本文引用格式：唐其伟,杜卉然,程伟,等.基于深度图像的人体目标检测方法[J].自动化与信息工程,2024,45(6):73-79;92.
TANG Qiwei, DU Huiran, CHENG Wei, et al. Human object detection method based on depth images[J]. Automation & Information Engineering, 2024,45(6):73-79;92.

基于深度图像的人体目标检测方法

唐其伟　杜卉然　程伟　陈刚
（日立楼宇技术（广州）有限公司，广东 广州 510700）

摘要：针对电梯轿厢场景下，基于传统 RGB 图像实现人体目标检测易受电梯轿厢内饰、光照、乘客外貌特征等因素影响，难以实现稳定、准确的乘客数量检测的问题，提出一种基于深度图像的人体目标检测方法。首先，设计深度图像采集装置，构建人体目标深度图像数据集；然后，提出一种适配专用推理架构的限制采样尺度及线性激活的 YOLOv8 网络优化方法。经实验验证，该方法在边缘计算平台上具有 0.984 的 AP@0.5 准确率及 44 ms 的推理延时，满足电梯轿厢场景下人体目标实时检测的需求，提升了电梯的运行效率及安全性。

关键词：电梯轿厢场景；深度图像；人体目标检测；边缘计算
中图分类号：TP391.41　　**文献标志码：**A　　**文章编号：**1674-2605(2024)06-0011-08
DOI：10.3969/j.issn.1674-2605.2024.06.011　　开放获取

Human Object Detection Method Based on Depth Images

TANG Qiwei　DU Huiran　CHENG Wei　CHEN Gang
(Hitachi Building Technology (Guangzhou) Co., Ltd., Guangzhou 510700, China)

Abstract: Aiming at the problem that human object detection based on traditional RGB images in elevator car scenes is easily affected by factors such as elevator car interior, lighting, and passenger appearance characteristics, making it difficult to achieve stable and accurate passenger quantity detection, a human object detection method based on depth images is proposed. Firstly, design a depth image acquisition device and construct a dataset of human target depth images; Then, a YOLOv8 network optimization method with limited sampling scale and linear activation adapted to a dedicated inference architecture is proposed. Experimental results show that this method has 0.984 AP@0.5 The accuracy and 44 ms inference delay meet the real-time detection requirements of human targets in elevator car scenes, improving the operational efficiency and safety of elevators.

Keywords: elevator car scene; depth image; human object detection; edge computing

0　引言

在电梯轿厢场景下，准确地检测乘客数量及其属性，可改善搭乘电梯的舒适感与安全性，提升电梯群的调度效率[1-2]。以往乘客检测通常采用基于 RGB 图像的人工智能算法模型。但在电梯轿厢内安装 RGB 摄像头存在位置局限性，且不同电梯轿厢的装饰、光照方式[3]具有多样性，以及乘客穿着颜色、佩戴物各不相同，导致基于 RGB 图像的乘客检测算法泛化能力不足[4]。尤其在电梯轿厢内乘客较为拥挤的情况下，因乘客相互局部遮挡[5]，致使检测准确率进一步降低。

基于飞行时间（time of flight, ToF）[6]的传感器受被测对象的纹理、颜色影响较小[7]，可在复杂的纹理和颜色场景下提取目标的深度图像。即使电梯轿厢内的乘客相互局部遮挡，其可见部分在深度图像中仍具有较为清晰的人体特征，与不具备空间信息的色彩数据相比，更有利于实现人体目标的准确检测。

在电梯轿厢场景下，目前尚未发现 ToF 传感器应用于人体目标检测的相关研究。而在非电梯轿厢场景下，ToF 传感器应用于人体目标检测的相关研究主要有基于概率密度函数和直方图的切割算法[8]、基于水

