再进行推理。模型运行态训推一体同步模式不存储模型，每次推理前都会先训练模型。本模式适用于小样本数据实现的轻量化 AI 模型，比如时序指标预测、指标异常检测等。模型运行态训推一体同步模式如图 2 所示，主要步骤如下。

步骤 1 应用网关将训练数据和推理数据同步发送给 AI 能力。

步骤 2 AI 能力的训练模块使用训练数据进行模型训练，生成运行态的新模型参数。

步骤 3 AI 能力使用步骤 2 生成的新模型参数，对步骤 1 推理数据进行模型推理，输出结果返回给应用网关。

2.1.2 模型运行态训推一体异步模式

对于模型运行态训推一体异步模式，AI 能力既包含训练模块又包含推理模块。能力首次调用时，应用网关发送训练数据和推理数据给 AI 能力。但是与同步模式不同，后续除了能力更新，能力被调用时只需要发送推理数据，这种训推一体模式被称为异步模式。异步模式的 AI 能力会以序列化文件形式存储在外部介质。每次能力被调用时，应用网关只发送推理数据和模型文件索引，AI 能力通过索引把对应的模型文件反序列化后，对推理数据进行计算返回结果。本模式适用于中等规模 AI 模型，如网络故障识别预测等。模型运行态训推一体异步模式如图 3 所示，主要步骤如下。

步骤 1 应用网关将推理数据和模型文件索引发送给 AI 能力。

步骤 2 AI 能力的推理模块根据索引，检索存储介质中的模型文件。如果存在模型文件，则跳转到步骤 4；否则，跳转到步骤 3。

步骤 3 如果用户首次使用 AI 能力，或用户发

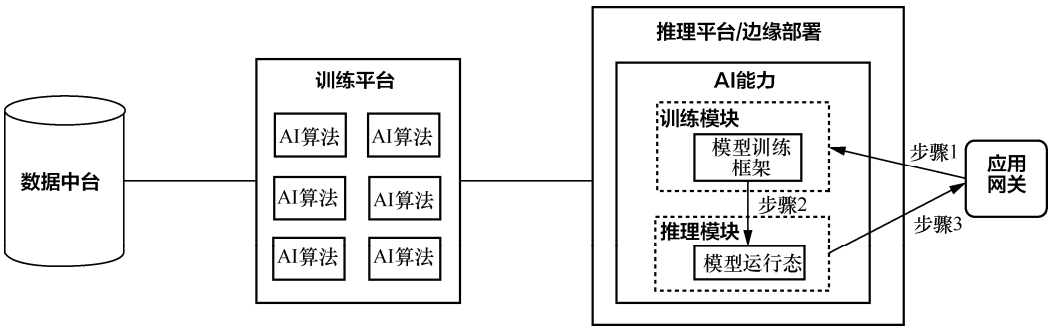

图 2 模型运行态训推一体同步模式

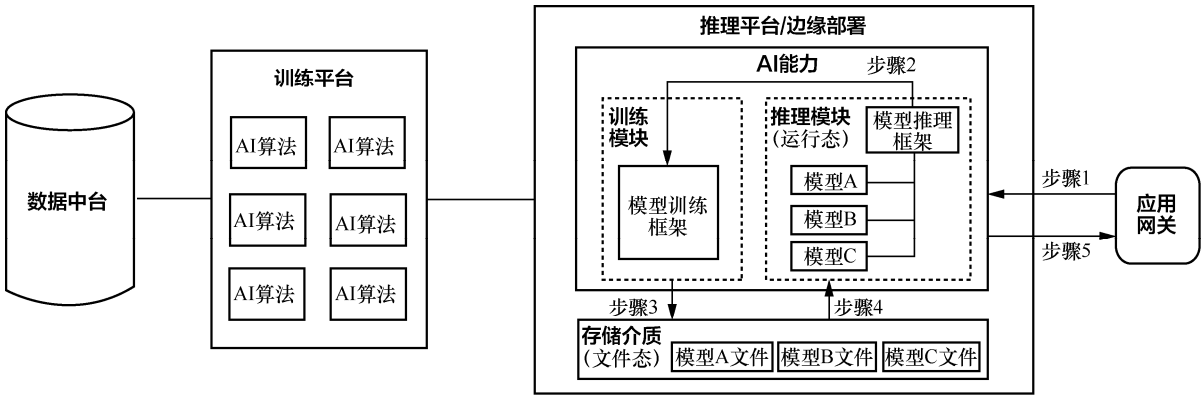

图 3　模型运行态训推一体异步模式

现 AI 能力性能下降，按需发起模型重训练请求，并提供训练数据。训练模块使用训练数据重新训练模型，并将模型文件及其对应的索引存储在外部介质中。

步骤 4　按照模型索引，在存储介质中检索对应的模型文件，并反序列化为运行态的模型框架。

步骤 5　使用运行态的模型框架，对步骤 1 的推理数据进行推理，并将结果返回给应用网关。

2.1.3　模型训练态训推一体模式

对于模型训练态训推一体模式，AI 能力只包含推理过程。当 AI 能力性能下降或首次调用，训练平台会从数据中台获取最新的样本数据，用于重新训练模型，并在推理平台自动或半自动部署为 AI 能力。本模式适用于大参数量的模型，或者需要图形处理器（graphics processing unit，GPU）等算力长时间训练迭。模型训练态训推一体模式如图 4 所示，主要步骤如下。

