

基于深度学习的轻量化目标检测模型综述

邓 览¹

DENG Lan

摘 要

针对基于深度学习的轻量化目标检测模型研究现状,文章系统介绍了深度学习目标检测分类器及轻量化目标检测模型的核心内容。重点分析了近年来该领域的研究热点,涵盖轻量化目标检测模型的架构设计方法、骨干网络架构等关键方向;结合当前主流检测数据集,对模型的核心评估参数及综合性能进行了全面评估。最后,对现有深度学习轻量化目标检测模型进行综合评述,并为优秀深度学习轻量化模型的设计提供了针对性建议,为该领域的后续研究与工程应用提供参考。

关键词

深度学习;轻量化模型;目标检测;计算机视觉

doi: 10.3969/j.issn.1672-9528.2026.06.038

0 引言

随着移动设备与高性能计算单元飞速发展,加之具身智能等边缘应用场景的现实需求^[1],深度学习目标检测^[2]算法迎来了全新的发展要求。现有多数检测模型虽在精度与实时性上表现优异,但普遍依赖 CPU 与 GPU 充足算力,难以直接部署于边缘终端。尽管借助模型拆分等技术可实现部分程序边缘落地,轻量化目标检测模型却更适配资源受限场景,由此成为当下热门研究方向。本文系统梳理轻量化目标检测模型类别与骨干网络架构,结合边缘计算应用场景构建评价指标体系并给出设计思路,可为相关后续研究与工程实践提供理论依据与实践参考。

近年来,已有若干关于目标检测或轻量化模型的综述文献。例如, Jiao 等^[3]系统梳理了深度学习目标检测的基础范式; Kamath 等^[4]则聚焦于资源受限设备上的目标检测挑战。与这些工作相比,本文的创新点与侧重点在于:

(1) 场景驱动的架构分析: 不仅按“两阶段/一阶段”分类,更紧密结合边缘计算的资源约束(算力、内存、功耗、带宽),分析各类模型轻量化改造成本与部署可行性,指出“一阶段及无锚框(Anchor-Free)模型是轻量化设计的优化起点”^[5]。

(2) 性能评估体系化: 超越单纯对比 mAP,构建了包含精度(AP/mAP)、推理效率(FLOPs,参数量)、速度(FPS,推理时间)及实时性判定的多维评估框架,并结合 MS-COCO、PASCAL VOC、KITTI 等数据集的实际表现进行交叉分析,对比分析不同模型在多种场景下的优劣。

(3) 面向实践的设计建议: 本文并非简单罗列技术,

而是从模型压缩(剪枝、量化、蒸馏)、训练策略(联邦学习)、架构创新(注意力、Transformer 轻量化)等多维度,提炼出面向工业界部署的、可操作的设计原则与未来方向的建议。

因此,本文旨在为边缘 AI 应用开发者与研究者提供一个从模型选型、性能评估到优化设计的全景式指南,填补了在轻量化目标检测领域兼顾系统性综述与工程实践视角的空白。

1 基于深度学习的目标检测器

近年来,轻量化目标检测研究领域迅速,通过开发新的表征方式、采用最新的机器和深度方法,取得可观的成就^[6]。通用的基于深度学习的目标检测模型已被分为两阶段、一阶段、高级阶段及轻量化目标检测器。

1.1 两阶段目标检测器

两阶段算法或称为双阶段算法,具有区域提议(Region Proposals)和检测头(detection head)两个不同的阶段。第一阶段是使用外部区域提议技术(如 Edge Bo 或 Selective Search)中的锚点计算 ROI 候选;第二阶段包括将提取的 RoI 处理成最终的边界框、坐标值和类别标签。两阶段算法的示例包括 Faster R-CNN、Cascade R-CNN、R-FCN 等。(两阶段目标检测器的优点包括通过在既定阶段更好地分析对象,用于高效回归边界框值的多阶段架构以及更好地处理数据集中的类别不平衡。但当检测部分(即使是使用轻量化检测头)与小型骨干结合时,骨干和检测部分会变得不平衡。这种不平衡会增加过拟合的危险并导致重复计算。

两阶段算法具有较高的检测精度,但是模型复杂度高,推理速度慢,不适用于实时检测场景,且二阶段算法的训练时间较长,需要大量数据和算力资源。一般适用于对检测精

1. 深圳中集天达空港设备有限公司 广东深圳 518103

度要求较高的场景，如医学图像分析、卫星图像分析等。

1.2 一阶段目标检测器

为解决二阶段算法面临的不足，研究人员开始研究一阶段检测算法（也称为单阶段检测算法）。一阶段检测算法取消了生成候选框的阶段，将目标识别问题视为回归问题来处理^[7]，直接在卷积神经网络提取的特征上进行目标分类和边界框回归^[8]。

YOLOv1 是卷积神经网络时代的第一个一阶段检测算法，在推理速度上有了明显的进步^[9]，但在小目标检测和密集目标检测中表现较差^[10]。但随着不断地迭代，引入了残差连接和特征金字塔（feature pyramid networks, FPN）增强特征的表达能力并充分利用不同尺度特征（YOLOv3），使用了 CSPDarkNet53 作为主干网络，使用 SPP 结构来扩大感受野（YOLOv4），引入重参数化技术和 SimOTA 标签分配方法等手段提升模型性能，YOLO 的检测精度和推理速度得到了进一步提升^[11]，逐渐成为实时目标检测的重要算法之一。后来的 SSD、RetinaNet 等研究^[12]，明显提升了对小目标和密集目标的检测能力。

