DOI：10.11830/ISSN.1000-5013.202305020

采用空间依赖的 MTDPN 扶梯危险行为的姿态估计

林志鸿[1]，郑力新[1]，曾远跃[2]

（1. 华侨大学 工学院，福建 泉州 362021；
2. 福建省特种设备检验研究院 泉州分院，福建 泉州 362021）

摘要： 为实现自动扶梯场景下姿态估计的快速响应和准确估计，提出一种基于空间依赖的多任务解耦姿态网络（MTDPN）。首先，对姿态估计网络进行定位和分类任务分支的解耦，使每个任务分支能够自适应地调整特征关注方向；其次，提出一种空间依赖卷积，通道联合层和空间联合层作为中间层，以逐点卷积和逐深度卷积取代传统卷积，从而降低 MTDPN 的参数量和浮点计算量，使每张图片的检测时间仅为 73.3 ms。在扶梯危险行为关键点数据集和 COCO 关键点数据集上对 MTDPN 进行评估。结果表明：与原始网络 YOLO-POSE 相比，MTDPN 在扶梯危险行为关键点数据集和 COCO 关键点数据集上的准确性指标均有所提高。

关键词： 自动扶梯；人体姿态估计；危险行为检测；任务解耦；空间依赖

中图分类号： TP 391.41；TU 229　　**文献标志码：** A　　**文章编号：** 1000-5013(2023)06-0751-08

Pose Estimation of Escalator Dangerous Behavior Using Spatially-Aware Multi-Task Decoupled Pose Network

LIN Zhihong[1], ZHENG Lixin[1], ZENG Yuanyue[2]

(1. College of Engineering, Huaqiao University, Quanzhou 362021, China;
2. Quanzhou Branch, Fujian Special Equipment Inspection and Research Institute, Quanzhou 362021, China)

Abstract: In order to realize fast response and accurate estimation of pose estimation in escalator scenarios, a multi-task decoupled pose network (MTDPN) based on spatially-aware is proposed. Firstly, the localization and classification task branches are decoupled for the pose estimation network so that each task branch can adaptively adjust the feature focus direction. Secondly, a spatially-aware convolution is proposed, with the channel joint layer and the spatial joint layer as the intermediate layer, replacing traditional convolution with point wise convolution and depth wise convolution, thus reducing the number of parameters and the computation of floating point of the MTDPN, so that the detection time of each image is only 73.3 ms. The MTDPN is evaluated on the escalator dangerous behavior key point dataset and the COCO key point dataset. The results show that the MTDPN has improved accuracy metrics on both the escalator dangerous behavior key point dataset and COCO key point dataset compared to the original network YOLOPOSE.

Keywords: escalator; human pose estimation; dangerous behavior detection; task decoupling; spatially-aware

自动扶梯[1]已经成为商场、医院、车站等公共场所常见的载客设备。但自动扶梯在实际应用中，由于乘坐人员使用不当和应急救援（自救）不及时，容易造成乘坐人员的坠落、碰撞、挤压等事故[2-4]，对人

收稿日期： 2023-05-29

通信作者： 郑力新（1967-），男，博士，教授，主要从事图像分析、机器视觉和深度学习方法的研究。E-mail：zlx@hqu.edu.cn。

基金项目： 福建省科技计划项目（2020Y0039）；福建省泉州市科技计划项目（2020C042R）

体伤害大、危险性高[5]。因此,准确及时的危险行为检测是保障自动扶梯使用的重要前提。

危险行为检测以人体骨架序列[6-7]为研究对象,先提取图像中乘客的姿态信息,再将所提取的骨架序列蕴含的人类行为进行分类,直观地让模型理解目标的行为,进而分析乘客行为的安全性。因此,稳定准确的人体姿态估计(HPE)对自动扶梯的危险行为检测具有重要意义。

早期的HPE仅针对单人目标,主要是基于传统的计算机视觉算法。汤一平等[8]针对少数画面帧中人体对象漏检问题导致的姿态估计丢失,提出基于混合高斯背景差分的方向梯度直方图(HOG)特征匹配算法,该算法能够对粘连的人体对象进行有效分割,从而降低人体对象误检率。但是HOG特征匹配算法对姿态和尺度形变较为敏感,对人物尺度变化较大的自动扶梯监控图像的检测精度不佳。