步骤 1　应用网关将推理数据发送给 AI 能力。

步骤 2　如果用户首次使用 AI 能力，或用户发现 AI 能力性能下降，可以发起模型重训练请求，跳转到步骤 3；否则跳转到步骤 6。

步骤 3　训练模块向数据中台发出获取训练数据的请求。

步骤 4　数据中台通过鉴权后，返回对应训练数据给训练模块。

步骤 5　训练模块使用获取的训练数据对模型进行重新训练，并将重新训练好的模型文件或镜像发送给推理平台。

步骤 6　推理平台使用运行态的 AI 能力对步骤 1 中推理数据进行预测，将结果返回给应用网关。

2.2　关键技术 2：AI 能力编织技术

本文参考编织计算技术[15-17]，提出了 AI 能力编织（AI fabric）网络智能组合编排方式。AI 能力编织技术按照一定规则，编排组合局部、单点、单功能的 AI 能力，形成整体化、高可用、多功能的复合 AI 应用，实现网络智能从点状应用向面向复杂任务场景的演进，大幅降低了 AI 应用开发的难度和复杂性。

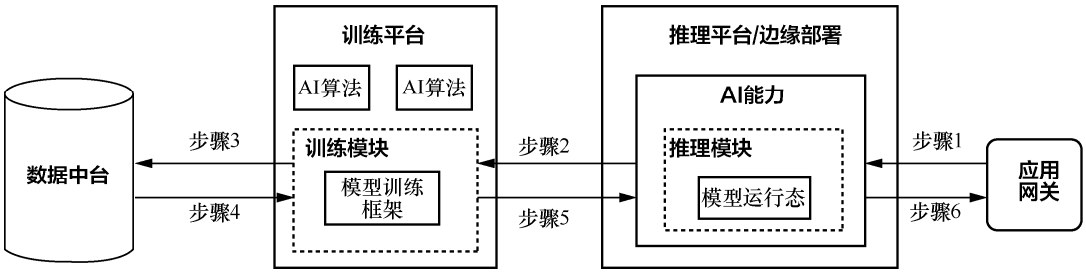

图 4　模型训练态训推一体模式

2.2.1 AI 能力编织技术

自智网络应用层获取能力层封装的 AI 能力开放接口，通过 AI 能力编织引擎，采用模型组合、服务编排等方式，按需组合多个 AI 微服务的接口，形成针对复杂应用场景的自定义 AI 流程，经权限校验后对外开放访问。该架构实现了网络 AI 能力的解耦和服务化架构，支持快速、灵活的 AI 应用开发和交付。AI 能力编织技术示意图如图 5 所示。

AI 能力编织的优势在于它通过一系列标准化的编排范式管理各个原子 AI 能力，能够快速、灵活地将稳定可靠的原子 AI 能力组合，生成针对特定业务场景的 AI 应用，实现开箱即用和灵活配置，为用户提供灵活的解决方案。为了实现 AI 能力编织，需要对每个 AI 能力进行标准化，规范能力输入、输出、算法框架和算力需求。

2.2.2 AI 能力编织规则

AI 能力编织主要包括 4 种典型规则：

（1）1-IN-1-OUT，单进单出；

（2）1-IN-*N*-OUT，单进多出；

（3）*N*-IN-1-OUT，多进单出；

（4）*N*-IN-*M*-OUT，多进多出。

AI 能力编织规则见表 1。

每种能力编织规则在接收一个或多个 AI 能力输出后，可进行直接转发、复制、聚合、最大最小值提取等数据转换，最终输出结果给特定的目标对象或流程。各个能力编织规则可自定义配置，以满足不同的网络智能场景需求。数据转换方式类型说明见表 2。