图的图像处理方法[9]等。这些方法的基本思路都是将 ToF 传感器捕获的原始深度图像转换为可视化图像，先求取连通域并计算连通域之间的关系，再根据连通域内的亮度变化（等效于深度图像的距离变化），判别其是否对应人体的头部及肩部，从而实现人体目标检测并进行数量统计。该思路通常能够实现直立人体的检测，但对于弯腰等其他姿态的人体可能存在漏检；同时，与人体头部及肩部深度特征相似的物体，如梯子、箱子等也可能被误检为人体。

为此，本文提出一种基于深度图像的人体目标检测方法。首先，设计基于 ToF 传感器的深度图像采集装置，构建人体目标检测深度图像数据集；然后，结合适配专用推理架构的目标检测模型，实现电梯轿厢内实时、高准确率的乘客检测。

1　深度图像采集装置

深度图像采集装置主要由结构件、MIPI 接口、ToF 传感器模组、图像处理单元等组成，其结构如图 1 所示。

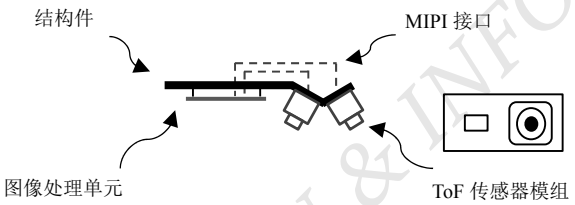

图 1　深度图像采集装置结构

结构件将 ToF 传感器模组固定在电梯轿厢门顶部；ToF 传感器模组安装在结构件的斜面上，镜头表面覆盖了红外滤光片，以降低电梯轿厢内其他波段（如可见光）的能量对 ToF 传感器测距性能的影响；ToF 传感器模组通过 MIPI 接口与图像处理单元连接；图像处理单元将 ToF 传感器模组采集的深度图像解码为每帧分辨率为 320×240 像素的 16 位深度图像，其中，深度轴为被测物体表面各像素点到传感器表面的距离，像素平面为视野范围。深度图像可视化效果如图 2 所示。

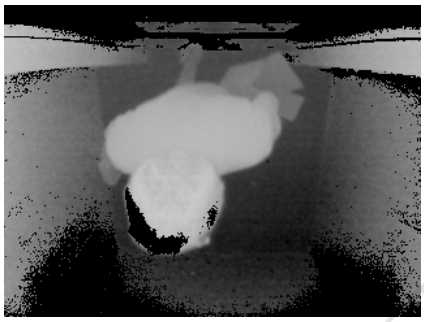

图 2　深度图像可视化效果

ToF 传感器的发光单元以设定的频次主动发射红外脉冲；红外脉冲受被测物体表面的漫反射或多重反射后，回传至 ToF 传感器；ToF 传感器根据发射与接收的红外脉冲相位差，估算红外脉冲从发射到被接收的传播时间，进而估算被测物体表面到 ToF 传感器表面的距离。ToF 传感器的测距原理如图 3 所示。

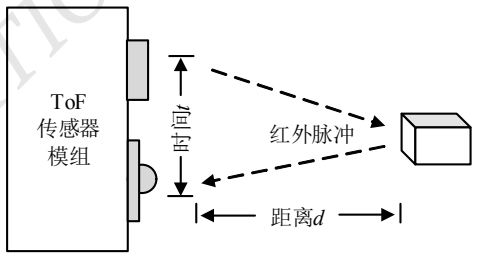

图 3　ToF 传感器测距原理

本文采用 ToF 面阵（分辨率为 320×240 像素）传感器按照设定的帧率采集深度图像，该传感器每个像素点都能估算对应的距离。若某个像素点接收到从被测物返回的红外脉冲时，则其像素值代表传感器表面该像素点到被测物体表面相应成像点的距离；若某个像素点未接收到红外脉冲，则其像素值取零，代表未能获取有效的距离信息。

2　人体目标检测深度图像数据集构建

2.1　数据集

在 ToF 传感器采集深度图像的过程中，被测物体成像的完整性与返回红外脉冲的强度有关。因此，在构建人体目标检测深度图像数据集时，需考虑电梯轿厢内地面材质、乘客穿着颜色、常见携带物品的反射特性差异。人体目标检测深度图像数据采集方案如表