随着卷积神经网络(CNN)的发展,HPE逐渐扩展到多人姿态估计领域。多人姿态估计根据骨骼关键点生成方式的不同分为自底向上方法和自上而下方法。自底向上方法提取画面中所有可能的骨骼关键点,生成高斯分布概率图[9-12],设计复杂的匹配策略组合人体姿态。自上而下方法[13-18]不同于自底向上方法一次性生成所有关键点,是由人类检测器和姿态估计器构成的两阶段方法。两种检测器将HPE划分为两个阶段:1)人类检测器在监控画面中检测可能存在的乘客类别和定位区域;2)姿态估计器中,对检测到的乘客所在区域使用单人姿态估计生成人体骨骼关键点。相比自底向上方法,自上而下方法具有较强的多人场景建模能力,能够有效避免不同人类骨骼之间的错误连接。不同于CNN模型,一些研究人员把Transformer[19]结构用于关键点位置的预测。TokenPose[20]从大量数据中学习关键点之间的统计约束关系,编码为关键点token。集联Transformer的姿态识别(PRTR)[21]利用自注意力层在Transformer中进行标记化表示,以捕获关键点的关节空间和外观建模。虽然基于Transformer的架构能够在空间和时间域中编码身体关节之间的远程依赖关系,但它们通常需要大规模的训练数据集来实现与卷积网络相比较的性能,这让Transformer的训练和推理变得昂贵。

自动扶梯场景中的实时姿态估计对模型的精度和速度具有一定的要求。YOLOPOSE[22]具有恒定的检测时间和精度优势,能满足扶梯场景中人体姿态估计对准确性和实时性的要求。结合应用环境和模型部署条件,以YOLOPOSE为基线模型,对自上而下方法进行研究。然而,自上而下方法受到2个限制:1)人类估计器依赖人类检测器的检测结果,未识别、错误识别和定位错误的人类乘客都会导致人类姿态估计的失效;2)2种检测器网络的参数量和计算量过于庞大,增加了训练量和推理成本,也增加了危险行为的检测耗时。这2个限制会影响自上而下方法的准确性和计算效率。基于此,本文提出一种基于空间依赖的多任务解耦姿态网络(multi-task decoupled pose network,MTDPN)。

1 基于空间依赖的多任务解耦姿态网络

耦合的人类检测器[23-24]导致不同任务的特征关注方向之间的混淆[25-26],因此,提出多任务解耦姿态网络(MTDPN),允许每个任务独立地学习和调整自己的偏置参数。

1.1 多任务解耦姿态网络

为了让具有不同特征关注方向的分类和定位任务实现各自最佳性能,提出一种多任务解耦姿态网络(MTDPN),将自上而下方法的检测网络解耦成多条不共享支路,以满足不同的视觉任务的特征关注需求。将包含分类、定位和姿态估计3种视觉任务信息的特征金字塔称为多任务耦合特征,在解耦头架构中,多任务耦合特征被拆解为3条不共享的任务分支,表示为

$$\boldsymbol{T}\in\boldsymbol{R}^{H\times W\times C\times(\mathrm{cls},\mathrm{box},\mathrm{conf},N_{\mathrm{kpt}})}, \tag{1}$$

$$\boldsymbol{P}_i=[\boldsymbol{C}_x,\boldsymbol{C}_y,\mathrm{anc_w},\mathrm{anc_h},\mathrm{conf},\mathrm{cls},\boldsymbol{K}_x^1,\boldsymbol{K}_y^1,\boldsymbol{K}_{\mathrm{conf}}^1,\cdots,\boldsymbol{K}_x^n,\boldsymbol{K}_y^n,\boldsymbol{K}_{\mathrm{conf}}^n],\quad i\in[1,\cdots,n]。 \tag{2}$$

式(1),(2)中:$\boldsymbol{T}$为多任务耦合特征;$\boldsymbol{R}$为张量;H,W,C分别为特征图的高、宽和通道数;cls为分类任务;box为定位任务;N_{kpt}为姿态估计任务;$\boldsymbol{P}_i$为特征金字塔中第i层特征,其对应的特征图尺度$F_i\in(H_i,W_i)$;$(\boldsymbol{C}_x,\boldsymbol{C}_y)$为定位任务回归框的中心点坐标;$(\mathrm{anc_w},\mathrm{anc_h})$为当前尺度特征层预设的锚框尺寸;(conf,cls)为当前检测对象的置信度及类别分数;$(\boldsymbol{K}_x^1,\boldsymbol{K}_y^1,\boldsymbol{K}_{\mathrm{conf}}^1,\cdots,\boldsymbol{K}_x^n,\boldsymbol{K}_y^n,\boldsymbol{K}_{\mathrm{conf}}^n)$为一组完整的17个人体骨骼关节坐标信息。