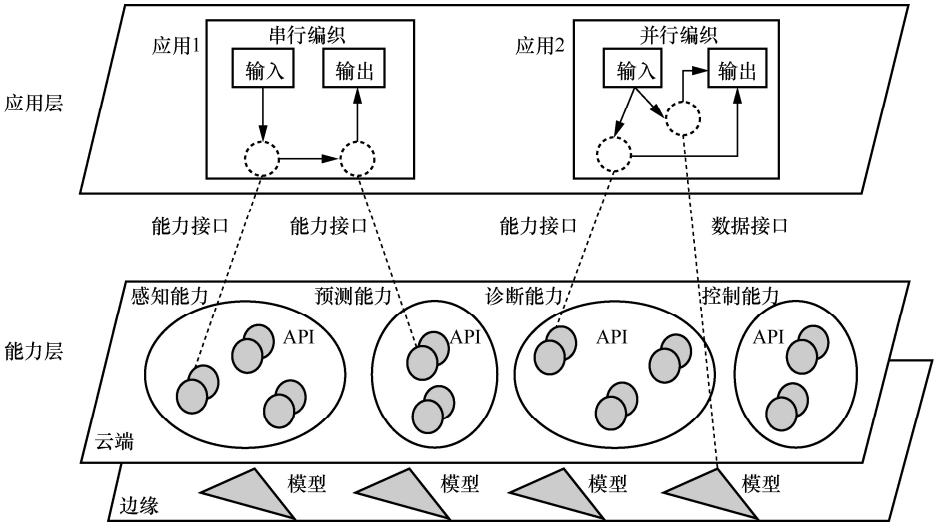

图 5 AI 能力编织技术示意图

表 1 AI 能力编织规则

结构名称	主要功能	使用场景	输入数据	数据转换方式	输出数据
1-IN-1-OUT	单进单出。接收一组数据，直接转发给特定目标对象	主要用于 AI 应用输入输出，或串行连接两个 AI 原子能力	一组数据	直接转发	与输入相同的一组数据
1-IN-*N*-OUT	单进多出。接收一组数据，多副本复制给特定目标对象	主要用于并行连接多个 AI 原子能力	一组数据	多副本复制	与输入相同的多组数据
N-IN-1-OUT	多进单出。接收多组数据，通过转换后，输出一组数据给特定目标对象	主要用于 AI 应用输出	多组数据	累加、最大、最小、平均值、众数等	转换后的一组数据
N-IN-*M*-OUT	多进多出。接收多组数据，通过转换后，输出为多组数据给特定目标对象	主要用于 AI 应用的多数据转换	多组数据	多副本复制	多组数据

表 2　数据转换方式类型说明

数据转换类型	转换逻辑	主要功能
直接转发数据转换	$f(x)=x$ x 为单个或单组数据	接收输入单个或单组数据，直接转发到目标对象
多副本复制数据转换	$f(x)=(x,x,\cdots,x)$ x 为单组或多组数据	接收输入单组或多组数据，复制为多个副本，把每个副本分别转发到不同的目标对象
多输入合并数据转换	$f(x,y)=(x,y)$ x 和 y 为单组或多组	接收输入单组或多组数据，按照某个维度进行数据合并
累加数据转换	$f(x)=\sum_i x_i$ x 为多个数据	接收输入的多个数据，对所有数据累加后，把结果转发到目标对象
统计函数转换	$f(x)=\text{statistics}(x)$ x 为多个数据，statistics(·) 表示统计函数，用于取最大值、最小值、平均值或众数	接收输入的多个数据，对所有数据取最大值，或最小值、平均值或众数后，把结果转发到目标对象

2.2.3　网络 AI 应用编织类型

AI 应用需要接收外部程序发送的输入数据，经过内部的模型推理计算，再返回输出结果。将 AI 应用接收外部调用请求的接口称为北向接口，返回推理结果的接口称为南向接口。通过组合不同的 AI 能力编织结构与原子 AI 能力，可以构建不同类型的 AI 应用。

网络 AI 应用一般可分为以下 4 类：单点 AI 应用、并行 AI 应用、串行 AI 应用和混合 AI 应用。以上 AI 应用类型的组合开发，支持快速高效地构建适应多样化场景需求的网络 AI 解决方案。网络 AI 的 4 种组合应用类型示意图如图 6 所示，网络 AI 的 4 种组合应用类型见表 3。

2.3　关键技术 3：AI 能力的云边协同部署

对于数据安全敏感、数据传输成本高昂或者需要实时响应的场景，不宜采用将 AI 模型集中部署在云端的方式[18]。这主要是因为：（1）数据传输到云端时存在泄露风险，不适用于安全要求极高的场景[19-20]；（2）大量数据上传到云端需要占用带宽，可能导致传输成本高昂[21]；（3）模型云