1 所示。其中，ToF 传感器的安装高度和主动光频次作为该采集方案的补充条件，以模拟不同开门高度的电梯轿厢场景。

表 1　人体目标检测深度图像数据采集方案

维度	类型	属性
1	传感器安装高度/m	2.0~2.5，每级 0.1
2	主动光频次/（次/帧）	60 000 或 80 000
3	乘客身高/m	1.0 / 1.2 / 1.6 / 1.8
4	乘客穿着颜色	深色、浅色、白色
5	乘客头饰	帽子、假发
6	乘客数量/个	0~12
7	地面材质	绒面、光滑等
8	常见携带物品	手提物品、行李箱等
9	乘客姿态	直立、弯腰、摆臂等

根据表 1 的采集方案，ToF 传感器以 15 F/s 的输出帧率采集电梯轿厢场景下的深度图像样本，并去除图像相似度较高的样本（图像缓慢变化过程中的前后几帧），形成包含 162 368 个样本的人体目标检测深度图像数据集。

2.2　深度图像预处理

人体目标检测深度图像数据集中的深度图像数据为 16 位，精度为 mm 级。考虑到在电梯轿厢场景下，人体头部及肩部与 ToF 传感器之间的距离通常在 0.5~1.5 m 范围内，根据目标检测算法对输入数据的精度要求，以及电梯轿厢场景下的有效测量范围和精度要求，对 16 位深度图像进行线性变换，得到的 8 位单通道图像如图 4 所示。

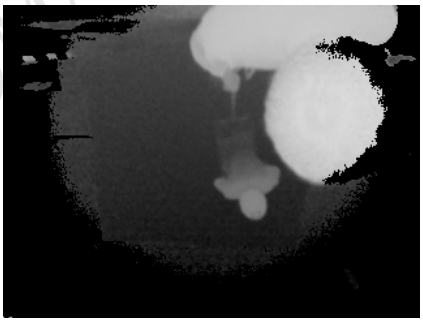

图 4　深度图像预处理后的 8 位单通道图像

图 4 中，深度图像的像素精度由 mm 变为 cm，在满足人体目标检测精度的同时，降低了算力消耗。

2.3　标注方式

深度图像预处理后的 8 位单通道图像以矩形框的方式标注人体目标，如图 5 所示。

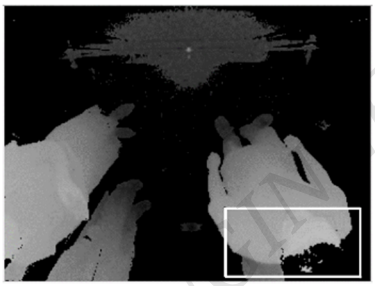

图 5　深度图像标注人体目标

当深度图像中的人体头部及肩部同时可见，且可见部分超过头部及肩部整体区域的一半时，以矩形框进行标注，而携带物品、轿厢背景等其他目标则不标注。根据整幅图像及矩形框的尺寸，将 XYXY（矩形框的左上、右下角坐标）格式的标注信息转换为目标检测算法所需的 XYWH（矩形框的中心点坐标及长、宽）格式。

3　适配专用推理架构的目标检测模型

3.1　边缘计算平台

主流的边缘计算平台以神经网络处理单元（neural processing unit, NPU）[10]为核心，能够实现高吞吐量的深度学习模型推理。而目标检测模型推理过程的计算资源消耗主要集中在乘加运算上。NPU 在设计上兼顾了计算速度和推理精度，通常采用 8 位整型精度进行乘加运算。而在服务器上训练目标检测模型时，通常会结合 16 位或 32 位浮点精度（混合精度）进行运算，当模型转换为 8 位整型精度后[11]，其准确率可能会下降。因此，目标检测模型适配 NPU 时，应考虑其准确率是否满足实际的使用需求。