1.3 高级阶段目标检测器

最近出现的高级阶段目标检测器（也称为无锚框检测器）移除了一阶段检测器中用于检测对象的锚点概念。CornerNet 将对象检测为配对的关键点，并引入了新的角点池化层以更好地定位角点^[13]。CenterNet 将对象检测为三元组，而不是一对关键点。Foveabox 预测对象的类别敏感语义图和类别无关边界框。高级阶段检测器在定位具有小尺寸、复杂背景和慢速检测速度的多个目标时也遇到困难。这种设计免去了锚框的超参数调优，设计更简单灵活，在某些场景下能达到更高的精度上限，但在关键点匹配、中心点定位等后处理可能较复杂，训练有时会更困难。

1.4 轻量化目标检测模型

在带宽和资源利用方面计算量低的是轻量化目标检测器，这类检测器目前较少，包括 ThunderNet、PP-YOLO、YOLObile、Trident-YOLO、YOLOv4-tiny、Trident-FPN 等。

基于深度学习的目标检测算法已被分为两阶段、一阶段、高级阶段和轻量化检测器，如图 1 所示。Faster R-CNN、Mask R-CNN、Cascade R-CNN、FPN 和 R-FCN 等算法属于两阶段检测器，而 YOLO、SSD、RefineDet 和 RetinaNet 属于一阶段检测器。高级目标检测器有 CornerNet、Objects as Points 和 FoveaBox 等。然而，上述列出的算法通常包含大量通道和卷积层，这需要在边缘设备上大量计算能力进行部署。图 1 中列出的基于深度学习的轻量化目标检测器是专门为资源受限环境设计的^[14]。由于其效率和紧凑性，一阶段

和高级阶段检测器的流水线是设计轻量化目标检测器的行业标准。

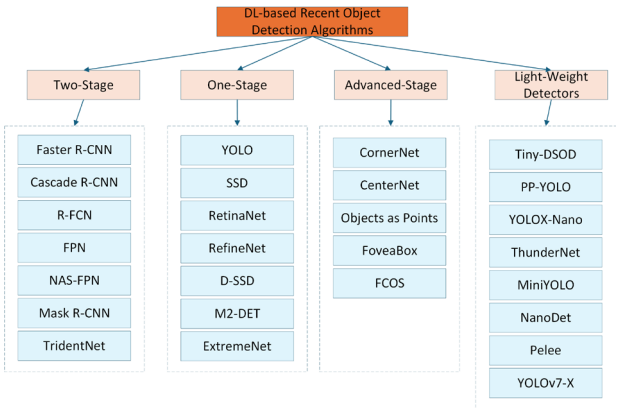


图 1 当前主流基于深度学习的目标检测器分类

2 基于深度学习的轻量化目标检测模型

随着自动驾驶、机器人视觉、智能交通、工业质量检测、目标跟踪等领域和应用的需求暴增，计算机视觉已在很大程度上使用了基于深度学习的目标检测。深度模型通常能提高性能，但其资源密集型的网络限制了将实际应用程序部署到边缘设备上或小型移动场景的设备上。以轻量化移动目标检测器创建极其有效的目标检测作为此问题的解决方案吸引了业界的研究兴趣。

在内存和处理能力受限的设备上执行骨干（backbone）设计的必要性驱动了基于深度学习的轻量化目标识别模型的研究和开发^[15]。近年来，已经提出了许多高效的轻量化骨干架构，例如 MobileNet、ShuffleNet、DetNAS 等。然而，所有这些架构都严重依赖于广泛部署的深度可分离卷积方法^[16]。关于基于深度学习的轻量化目标识别模型，本文将在本节介绍方法和每个组件。

2.1 轻量化目标检测模型的架构方法及综合分析

边缘设备上基于深度学习的轻量化目标检测算法由多个组件组成，包括输入、骨干、颈部和检测头。每个组件的定义和详细信息及常用示例如表 1 所示。

表 1 基于深度学习的轻量化目标检测器的构建块

组件	描述	示例
输入	传递到骨干架构的图像数据	图像，图像块，图像金字塔
骨干	骨干是在标准数据集（如 ImageNet）上预训练的网络，用于特征提取	CSPDarkNet, ShuffleNet, MobileNet, mNASNet, MobileViT
颈部	在骨干和头部之间，用于转换特征图	FPN, PANet, Bi-FPN
轻量化检测头	用于预测对象的类别和边界框	Tiny YOLOv4, NanoDet, Scaled YOLOv4, PP-PicoDet