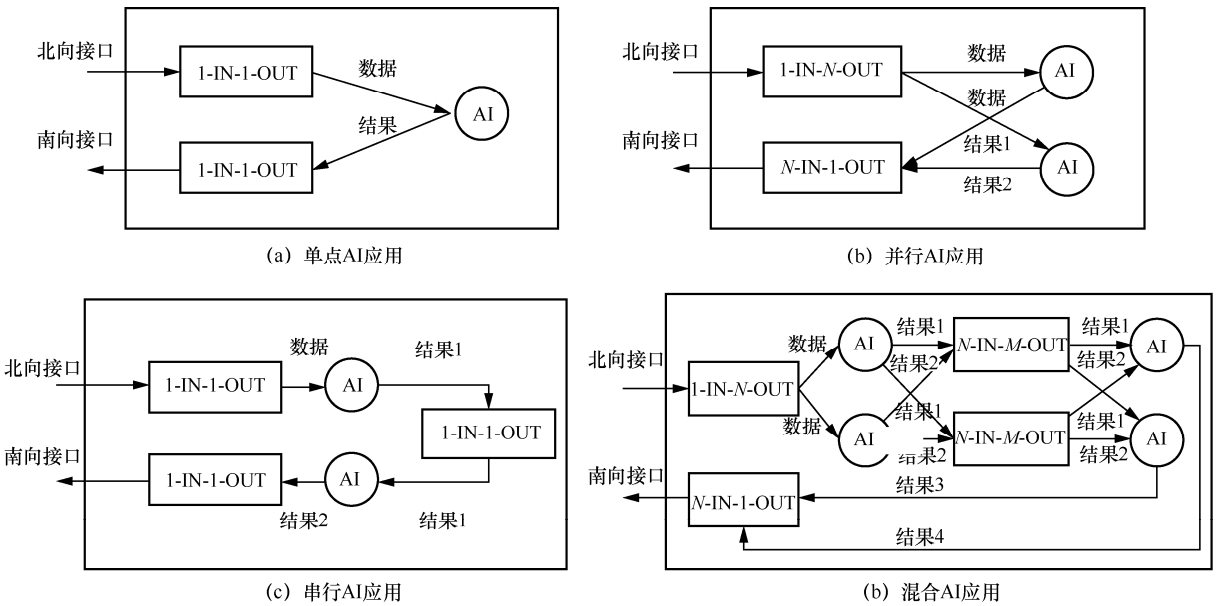

图 6　网络 AI 的 4 种组合应用类型示意图

表 3　网络 AI 的 4 种组合应用类型

应用类型	网络 AI 应用的内部结构	网络 AI 应用的适用场景
单点 AI 应用	简单的单一 AI 模型服务，具有一个北向输入接口和一个南向输出接口。北向和南向接口分别通过两个 1-IN-1-OUT 结构与单个 AI 模型连接	这种结构用于简单的单任务场景，如人脸识别和流量预测等
并行 AI 应用	把多个 AI 服务并行执行，各自产生输出，组合成一个统一输出。北向接口通过 1-IN-*N*-OUT 结构连接一组 AI 模型，每个模型并行执行一个任务；南向接口通过 *N*-IN-1-OUT 结构聚合各模型结果	可用于模型集成提升精度，也可用于功能分解后串行执行
串行 AI 应用	通过多个AI服务串行执行，后续服务使用前序服务输出作为输入，最终形成整体输出。北向和南向接口通过两个 1-IN-1-OUT 结构连接，中间通过多组 1-IN-1-OUT 串行连接不同 AI 模型，前一模型输出是后一模型输入	可用于多阶段处理，也可用于模型串行集成
混合 AI 应用	支撑复杂应用场景，串行、并行、分支等结构复合使用，形成聚合型输出。北向接口通过 1-IN-*N*-OUT 连接多个模型，南向接口通过 *N*-IN-1-OUT 聚合输出	模型之间连接形式混合并行和串行，用于复杂全流程场景

端部署需要进行数据上传和结果返回，整个过程时延较高，无法满足对实时性要求极高的应用，如流式处理类[22]。

本文重点是不同的数据特性和生产场景需求，通过网络 AI 能力云端部署和边缘部署模式，可以在发挥强大计算能力的同时，降低 AI 能力在云端和边缘端部署的门槛，使网络智能化更具弹性、敏捷性。AI 能力云边协同部署如图 7 所示。

2.3.1　AI 能力云端部署模式

AI 能力云端部署模式是将训练好的 AI 模型集中部署在推理平台的服务器集群中，形成提供服务的 AI 能力[21]。这种部署模式的主要技术要点包括模型在线部署运维、弹性计算和资源调度、高性能计算和自动化运维监控等。

（1）对于模型在线部署运维，云端部署需要建立模型的在线部署和运维机制，包括模型的编译、打包和自动化部署等。这些技术能够实现模型的可靠部署、动态扩展和快速升级，保证云端能力的高可用性和可扩展性。

（2）对于弹性计算和资源调度，云端需要具备弹性计算和资源调度的能力，以保证在高负载时能够高效利用计算资源，并合理分配给不同的任务。这包括动态资源调度算法、容器技术和虚拟化技术的应用。

（3）对于高性能计算，云端部署需要具备高性能计算能力，能够处理大规模数据和复杂计算任务。使用图形处理器、张量处理器（tensor processing unit，TPU）和神经网络处理器（neural processing unit，NPU）等提高模型训练和推理的速度和效率，从而提供更快的响应时间和更好的用户体验。

（4）对于自动化运维监控，云端部署需要建立完善的运维和监控系统，自动化地管理模型的部署、升级和监控。这包括自动化运维工具、性能监控和故障检测等技术，以及对模型运行情况的实时监控和分析。