NPU 的计算资源有限，在进行卷积或池化等算子运算时，若核尺寸不合适，会降低这些算子的计算效率。此外，若目标检测模型中包含复杂的非线性函数，将增加模型的推理延时，甚至可能因精度下降而无法

得出准确的结果。因此，目标检测模型适配 NPU 时，需兼顾准确率并尽可能地提升推理效率。

本文综合考虑目标检测算法的准确率和计算效率需求，结合 NPU 对算子的支持特性，以 YOLOv8 网络[12]为基础进行优化设计，旨在不显著影响准确率的前提下，提升 YOLOv8 模型在 NPU 上的推理性能。

3.2 YOLOv8 网络

YOLOv8 网络主要包含主干网络（Backbone）、颈部网络（Neck）、头部网络（Head）3 部分，其结构如图 6 所示。

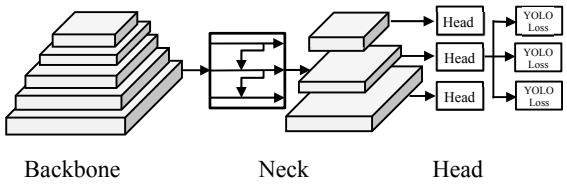

图 6　YOLOv8 网络结构示意图

在 YOLOv8 网络中，计算量占比较大的主干网络采用了较多的类平衡采样（class-balanced sampling, CBS）模块和 1 个空间金字塔池化（spatial pyramid pooling, SPP）模块。这些模块中包含卷积层、批量正则化层[13]及非线性激活函数，导致推理效率较低。为此，本文提出一种限制采样尺度及线性激活的 YOLOv8 网络优化方法，以减少计算量，提高推理效率。

YOLOv8 网络具有不同的规模，以对应不同的算力及准确率需求。本文结合 NPU 的性能及推理实时性，对不同规模的 YOLOv8 网络进行训练及推理。

3.3 YOLOv8 网络优化

YOLOv8 网络的快速空间金字塔池化（spatial pyramid pooling fast, SPPF）模块结构[14]与传统的 SPP 模块结构[15]类似，利用多个池化层对特征图采样，以提升模型对跨尺度目标的鲁棒性。然而，大尺寸的池化核会降低 NPU 的吞吐量。

在电梯轿厢场景下，人体检测目标的尺寸相对有限，且跨尺度的特征提取并非必需。因此，本文利用 3 个 3×3 的卷积层来替代 SPPF 模块（措施 1），在 NPU 上实现较高吞吐量的人体目标特征提取。

YOLOv8 网络中 CBS 模块的 SiLU 激活函数增加了 NPU 的推理时间[16]。SiLU 激活函数的计算公式为

$$SiLU\left(x\right)=x\cdot Sigmoid\left(x\right) \tag{1}$$

$$Sigmoid\left(x\right)=\frac{1}{1+e^{-x}} \tag{2}$$

本文利用 ReLU 激活函数[17]替换 SiLU 激活函数（措施 2）。ReLU 激活函数的计算公式为

$$ReLU=\max\left(0,x\right) \tag{3}$$

经实验验证，相对于 SiLU 激活函数，ReLU 激活函数在保持近似检测准确率的同时，可有效降低模型在 NPU 上的推理延时。SiLU 激活函数与 ReLU 激活函数的输出曲线差异如图 7 所示。