轻量化目标检测器的输入是图像、图像块或图像金字塔，最先送到轻量化骨干架构中以计算特征图，例如 CSP-DarkNet、ShuffleNet、MobileNet、PeeleeNet^[17]。骨干是基于深度学习的轻量化目标检测架构的一部分，将图像转换为特征图，而颈部通过连接骨干和检测头来转换特征图。输入图像被传递到轻量化骨干架构以计算对象的初始特征向量。骨干网络可以是预训练的网络，也可以是为特征提取而从头开始构建的神经网络，骨干架构的目的是特征提取并输出特征图。随后颈部组件将此特征图转换为所需的特征向量，以根据应用处理各种目标检测任务。轻量化检测头可以被看作专注于提取 RoI 的深度神经网络。此外，某些池化层固定计算出的 RoI 的大小，计算检测到的对象的最终特征。然后，最终特征被传递到分类和回归损失函数，以分配类别标签并回归边界框的坐标值。重复整个过程直到获得具有所需类别标签的边界框的最终回归值^[18]。如图 2 所示的详细方法是一个典型，基于深度学习的轻量化目标检测由三个部分组成，即骨干架构、颈部组件和轻量化头部预测。输入图像被馈送到骨干，其架构将输入图像转换为特征图。

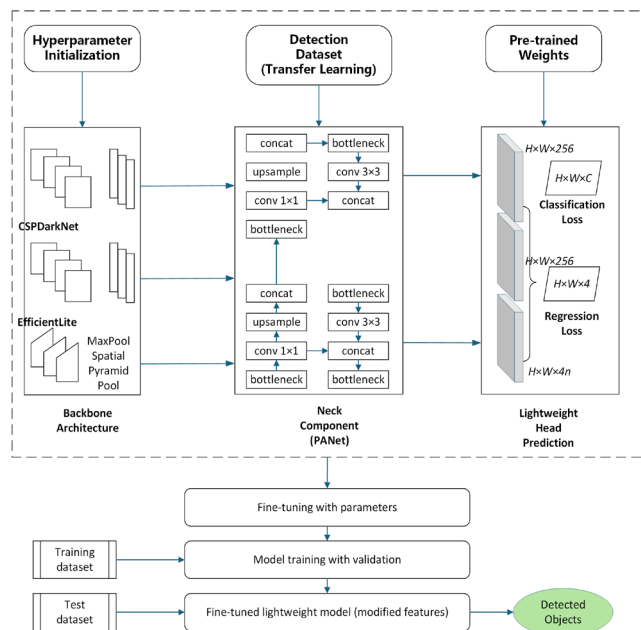


图 2 基于深度学习的轻量化目标检测模型的方法论

对于基于深度学习的轻量化模型，骨干架构可从表 1 中给定的类别中部署。由 Conv2D + 批量归一化 + ReLU 激活函数表示的基本卷积模块构成了骨干架构，通过消除 CNN 优化过程中的冗余梯度信息并将梯度修改集成到特征图中，减少了输入参数和模型大小^[19]。例如，在瓶颈交叉阶段部分暗网模型中，一个 640 px×640 px 的图像被分成四个 320 px×320 px 的图像，然后组合成一个 320 px×320 px 的特征图。此外，还包括 SPP 模块以添加不同大小的特征并增加网

络的感受野。通过增强骨干架构和检测头之间的信息流，颈部改变特征图。颈部 PANet 基于 FPN 拓扑构建，用于提供从顶部到底部的强大语义特征。从底部到顶部的 FPN 层也表达了重要的位置特征。

上述架构是轻量化模型的“骨架”，而实现轻量化的“血肉”则依赖于一系列其他核心技术^[20]。当前业界主流的轻量化方式如表 2 所示。

表 2 主流轻量化核心技术对比

技术	核心思想	代表应用 / 模型	优点	挑战
高效卷积设计	替换标准卷积，大幅减少计算量。	MobileNet（深度可分离卷积）、ShuffleNet（通道混洗 + 组卷积）	计算效率提升显著，是骨干网络轻量化的基石	可能带来特征表达能力下降，需精细设计
			不改变训练流程，部署友好，大幅减少存储与计算	
模型压缩（后处理）	对训练好的模型进行“瘦身”。	TensorRT/Paddle Lite 部署时的量化；通道剪枝	友好，大幅减少存储与计算	存在精度损失，需要重新训练或校准
神经网络架构搜索 (NAS)	自动化搜索最优轻量架构。	MobileNetV3, EfficientNet-Lite, DetNAS	在给定约束（如延迟）下找到帕累托最优解（Pareto optimal solution）	搜索成本极高，依赖强大算力
	用大模型（教师）蒸馏指导小模型（学生）训练。	PP-YOLOE, 许多改进 YOLO 的工作	能提升小模型的精度上限，实现“小而精”	需要预先训练好的大模型，训练过程更复杂

需要注意的是，训练、推理和隐私与安全，都是基于深度学习的轻量化模型需要关注的重要内容。

用于评价轻量化目标检测模型的指标通常有以下几个：

（1）FLOPs：进行一次前向推理所需的浮点运算次数。此指标主要衡量模型的理论计算复杂度，FLOPs 与能耗强相关，对轻量化模型的应用非常重要。但需要注意的是 FLOPs 低不等于推理快，最终速度取决于模型所运行的硬件架构、内存带宽、软件优化水平，在实际目标平台上测试推理时间更具有实际指导意义。