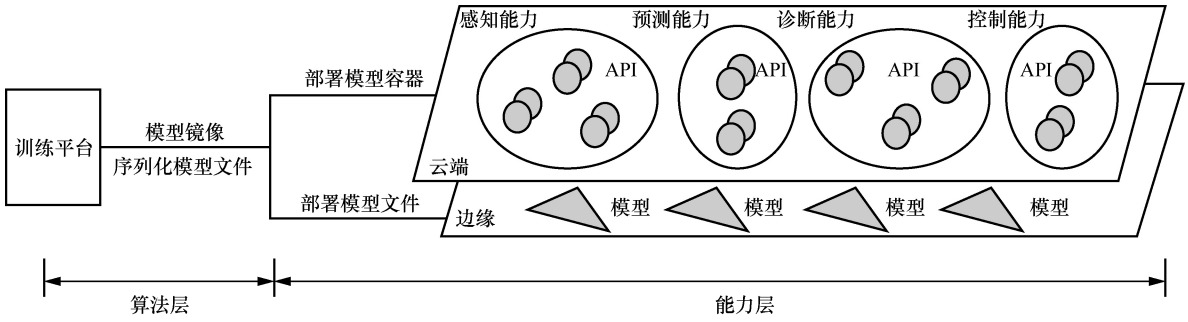

图 7　AI 能力云边协同部署

通过上述关键技术，能够实现 AI 能力在云端的高效部署和运行，提供稳定可靠的服务，并支持大规模的数据处理和复杂计算任务。

2.3.2　AI 能力边缘部署模式

为实现 AI 能力边缘部署，核心在于 AI 模型的快速迁移。算法层需要能够高效地将增量更新后的模型推送到海量边缘设备部署，保证边缘端模型及时升级；能力层的边缘端需要能采集本地数据，反馈到云端持续改进全局模型。因此，模型压缩与优化、增量模型更新以及与之配套的数据采集反馈机制是实现 AI 能力边缘部署的关键。

（1）对于模型压缩与优化，需要使用模型压缩技术对训练的大型模型进行轻量化改造，生成对计算资源需求更低的轻量级模型。常用的模型压缩技术包括参数剪枝、知识蒸馏等。参数剪枝通过移除冗余参数，降低模型大小；知识蒸馏则使用模型去训练小模型，传递知识。此外，还可以采用量化、裁剪等方式优化模型。这些方法可以将云端模型大小压缩至 1/10 以下，降低计算需求，使得模型可以高效部署到边缘端。

（2）对于增量模型更新，需要增量地将新模型部署到边缘端，而不是完整重新部署，以提高效率。关键技术包括基于模型剪枝的增量训练，其中只重新训练修改部分，以及使用补丁技术仅更新模型差异部分。此外，采用直传法直接发送增量文件，或采用点对点方式分发等，都可以实现模型快速增量更新。

（3）对于数据采集反馈机制，需要边缘端实时收集本地数据，反馈到云端，包括对接口调用、模型输出、运行指标等进行采集。反馈的数据既可直接用于算法层模型的增量训练，也可进行汇总后供模型质量监测之用。同时，需要建立反馈的数据管道与存储。还需要考虑数据安全、合规等问题。

模型压缩优化、增量更新和数据反馈机制是实现云边模型快速迁移与协同的核心技术，能够保证云边端部署的及时升级和模型的不断优化。这对于构建弹性、智能的网络基础环境具有重要意义。此外，还需要统一的服务注册发现机制，对模型资产进行统一管理。

3　自智网络全栈部署的应用实践

3.1　基于多源异构数据的网络异常检测应用

基于多模态数据的网络异常检测应用通过组合使用 KPI 异常检测、告警异常检测和日志异常检测能力进行 AI 能力编织，从多数据源联合识别判断网络异常。同时，采用集训练与推理于一体的云端 AI 部署方案，使检测模型可以实时响应网络数据，显著提升了检测的准确率和故障响应的时效性。相比传统监控，该方案最大限度地发挥了多源异构数据的优势，实现了对网络状况的主动感知、解析与响应，并能及时更新模型参数，是网络故障运维智能化的重要突破。基于多源异构数据的网络异常检测应用示意图如图 8 所示。

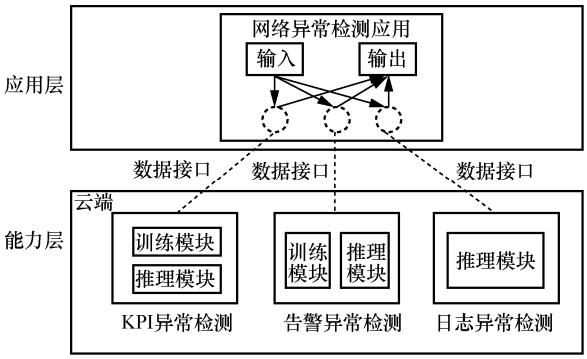

图 8　基于多源异构数据的网络异常检测应用示意图

3.2　智慧机房应用

机房管理是网络运维的基础工作，包括日常巡检、电源验收、设备管理和上下电管控等。智慧机房应用通过 AI 编排技术，整合人员管理、工程管理、机房管理、自动验收和资源管理等 AI 能力，聚焦一线维护人员的生产运维需求，解决网络机房在安全管理方面的短板。智慧机房应用围绕机房安全、生产运营、资产数智化等场景，结合实际机房生产运维需求，实现机房人员、事件、资产的安全高效管理，使生产流程贯通、可管控可追溯。智慧机房应用示意图如图 9 所示。