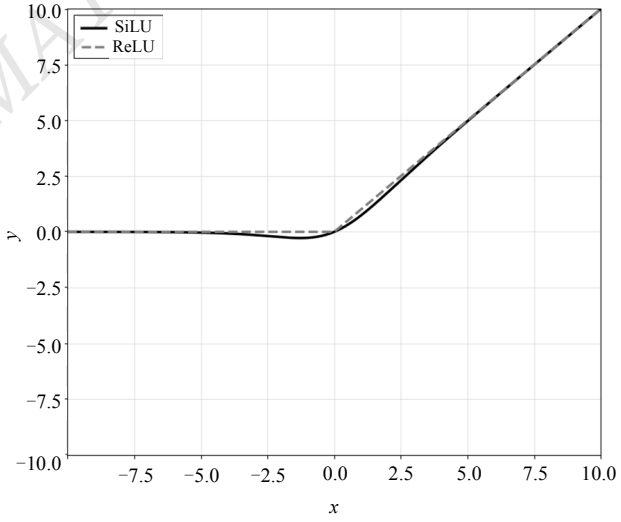

图 7　SiLU 与 ReLU 激活函数的输出曲线差异

利用图形处理单元（graphics processing unit, GPU）训练的 YOLOv8 模型为 16 位浮点精度，需要将其量化为 8 位整型精度，才能最大程度地发挥 NPU 的计算效率。

首先，对 YOLOv8 模型的卷积层与批量正则化层进行线性融合，以减少乘加运算的次数。

卷积层表示为

$$Y=X\cdot W+B \tag{4}$$

式中：Y 为卷积层的输出特征图，X 为卷积层的输入特征图，W 为卷积层的卷积核权重，B 为卷积层的偏移权重。

批量正则化层表示为

$$X' = \frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}}$$
$$Y' = \gamma X' + \beta \tag{5}$$

式中：X'为标准化后的输入特征图，μ为批量特征图的均值，σ^2为批量特征图的方差，ε为浮点数的最小正值（防止分母为零），Y'为正则化后的输出，γ和β分别为可训练的缩放参数和偏移参数。

融合批量正则化层后的卷积层可表示为

$$Y' = X \cdot W' + B' \tag{6}$$

其中

$$W' = \frac{\gamma}{\sqrt{\sigma^2+\varepsilon}} W$$
$$B' = \beta + \frac{\gamma}{\sqrt{\sigma^2+\varepsilon}}(B-\mu) \tag{7}$$

在卷积层与批量正则化层线性融合后，针对每层卷积核的权重或每层卷积后输出特征图的各个通道，单独采用非对称（零点偏移）的量化方式，统计每个通道中 16 位浮点数的最大值、最小值，并将最小值映射为 0，最大值映射为 255，其他浮点数以此为基础进行等比例变换并形成缩放系数。缩放系数 Scale 的计算公式为

$$Scale = \frac{FP16_{\max} - FP16_{\min}}{255-0} \tag{8}$$

式中：$FP16_{\max}$、$FP16_{\min}$分别为某通道浮点数的最大值、最小值。

由于变换后的8位整型精度在后续的计算中需要重新逆变回 16 位浮点精度，因此，还需要进行零点偏移。变换前的浮点数零点以偏移量的方式在 8 位整型上表示，其计算公式为

$$Zero = Round\left(\frac{0 - FP16_{\min}}{255}\right) \tag{9}$$

式中：*Zero* 为零点偏移量，*Round* 为四舍五入取整函数。

上述非对称量化（以偏移量表示浮点数零点）将 16 位浮点数的最大值、最小值分别映射到 8 位整型的 255、0 上，充分利用 8 位整型精度的数值范围，提高变换精度。由于各通道浮点数的最大值、最小值的绝对值通常不相等，此方法可避免对称量化（以 127 表示浮点数零点）时，最大值或最小值无法达到 8 位整型极值的问题。通过多个样本推理，得到每层各通道的零点偏移量及缩放系数，并固化到 YOLOv8 模型中。

在 NPU 上推理时，利用 YOLOv8 模型中已固化的零点偏移量及缩放系数，将每层卷积层的各通道权重及上一层输出的 16 或 32 位浮点精度的特征图变换为整型精度；结合当前卷积层的 8 位权重，使卷积层的乘加运算以 8 位整型精度执行；卷积层的输出根据零点偏移量及缩放系数进行逆变换，恢复至 16 位浮点精度。通过上述过程，可提升 YOLOv8 模型在 NPU 上的推理效率，并降低因计算精度从 16 位下降到 8 位而对模型准确率造成的影响。