（2）推理时间：处理单张图片从输入到输出所需的时间（单位：ms），推理时间通常随着网络复杂性和规模而增加，也是最直接的用户体验指标。

（3）平均精度：AP（average precision）是评估目标检测模型在单一特定类别上准确性的最核心指标，包括模型的两个关键能力。

(4) 查准率 (Precision)：也称准确率，模型预测出的边界框中，有多少是正确的（即预测框与真实框的重叠度足够高）。

(5) 查全率 (Recall)：也称召回率，数据集中所有真实的目标，模型找出了多少。AP 的数值可以理解为：在召回率从 0% 到 100% 的整个过程中，模型所能保持的平均精度水平。AP 值越高（最高为 100 或 1.0），代表该类检测得越准、越全。

(6) 平均精确率均值：mAP (mean average precision) 是反映模型在各个类别上识别正类样本能力的均衡性和总体水平。

(7) FPS (帧/s)：每秒能处理的图片帧数，与推理时间强相关，用来衡量模型在实际应用中单位时间内处理图像的数量，数值越高意味着模型处理图像的流畅度越高，实时响应性能越好，是评估目标检测模型实时效能的重要依据，30 帧/s 常被视为实时性的门槛。

(8) 参数量：权重和偏差是深度学习中的模型参数，是在整个学习过程中将发现的训练数据的特征。参数总数是神经网络上所有权重和偏差的总和，是模型性能的常见指标。而在存储空间有限或是对边缘计算设备适用的场景下，模型大小是一项关键指标，因此参数量常被用作衡量轻量化目标检测模型性能的一个标准。

以 YOLO 为例，近几年已经开发了基于 YOLO 的轻量化目标检测器^[21]，AP 也在不同的测试中取得了良好的成绩。表 3 为 2023 和 2024 年基于 YOLO 的轻量化目标检测器的性能参数（由于测试集不同，表中未体现 AP 指标）。

表 3 基于 YOLO 的轻量化目标检测器性能参数

检测器	FLOPs /10 ⁹	推理时 间/ms	FPS /(帧/s)	参数量 /10 ⁶	实时性
DSP-YOLO	15.8	—	71.4	7.02	暂不 满足
YOLO-NL	154.6	10.4	96.1	54.76	暂不 满足
EL-YOLO	—	41.0	—	7.56	√
YOLO-S	34.59	—	75.0	7.80	√
GCL-YOLO	10.7	—	53.0	1.64	√
Light YOLO	—	73.0	102.04	3.23	√
Edge YOLO	14.2	—	80.7	5.80	√
GL-YOLO-Lite	3.3	70.51	52.63	4.41	√
LC-YOLO	17.6	15.6	—	7.30	√

2.2 基于深度学习的轻量化目标检测模型的骨干架构

用于图像处理的基于深度学习的模型已经取得了长足的发展，并且在目标分类方面超越了传统方法。

最有效的深度学习目标分类架构是卷积神经网络 (CNN)，其功能类似于人脑，包括实时响应其环境的神经元。基于深度学习的 CNN 架构已被用作基于目标分类的特征提取器微调分类器^[22]。使用随机种子对过滤器和参数进行训练，并通过前向传播进行处理。然而，由于严重资源受限的条件，特别是在内存带宽方面，用于轻量化目标识别模型的专用 CNN 架构的发展受到的关注比预期的要少。在本节中，总结了用于基于深度学习的轻量化目标检测模型的特征提取器。骨干架构用于提取特征以执行轻量化目标识别任务，其中图像作为输入提供，特征图作为输出产生。大多数用于检测任务的骨干架构本质上是用于分类问题的网络，其中考虑了最终的完全连接层，轻量化目标检测网络在视觉上则不同于传统的检测和分类网络。

表 4 列出了近几年出现的骨干架构的参数数量、来源、优势 / 特点和 Top-1 准确性。根据基于深度学习的骨干架构的研究，SqueezeNet 和 ShuffleNetV2 是当今边缘设备中使用最广泛的轻量化骨干架构。由深度可分离卷积、具有线性瓶颈的倒置残差拓扑和自动互补搜索结构构建的模型的性能逐渐由 MobileNet 系列增强^[23]。

表 4 基于深度学习的轻量化目标检测方法中常用的骨干架构

骨干框架	参数 /10 ⁶	Top-1 准 确率 /%	优势 / 特点
SqueezeNet	4.8	60.4	参数比 AlexNet 少 50 倍
ShuffleNetV2	5.2	70.9	引入了通道混洗操作
PeelecNet	5.9	72.1	使用传统卷积构建
DetNAS	5.2	73.8	利用 NAS 设计更好的骨干
MNasNet	5.3	74.0	纳入模型延迟和分解的分层 搜索空间，鼓励层多样性
MobileNetV3	5.4	75.2	允许以顺序方式微调各个层
OFA	7.7	76.0	通过支持多样化的架构设置 降低成本
GhostNet	5.2	73.9	应用低成本变换生成幻影特 征图
MobileViT	6	78.4	面向移动设备的轻量化视觉 Transformer