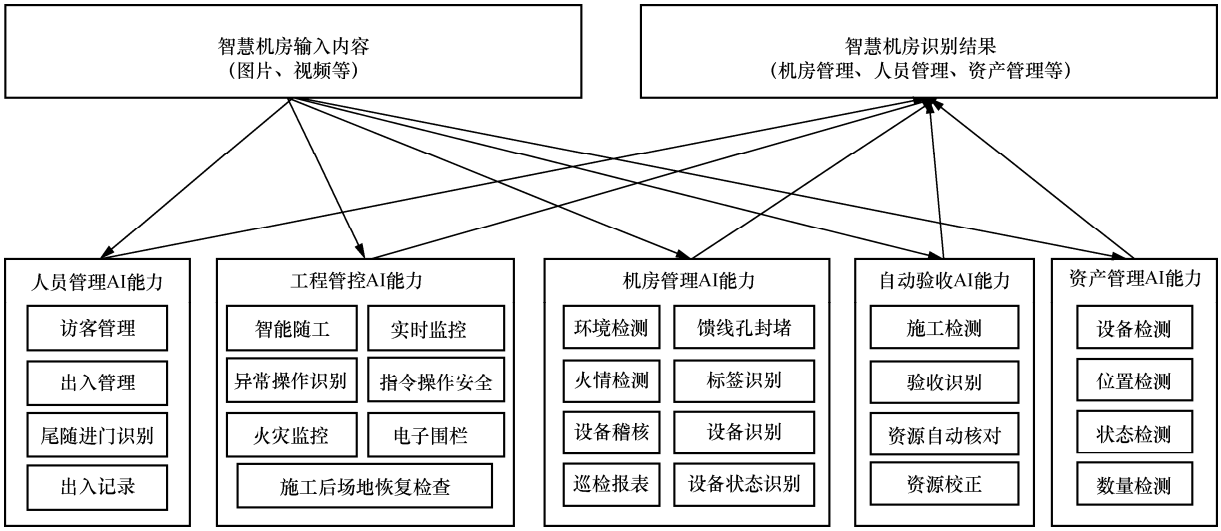

图 9　智慧机房应用示意图

3.3　基于视频的网络设备巡检应用

网络中存在海量无源设备，需要进行定期的资产盘点、资源录入和信息核对。传统的人工盘点方式存在效率低下、质检效果差、工作量大等问题。为了提高网络设备巡检效率，降低人员工作量，打造基于视频的网络设备巡检应用。该应用的主要流程是移动终端上部署设备巡检的 AI 模型和软件开发工具包（software development kit，SDK），形成边缘计算能力。巡检人员使用终端对现场设备进行拍摄录像。部署在终端的 AI 模型对视频流进行智能分析，自动识别设备资产信息。将识别结果通过数据接口上传至巡检系统的云端平台。平台数据后台接收结果，形成设备巡检结果数据。基于视频的网络设备巡检应用示意图如图 10 所示。