4 检测性能试验

4.1 实验数据

本文的实验数据来源于自行采集并构建的人体目标检测深度图像数据集（包含 162 368 个样本），并按照约 9∶1 的比例，随机抽取其中的 146 250 个样本作为训练集，剩余的 16 118 个样本作为验证集。

4.2 实验平台

本实验的服务器端主要用于模型训练及准确率验证，其处理器为双路 AMD EPYC 7763，搭配 16 通道的 DDR4-3200，共 2 TB 的系统内存；图像处理器为八路 Nvidia RTX 3090，共 192 GB 的图形内存。

边缘计算端主要用于推理实验，处理器为 Rockchip RK3568 SoC，搭配单通道 1 GB DDR4-2133 的系统内存，其中 SoC 的 CPU 部分由 4 个 ARM

Cortex-A55 核心组成，NPU 部分为 RKNPU 核心。

实验选取 N、S、M 3 种规模的 YOLOv8 网络进行训练。输入图像的分辨率为 320×240 像素，训练批量为 512 个样本，训练轮次为 300 次。基于 PyTorch 框架结合 8 路 GPU 训练 YOLOv8 网络。以相同的参数对优化前后的 YOLOv8 网络进行训练和对比测试。训练完成后，通过 Rockchip RK3568 配套的工具链，将 YOLOv8 模型转换为 NPU 支持的格式和计算精度。

4.3 性能指标

实验分为目标检测性能验证和乘客数量统计两部分。其中，目标检测的性能指标为 AP@0.5 及 mAP。通常情况下，召回率（recall）上升，会伴随着精度（precision）的下降。平均精度（average precision, AP）是指从小到大统计不同召回率对应的精度并进行积分。AP@0.5 表示样本目标检出区域与样本标注区域的交并比为 0.5 时的 AP。mAP 表示交并比从 0.5 到 0.95（按 0.05 的步长递增）时 AP 指标的平均值，该值越大表示模型的推理结果越接近于样本的标注结果。乘客数量统计是根据每幅图像的检测结果，结合真实结果得出的正确检出正样本（true positive, TP）、错误检出正样本（false positive, FP）、漏检出负样本（false negative, FN）、正确检出负样本（true negative, TN）的数量，计算每幅图像的准确率（accuracy），计算公式为

$$A_{\text{accuracy}} = \frac{TP}{TP + FP + FN} \tag{10}$$

整个验证集的平均准确率（mean accuracy）的计算公式为

$$M_{\text{mean accuracy}} = \frac{\sum_{i=1}^{n} A_{\text{accuracy}}}{n} \tag{11}$$

式中：i 为验证集中第 i 个样本，n 为验证集样本总数。

4.4 实验结果分析

训练完成后，将 YOLOv8 模型分别部署到服务器端及边缘计算端，利用人体目标检测深度图像数据的验证集进行推理并统计 AP 指标，结果如表 2、3 所示。

表 2 验证集服务器端的推理结果

实验序号	YOLOv8 模型规模	优化措施	AP@0.5	*mAP*
1	N	—	0.985	0.701
2	N	措施 1	0.984	0.696
3	N	措施 2	0.982	0.695
4	N	措施 1+2	**0.981**	**0.692**
5	S	—	0.991	0.773
6	S	措施 1	0.991	0.772
7	S	措施 2	0.989	0.770
8	S	措施 1+2	**0.989**	**0.770**
9	M	—	0.993	0.798
10	M	措施 1	0.993	0.796
11	M	措施 2	0.993	0.795
12	M	措施 1+2	**0.992**	**0.795**

由表 2 可知，N、S、M 三种规模的 YOLOv8 模型被优化后，其 AP@0.5、*mAP* 仅有轻微下降，说明优化措施未显著影响模型推理的准确率，且适用于不同规模的模型。