2.3 基于深度学习的轻量化目标检测器的性能分析

基于深度学习的轻量化目标检测器在准确性和效率之间取得了平衡。尽管前文提到的轻量化检测器具有较快的推理速度，但对于某些任务，准确性并非每次都能达到标准^[24]。

如何客观公正正确地评测轻量化检测器的性能，是非常重要的。数据集对于轻量化目标检测至关重要，因为它们可以对竞争算法进行标准比较，并为解决方案建立目标。

目前比较流行且得到业界公认的数据集如表 5 所示。

表 5 常用轻量化目标检测模型的基准检测数据库

数据集	数量	种类	特点
PASCAL VOC	VOC2007 由 2 501 张训练图像、2 510 张验证图像和 5 011 张测试图像组成	四个大类：动物、车辆、人物和家用物品，20 个类别	具有语义相似性的对象分类，如卡车和公共汽车，增强了检测的复杂性级别。
MS-COCO	包含 328 000 张日常物品和人物的照片	80 个类别	提供了广泛的真实图像，展示了具有各种背景、重叠对象等的混乱场景
KITTI	括 7 481 张用于训练和 7 518 张用于测试的已标记图像。有 100 000 个行人实例、6 000 个 ID		人类类别有两个子类：行人和骑自行车的人。根据对象被遮挡和截断的程度，对象标签分为简单、中等和困难级别

图 3 展示了基于深度学习的轻量化检测模型在上述检测数据集上的性能。轻量化目标检测器 YOLOv4-dense 在 KITTI 数据集上 mAP 值达 84.30%，在 PASCAL VOC 数据集上 mAP 值达 71.60%。L4Net 检测器在 KITTI 上 mAP 值达 71.68%，在 PASCAL VOC 上 mAP 值达 82.30%，在 MS-COCO 上达 42.90%。RefineDet-Lite 检测器在 MS-COCO 数据集上 mAP 值达 26.80%。进一步比较性能，FE_YOLO 在 KITTI 数据集上表现最佳，如图 3 所示，而 L4Net 检测器在 MS-COCO 数据集上表现最佳，最后，轻量化 YOLO-Compact 检测器在 PASCAL VOC 数据集上优于其他检测器。

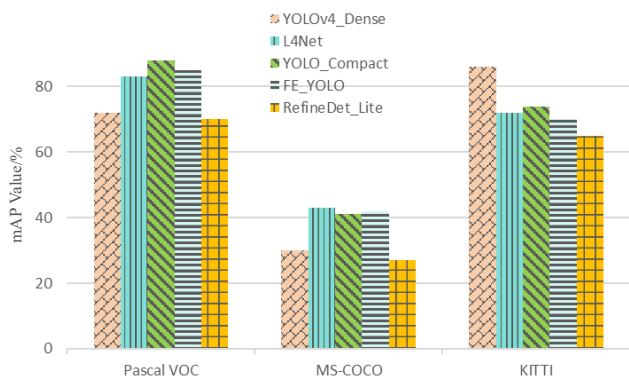


图 3 常见轻量化检测模型在不同数据集上的性能

通过对表 3、表 5 及图 3 的综合分析，可以得出关于轻量化目标检测模型性能权衡的几点关键观察，这些观察对模型选型具有重要指导意义：

精度与速度的固有矛盾：在相同技术代际下，高 mAP 通常以低 FPS 或高 FLOPs（或低实时性或者高能耗）为代价。以 YOLOv7-X 为例（未纳入表 2 但也属前沿技术），mAP 在 COCO 上超 50%，但 FLOPs 高达数百吉，无法直接部署于低端边缘设备。而 PP-PicoDet 或 NanoDet 等模型，通过在精度上做出适度妥协（mAP 约 20%~30%），实现了极低的参数量（ $<1 \times 10^6$ ）和计算量，适用于 MCU 级设备^[25]。

数据集特性对性能评估的影响：同一个模型在不同数据集上可能表现差异显著。如图 3 所示，L4Net 在 PASCAL

VOC 上表现优异（mAP 为 82.3%），但在更复杂、小目标更多的 MS-COCO 上表现一般（mAP 为 42.9%）。这也说明了轻量化模型的设计必须紧密贴合目标应用场景的数据特性。如在车辆检测（KITTI）上优秀的模型（如 YOLOv4-dense），其设计可能侧重于此场景，而非通用检测^[26]。

“实时性”标准的相对性：表 3 中标注“√”的模型均声称满足实时性。但实时性标准因场景而异：自动驾驶要求至少 >30 帧/s，视频监控可能 10 帧/s 即可满足需求，而某些物联网设备仅需 1~2 帧/s。因此，评估时需结合具体业务场景对延迟的要求，而非盲目追求高 FPS^[27]。

3 技术发展趋势

3.1 当前技术状态小结

本文通过对基于深度学习的轻量化目标检测模型的系统梳理，得出以下几点结论：

技术范式收敛：一阶段检测架构（尤其是 YOLO 系列及其变种）因为在速度与精度平衡上的天然优势，已然成为轻量化目标检测的事实标准。无锚框（Anchor-Free）方法因其设计简洁而逐渐成为轻量化设计的新兴主流。

轻量化技术多元化协同：单一的轻量化技术已难以满足业务需求，“高效骨干网络设计（如 MobileNetV3）+ 模型压缩（量化/剪枝）+ 训练优化（知识蒸馏）”的三位一体方案，是当前实现优秀轻量化设计的常见技术路径。

评估维度全面化：对于边缘设备，必须建立涵盖精度（mAP）、效率（FLOPs/参数量）、速度（FPS/延迟）、功耗的多维评估体系。当然，在目标硬件平台上的实测性能才是最终的验收标准。