相比传统人工巡检，基于视频的网络设备巡检应用有很多优点，具体如下。

（1）提高了设备巡检效率。一个巡检员携带终端就可以完成视频采集，无须记录纸质表格，大大减少了人工录入环节。

（2）提高了巡检质量。通过 AI 模型的精确识别，相比人工识别可以大幅降低错误率。同时，利用视频检测，解决了单张照片识别时设备被遮

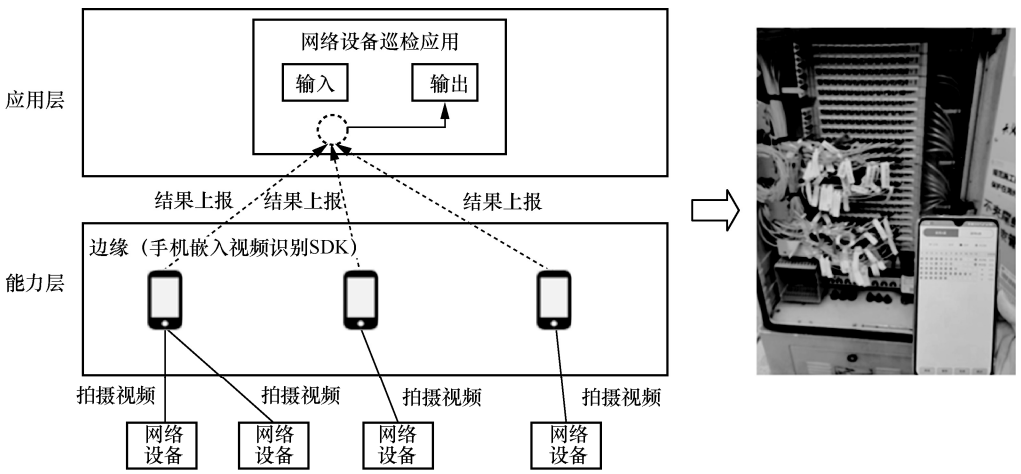

图 10　基于视频的网络设备巡检应用示意图

挡而导致精度下降的问题。

（3）实时性强，边缘部署避免了大量视频的传输时延，可以对视频流进行实时分析。

（4）保证了数据安全性，视频和数据均存储在内部系统，不会外泄。

（5）易于部署，通过 SDK 方式在现有的终端上部署应用，无须额外采购专用设备。

4　结束语

随着通信网络向自智网络演进，网络运营商面临着日益复杂的网络环境和用户需求。如何实现网络基础设施的智能化升级，是运营商必须解决的核心问题[23-24]。本文系统地研究和探讨了自智网络全栈部署的关键技术，具有重要的理论价值和实践意义。

在当下 5G 和数字化转型的背景下，运营商亟须加快自智网络建设的步伐。本文的研究成果为运营商搭建自智网络全栈提供了科学指导。随着以 ChatGPT 为代表的大模型等新型人工智能技术出现，自智网络还需要持续考虑研究新技术如何与网络架构深度融合[25]，以支持网络的自动化、智能化和可编程化。同时，随着时间推移，运营商的网络环境和业务需求也会发生变化，所以自智网络全栈化部署方案也需要持续迭代更新，以适应新出现的网络场景需要。

参考文献：

[1] TMF. A white paper on autonomous networks[R]. 2022.

[2] 李爱华, 吴晓波, 陈超, 等. 5G 网络大数据智能分析技术[J]. 电信科学, 2022, 38(8): 129-139.
LI A H, WU X B, CHEN C, et al. Big data intelligent analysis technology for 5G network[J]. Telecommunications Science, 2022, 38(8): 129-139.

[3] 欧阳晔, 王立磊, 杨爱东, 等. 通信人工智能的下一个十年[J]. 电信科学, 2021, 37(3): 1-36.
OUYANG Y, WANG L L, YANG A D, et al. Next decade of telecommunications artificial intelligence[J]. Telecommunications Science, 2021, 37(3): 1-36.

[4] 李涛, 王春佳, 李姗姗. 电信网络智能化方案研究[J]. 电信科学, 2023, 39(3): 162-172.
LI T, WANG C J, LI S S. Research on intelligent scheme of telecommunication network[J]. Telecommunications Science, 2023, 39(3): 162-172.

[5] 中国联通. 中国联通自智网络白皮书[R]. 2021.
China Unicom. China Unicom autonomous networks white paper[R]. 2021.

[6] 中国移动. 中国移动自动驾驶网络白皮书[R]. 2021.
China Mobile. China Mobile practice on autonomous network white paper[R]. 2021.

[7] RAFIQUE D, VELASCO L. Machine learning for network automation: overview, architecture, and applications[J]. Journal of Optical Communications and Networking, 2018, 10(10): 126-143.

[8] LUO G Y, YUAN Q, LI J L, et al. Artificial intelligence powered mobile networks: from cognition to decision[J]. IEEE Network, 2021, 36(3): 136-144.

[9] NI J B, LIN X D, SHEN X S. Efficient and secure service-oriented authentication supporting network slicing for 5G-enabled IoT[J]. IEEE Journal on Selected Areas in Communications, 2018, 36(3): 644-657.

[10] CHEN M Z, CHALLITA U, SAAD W, et al. Machine learning for wireless networks with artificial intelligence: a tutorial on neural networks[J]. 2017.

[11] 邓灵莉, 张新鹏, 顾宁伦, 等. 开源与产业研究[J]. 电信科学, 2021, 37(10): 12-21.
DENG L L, ZHANG X P. GU N L, et al. Network intelligence standards, open source and industry research[J]. Telecommunications Science, 2021, 37(10): 12-21.

[12] KREUZBERGER D, KÜHL N, HIRSCHL S. Machine learning operations (MLOps): overview, definition, and architecture[J]. IEEE Access, 2022(11): 31866-31879.

[13] GARG S, PUNDIR P, RATHEE G, et al. On continuous integration/continuous delivery for automated deployment of machine learning models using MLOps[C]//Proceedings of 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). Piscataway: IEEE Press, 2022: 25-28.

[14] ZHOU Y, YU Y, DING B. Towards MLOps: a case study of ML pipeline platform[C]//Proceedings of 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE). Piscataway: IEEE Press, 2021: 494-500.