表 3 验证集边缘计算端的推理结果

实验序号	YOLOv8 模型规模	优化措施	AP@0.5	推理延时/ms
1	N	—	0.973	10
2	N	措施 1	0.971	10
3	N	措施 2	0.971	9
4	N	措施 1+2	**0.971**	**9**
5	S	—	0.980	31
6	S	措施 1	0.980	30
7	S	措施 2	0.979	28
8	S	措施 1+2	**0.979**	**28**
9	M	—	0.986	51
10	M	措施 1	0.985	49
11	M	措施 2	0.984	46
12	M	措施 1+2	**0.984**	**44**

由表 3 可知，N、S、M 三种规模的 YOLOv8 模型被优化后，在 NPU 上的推理效率有较大提升，且适用于不同规模的模型，有效降低了电梯轿厢场景下模型的硬件资源需求。

在电梯轿厢内乘客密度最大的情况下，根据 ToF 传感器的不同安装高度，统计不同乘客数量条件下，YOLOv8 模型在 NPU 上推理的平均准确率，结果如表 4 所示。

表 4　人体目标统计平均准确率

实验序号	YOLOv8 模型规模	可见乘客数量/个	样本数/个	平均准确率/%
1	N	1~2	58 023	98.37
2	S			99.47
3	M			99.70
4	N	3~5	45 461	97.92
5	S			99.13
6	M			99.56
7	N	6~8	23 536	96.97
8	S			98.81
9	M			99.43
10	N	9~10	19 230	94.54
11	S			97.16
12	M			98.78

由表 4 可知：利用深度图像的人体目标检测样本可有效训练优化后的 YOLOv8 模型；且本文提出的限制采样尺度及线性激活的 YOLOv8 网络优化方法，能有效提升边缘计算平台对不同规模模型的推理效率，满足电梯轿厢场景下 ToF 传感器深度图像的检测准确率及实时推理需求。

5　结论

本文针对电梯轿厢场景下的人体目标检测需求，设计基于 ToF 传感器的深度图像采集装置，并构建人体目标深度图像数据集；优化了 YOLOv8 网络中的计算模块以适配专用推理架构的特性，提高了人体目标检测的准确率及推理效率。

1） 针对电梯轿厢场景下安装及检测范围的特定需求，设计了深度图像采集装置。该装置可方便地安装在电梯门上，其视野范围完整覆盖检测区域，能有效获取电梯轿厢场景下的人体目标深度图像。

2） 构建的人体目标深度图像数据集，预设了较丰富的场景和目标属性变量；结合 ToF 传感器的成像特性，采集了电梯轿厢场景下不同属性的人体目标深度图像；具备跨电梯轿厢场景的特征，为训练具备泛化能力的目标检测模型提供了数据基础。

3） 根据专用推理架构特性，优化 YOLOv8 网络中的 SPP 模块及激活函数，适配 NPU 计算资源有限以及线性运算的特性，取得检测准确率与推理效率的平衡，并满足实际应用场景的需求。

本文提出的基于深度图像的人体目标检测方法在深度图像采集装置设计、深度图像数据集构建、YOLOv8 网络优化等环节均获得了预期的效果，实现了电梯轿厢场景下 AP@0.5 准确率为 0.984 的人体目标检测，有效缩短了 YOLOv8 模型在边缘计算平台上的推理延时，降低了人体目标检测的推理成本，可实现电梯轿厢场景下乘客信息的实时获取。该方法结合 YOLOv8 模型的推理结果及原始深度图像的三维变换，可估算人体身高，并实现幼童检测等功能。

参考文献

[1] LIU T D, CHEN J, JIANG H. Passenger volume estimation based on the relational model of visual density for elevator group-controlled system[J]. Applied Mechanics and Materials, 2012,127:338-343.

[2] CHUA S N D, LIM S F, S. LAI N, et al. Development of a child detection system with artificial intelligence using object detection method[J]. Journal of Electrical Engineering & Technology, 2019,14(6):2523-2529.

[3] FAN H, ZHU H, YUAN D. People counting in elevator car based on computer vision[J]. IOP Conference Series: Earth and Environmental Science, 2019,252(3):032131.

（下转第 92 页）