3.2 现存挑战与未来方向

尽管技术的发展已取得显著进展，但轻量化目标检测在落地边缘设备时仍面临严峻挑战，解决这些挑战也正是未来的研究方向：

小目标检测性能瓶颈：轻量化模型在减少参数和计算的同时，往往首先牺牲了对小目标和密集目标的特征分辨能力。

这是轻量化模型在监控、遥感等场景应用的主要技术瓶颈。

动态环境自适应能力不足：当前模型多为静态优化，难以适应边缘设备上变化的计算资源、输入数据分布（如光照、天气）和任务需求，针对这个挑战，动态神经网络、条件计算是值得探索的方向。

软硬件协同设计深度不足：大多数研究仅在通用 CPU/GPU 上评估，真正的极致效率需要从软件算法层面同时考虑特定硬件架构（如 NPU、DSP、FPGA）的特性，进行硬件感知的算法设计。

数据隐私与分布式学习的矛盾：联邦学习虽能保护隐私，但通信开销和异构数据分布下的收敛问题，在资源受限的边缘网络中是棘手的挑战。

3.3 未来发展趋势分析

随着移动用户和物联网设备产生海量数据，数据增长也将带来诸如延时和数据污染等新问题。此外，如果要从基于深度学习的目标检测和识别算法中实时获取信息，传统方法存在明显短板。相关技术的发展和实际应用的急迫需求使得边缘计算设备引起了广泛关注。随着物联网、5G 和便携式处理设备时代的到来，开发人员如何能够快速完成轻量化检测模型设计和部署边缘应用也就显得至关重要。

由于深度学习领域的显著进步，已经提出了许多针对边缘设备应用的目标识别模型的增强。DeepSense、TinyML、DeepThings 和 DeepIoT 是近年来发布的几个框架，在为物联网边缘设备压缩深度模型取得较好的进步。为了满足基于深度学习的轻量化目标检测器的处理需求，模型在工程应用上必须能够克服几个工程场景下的限制，如严酷的工业环境、有限的电池、有限的计算能力和受限的内存，同时保持一定的准确性水平和合理的续航时间。因此需要创建一个框架，使机器学习模型能够快速在物联网设备中实施。著名的 TinyML 框架包括谷歌的 TensorFlow Lite、微软的 ELL、ARM 的 ARM-NN 和 CMSIS-NN、意法半导体的 STM32Cube-AI 以及弗劳恩霍夫 IMS 的 Alfes，使得在边缘使用深度学习成为可能。当与其他基于微控制器的任务结合时，低延迟、低功耗和低带宽的 AI 算法可以作为智能系统的一部分以低成本运行，大多数轻量化模型采用 TensorFlow Lite 量化，易于在边缘设备上部署，如手机，物联网边缘设备和嵌入式板卡。在交通车道和车牌检测、车辆、驾驶员疲劳检测、行人和障碍物检测、驾驶辅助）、农业（植物病害检测、食品生产、作物收获、害虫管理、入侵检测、智能农业）、军事（目标检测、飞行辅助、军事训练、危险检测）、医疗保健（肿瘤的检测和诊断、手术和训练辅助、细胞体结构检查、患者健康监督和护理）、工业（虚拟助手接入、故障检测、智能工厂、具身机器人），以及室内机器人等方面取得了广泛的应用。

迫切的需求推动了研究的发展，除了用于网络边缘深度学习模型训练过程的专用硬件外，研究人员还创建了新的训练方法，以减少边缘设备中的内存占用并加速低资源设备上的训练。其中用于压缩预训练网络的四个关键类别尤为关键：包括剪枝、量化、知识蒸馏和低秩分解^[28]。

未来轻量化模型研究和发展重要方向主要从以下方式开展。

（1）剪枝：也称为网络剪枝，是减小目标检测模型大小并加速模型推理的一项有用技术。通过剪掉与应用程序无关的神经元之间的连接，此方法降低了分析新输入所需的计算量。除了消除连接外，还可以消除当其大部分权重相对于深度神经网络的整体上下文较低时被视为无关紧要的神经元。利用这种方法，可以在资源优先的设备中使用具有减小尺寸、更高速度和提高内存效率的深度神经网络。如图 4 所示，左侧展示了一个稠密的网络，包含大量冗余连接；右侧展示了稀疏化后的网络，连接变少，结构更紧凑。

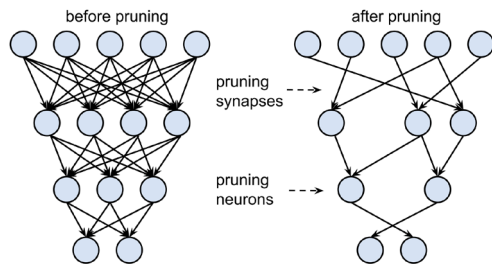


图 4 剪枝前后的突触与神经元

（2）权重量化：权重量化方法以精度换取速度，通过减少浮点参数的数量来缩小模型的存储容量。除了消除不必要的关联外，每个权重都保存为单独的值。权重量化技术旨在通过将相关的权重值聚类为一个值，将这些值压缩为整数或占用尽可能少的空间。因此，权重将重新调整，表明精度也发生了修改。这导致了一个循环实现，其中权重在每次训练后被量化。