[15] MARCHETTI N, PRASAD N R, JOHANSSON J, et al. Self-organizing networks: state-of-the-art, challenges and perspectives[C]//Proceedings of 2010 8th International Conference

on Communications. Piscataway: IEEE Press, 2010: 503-508.

[16] CHEN M, LIU J, LI P, et al. Fabric computing: concepts, opportunities, and challenges[J]. Innovation (Cambridge (Mass)), 2022, 3(6): 100340.

[17] DING Z, PENG Y, SONG H. Cloud-native data fabric for large-scale machine learning[J]. Frontiers of Computer Science, 2020, 14(2):353-373.

[18] 朱明伟. 网络智能化中的 AI 工程化技术方案[J]. 电信科学, 2022, 38(2): 157-165.
ZHU M W. AI engineering technology solutions in network intelligence[J]. Telecommunications Science, 2022, 38(2): 157-165.

[19] 王晴天, 刘洋, 刘海涛, 等. 面向 6G 的网络智能化研究[J]. 电信科学, 2022, 38(9): 151-160.
WANG Q T, LIU Y, LIU H T, et al. Research on network intelligence for 6G[J]. Telecommunications Science, 2022, 38(9): 151-160.

[20] 詹勇, 吴枫. 5G 网络自动化: 从人工运维到全自治[J]. 电信科学, 2022, 38(8): 140-150.
ZHAN Y, WU F. 5G network automation: from manual operating to full autonomous[J]. Telecommunications Science, 2022, 38(8): 140-150.

[21] 程瑞营, 张攀, 肖雨, 等. 基于时序数据的云网协同平台人工智能运维体系[J]. 电信科学, 2022, 38(11): 24-35.
CHENG R Y, ZHANG P, XIAO Y, et al. Time series data based AI operation and maintenance system of cloud network collaboration platform[J]. Telecommunications Science, 2022, 38(11): 24-35.

[22] DAI L L, JIAO R C, ADACHI F, et al. Deep learning for wireless communications: an emerging interdisciplinary paradigm[J]. IEEE Wireless Communications, 2020, 27(4): 133-139.

[23] 杨震, 赵建军, 黄勇军, 等. 基于网络演进的人工智能技术方向研究[J]. 电信科学, 2022, 38(12): 27-34.
YANG Z, ZHAO J J, HUANG Y J, et al. Study on the direction of artificial intelligence technology based on network evolution[J]. Telecommunications Science, 2022, 38(12): 27-34.

[24] 郭泽华, 朱昊文, 徐同文. 面向分布式机器学习的网络模态创新[J]. 电信科学, 2023, 39(6): 44-51.
GUO Z H, ZHU H W, XU T W. Network modal innovation for distributed machine learning[J]. Telecommunications Science, 2023, 39(6): 44-51.

[25] 张嗣宏, 张健. 以 ChatGPT 为代表的生成式 AI 对通信行业的影响和应对思考[J]. 电信科学, 2023, 39(5): 67-75.
ZHANG S H, ZHANG J. Impact and countermeasures of generative AI represented by ChatGPT on the telecom industry[J]. Telecommunications Science, 2023, 39(5): 67-75.

[作者简介]

薛飞（1983− ），男，中国移动通信集团广东有限公司高级工程师，主要研究方向为自智网络高阶演进、网络 AI 应用开发和网络 AI 训练平台等。

陈彬（1980− ），男，现就职于中国移动通信集团广东有限公司，主要研究方向为自智网络高阶演进、网络智能化规划等。

刘静（1991− ），女，现就职于中国移动通信有限公司研究院，主要研究方向为九天网络智能化平台规划和需求管理、MLOps 训推一体应用等。

梁晓扬（1984− ），男，现就职于中国移动通信有限公司研究院，主要研究方向为自智网络平台架构、MLOps 以及大模型在自智网络领域的应用。

朱琳（1983− ），女，中国移动通信有限公司研究院网络智能化能力及应用产品线 CEO、高级工程师，长期从事通信网络智能化理论、算法、关键技术、产品研发及应用落地相关工作。

王凤（1985− ），女，现就职于中国移动通信有限公司研究院，主要研究方向为网络智能化一线综合应用样板间设计、通信网络大模型等。

李天（1981− ），男，现就职于中国移动通信集团广东有限公司，主要研究方向为网络异常检测算法和应用、九天网络智能化平台边缘节点建设、AI 能力编织技术等。

陈贞贞（1990− ），女，现就职于中国移动通信集团广东有限公司，主要研究方向为自智网络体系架构、自智网络成效评估、MLOps 训推一体等。

张靓（1980− ），女，现就职于中国移动通信集团广东有限公司，主要研究方向为自智网络全栈部署体系、MLOps 训推一体。

李潇（1982− ），女，现就职于中国移动通信集团广东有限公司，主要研究方向为自智网络应用开发等。