（3）知识蒸馏：知识蒸馏呈现为一种新的迁移学习模式。这种技术可以从一个大型且训练有素的深度神经网络（在这种情况下称为教师）中提取知识到一个缩小的深度网络（称为学生）。通过这样做，学生网络可以学会在减小尺寸和提高处理速度的同时实现与教师网络相同的结果。在模型压缩中，因为模型蒸馏可以用于任意网络（从大网络到小网络，从循环网络到卷积网络，从集成网络到单一网络等），极其吸引人，所以被广泛研究，并应用于具体的视觉计算任务之中^[29]。通过知识蒸馏的过程，信息从一个大型的、经过充分训练的端到端检测网络转移到多个更快的子模型。

图 5 是一个用于密集预测任务（语义分割、深度估计和物体检测）的结构化知识蒸馏模型，该模型整合了暗知识的

逐像素蒸馏 (pixel-wise distillation)、马尔科夫随机场的逐配对 / 特征块蒸馏 (pair-wise distillation) 和有条件生成对抗网络的整体蒸馏 (holistic distillation)。图 5 以语义分割任务为例的蒸馏框架。(a) 逐对蒸馏; (b) 逐像素蒸馏; (c) 整体蒸馏。在训练过程中, 该模型将繁琐的网络固定作为教师网络, 仅优化学生网络和判别器网络。具有紧凑架构的学生网络将通过三个蒸馏项和一个特定任务的损失进行训练, 如用于语义分割的交叉熵损失^[30]。

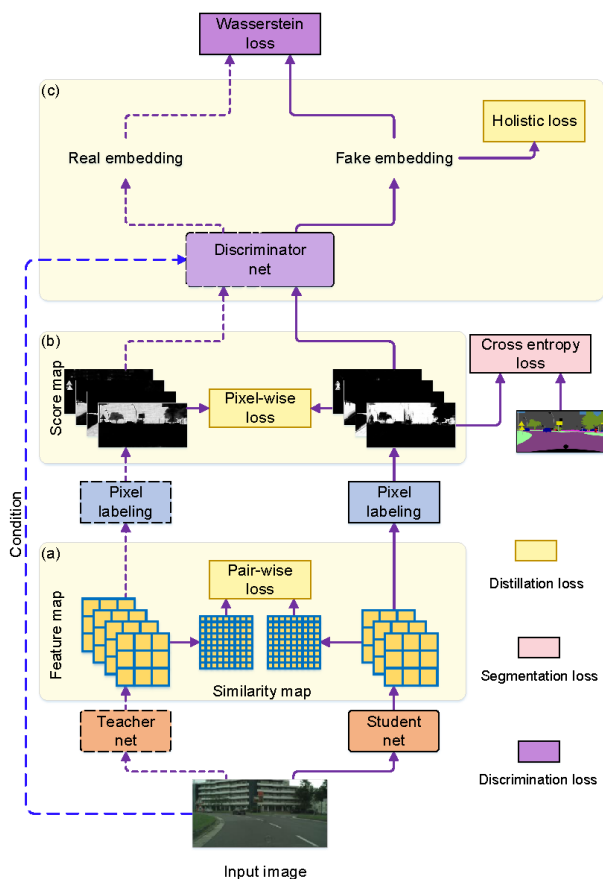


图 5 以语义分割任务为例的蒸馏框架

训练微型网络: 深度神经网络的初始卷积核主要的分解方式采用矩阵分解, 结果的准确性得以显著提高。直接训练微型网络可以大幅减少网络准确性损失并加速推理。

分布式学习和模型分区: 在资源不均的设备集群上, 可以采用模型分割的方式进行分布式联邦训练, 让算力弱的设备只训练模型的前端部分。用联邦学习为不同场景 (如室内 vs 室外摄像头) 协同训练出更通用的前端特征提取器或更适合特定任务的后端检测头。

4 结语

基于深度学习的轻量化目标检测模型是提高神经网络架构硬件效率的良好方式, 可以运行于边缘设备上的模型是未来的重要需求和应用场景。但由于边缘设备本身条件的限制,

对模型的轻量化提出了更高的要求。

良好的模型不仅要在测试数据集上取得优秀的成绩, 更要在边缘设备的硬件上可靠运行满足工程应用, 剪枝、知识蒸馏、权重量化、训练微型网络是基于深度学习的目标检测轻量化模型需要进行突破的关键技术路径。

本文列举和分析轻量化目标检测模型, 希望能够为相关领域的学者和技术人员提供有益的思考方向。

参考文献:

- [1] Chang Libo, Zhang Shengbing, Du Huimin, et al. Position-aware lightweight object detectors with depthwise separable convolutions[J]. Journal of Real-Time Image Processing, 2020, 18: 857-871.
- [2] Chen Long, Ding Qiwei, Zou Qin, et al. DenseLightNet: a light-weight vehicle detection network for autonomous driving[J]. IEEE Transactions on Industrial Electronics, 2020, 67(12): 10600-10609.
- [3] Jiao Licheng, Zhang Fan, Liu Fang, et al. A survey of deep learning-based object detection[J]. IEEE Access, 2019, 7: 128837-128868.
- [4] Kamath V, Renuka A. Deep learning based object detection for resource constrained devices: systematic review, future trends and challenges ahead[J]. Neurocomputing, 2023, 531: 34-60.
- [5] Tian Zhi, Shen Chunhua, Chen Hao, et al. FCOS: a simple and strong anchor-free object detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933.
- [6] Duan Kaiwen, Bai Song, Xie Lingxi, et al. CenterNet: keypoint triplets for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 14589-14604.
- [7] Gagliardi A, de Gioia F, Saponara S. A real-time video smoke detection algorithm based on Kalman filter and CNN[J]. Journal of Real-Time Image Processing, 2021, 18(6): 2085-2095.
- [8] Han S, Yoo J, Kwon S. Real-time vehicle-detection method in bird-view unmanned-aerial-vehicle imagery[J]. Sensors, 2019, 19(18): 3958.
- [9] Haque W A, Arefin S, Shihavuddin A S M, et al. DeepThin: a novel lightweight CNN architecture for traffic sign recognition without GPU requirements[J]. Expert Systems with Applications, 2021, 168: 114481.
- [10] Hu Mengzi, Li Ziyang, Yu Jiong, et al. Efficient-lightweight YOLO: improving small object detection in YOLO for aerial

- images[J]. Sensors, 2023, 23(14):6423.
- [11]Huang R, Pedoeem J, Chen Cuixian .YOLO-LITE: a real-time object detection algorithm optimized for Non-GPU computers[C]//2018 IEEE International Conference on Big Data (Big Data). Piscataway:IEEE,2018:2503–2510.
- [12]Jiang Lin, Nie Wenkang, Zhu Jianyang,et al. Lightweight object detection network model suitable for indoor mobile robots[J]. Journal of Mechanical Science and Technology, 2022, 36(2): 907-920.
- [13]Law H, Deng Jia. CornerNet: detecting objects as paired keypoints[PP/OL].(2019-03-18)[2026-05-22].<https://arxiv.org/abs/1808.01244>
- [14]Kim K, Jang S J, Park J, et al. Lightweight and energy-efficient deep learning accelerator for real-time object detection on edge devices[J]. Sensors, 2023, 23(3): 1185.
- [15]Li Xiang, Wang Wenhai, Wu Lijun, et al.Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook: Curran Associates, 2020: 15482-15493.
- [16]Li Tianyang, Wang Jian, Zhang Tibing.L-DETR: a lightweight detector for end-to-end object detection with transformers[J].IEEE Access, 2022, 10: 103485-103494.
- [17]Liu Jun, Li Qing, Cao Rui, et al.MiniNet: an extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 255-267.
- [18]Liu Yi, Zhang Changsheng,Wu Wenjing,et al. Mini YOLO: a lightweight object detection algorithm that realizes the trade-off between model size and detection accuracy[J]. International Journal of Intelligent Systems, 2022, 37(12): 12135 - 12151.
- [19]Paluru N, Dayal A, Jenssen H B, et al.Anam-Net: anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(3): 932-946.
- [20]S S J, P E R. LittleYOLO-SPP: a delicate real-time vehicle detection algorithm[J]. Optik, 2021, 225:165818.
- [21]Sun Yanwen, Pan Bo, Fu Yili .Lightweight deep neural network for real-time instrument semantic segmentation in robot assisted minimally invasive surgery[J]. IEEE Robotics and Automation Letters,2021,6(2):3870-3877.
- [22]Wang Chunzhi, Wang Zaoning, Li Ke, et al.Lightweight object detection model fused with feature pyramid[J]. Multimedia Tools and Applications,2022,82:601–618.
- [23]Yin Tianrui, Chen Wei, Liu Bo, et al.Light "You Only Look Once": an improved lightweight vehicle-detection model for intelligent vehicles under dark conditions[J]. Mathematics 2024, 12(1):124.
- [24]Zhang Yan, Zhang Haifeng, Huang Qingqing,et al.DsP-YOLO: an anchor-free network with DsPAN for small object detection of multiscale defects[J]. Expert Systems with Applications,2024,241:122669.
- [25]Zhou Long, Wei Suyuan, Cui Zhongma, et al.YOLO-RD: a lightweight object detection network for range doppler radar images[J]. IOP Conference Series: Materials Science and Engineering,2019, 563(4): 042027.
- [26]何明俊. 基于深度学习的目标检测算法轻量化研究 [D]. 广州: 广东工业大学,2025.
- [27]毛义海. 基于边缘计算设备的轻量级目标检测算法研究 [D]. 厦门: 厦门理工学院, 2024.
- [28]Han Song, Pool J, TRAN j, et al. Learning both weights and connections for efficient neural networks[C]//NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems.NewYork:ACM,2015:1135 - 1143.
- [29]Liu Yifan, Chen Ke, Liu C, et al.Structured knowledge distillation for semantic segmentation[C/OL] //2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition(CVPR).Piscataway:IEEE,2019[2026-05-23].<https://ieeexplore.ieee.org/document/8954081>.DOI: 10.1109/CVPR.2019.00271.
- [30]Liu Yifan, Shu Changyong, Wang Jingdong, et al.Structured knowledge distillation for dense prediction [PP/OL].(2020-06-14)[2026-05-23].<https://doi.org/10.48550/arXiv.1903.04197>.

【作者简介】

邓览（1985—），男，安徽宿松人，硕士研究生，高级工程师、研发总监，研究方向：智慧机场、人工智能、机器视觉、工业自动化。

（收稿日期：2026-03-29 修回日期：2026-06-03）