全卷积耦合网络和多任务解耦姿态网络架构,如图1所示。图1中:(a)是原始的全卷积耦合网络,

通过全卷积层输出所有任务的预测结果；(b)是多任务解耦姿态网络；$P_3 \sim P_6$ 分别表示第 3～6 层的特征层；$N_{类别}$ 为网络预测的类别数量，通常为 1；$N_{关键点}$ 为网络估计的人体关键点的数量，通常为 17。

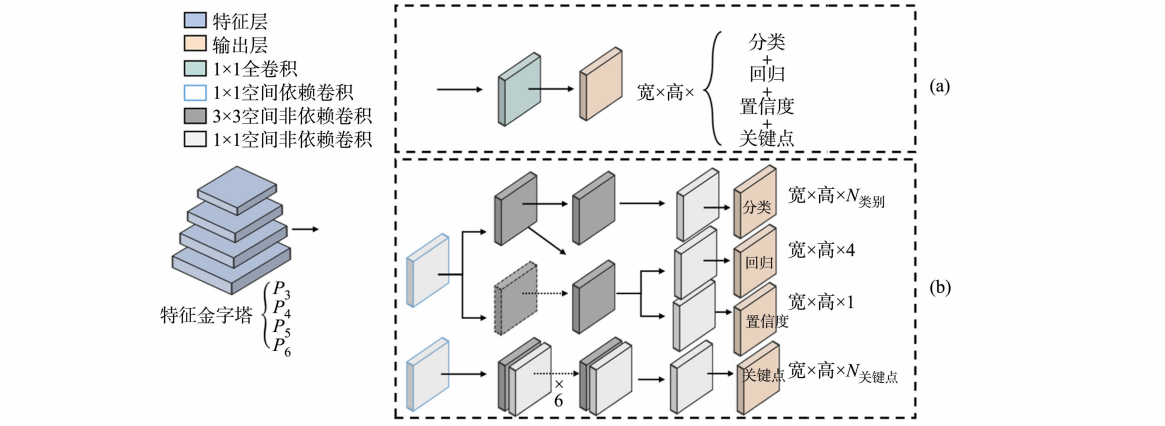

图 1　全卷积耦合网络与多任务解耦姿态网络的架构图

Fig. 1　Architecture of full convolutional coupling network and multi-task decoupled pose network

首先，多任务耦合特征以 1 个卷积核大小为 1×1 的空间非依赖卷积层作为根茎层；然后，连接 2 层卷积核大小为 3×3 的空间非依赖卷积和 1 层卷积核大小为 1×1 的空间非依赖卷积，以构建分类支路；最后，输出全部类别的预测结果。相似地，定位支路遵从分类支路的构建原则，定位支路复用分类支路中第 1 层空间非依赖卷积的特征信息，以减少参数冗余，再构建 1 层卷积核大小为 3×3 的空间非依赖卷积适配定位任务的特征关注方向。由于置信度分支与定位支路的交并比(IoU)判别原则有非常深的联系，因此，置信度分支与定位共享特征信息。

采用 2 层平行卷积核大小为 1×1 的空间非依赖卷积分别输出回归结果和置信度结果。姿态估计分支首先由一层卷积核大小为 1×1 的空间依赖卷积激活支路关注的特征信息；然后，使用重复堆叠 6 次卷积核大小为 3×3 的空间非依赖卷积和卷积核大小为 1×1 的空间非依赖卷积构成卷积块；最后，输出 57 个人体关键点结果。为了防止不同任务特征关注方向的相互影响，人类检测器分支应该不与姿态回归器分支共享任何参数。

1.2　空间依赖卷积

自上而下方法的两种检测器模型和多任务解耦姿态网络的多分支结构引入了庞大的计算成本和复杂的特征表达，增加了自动扶梯场景中自上而下方法应用的优化难度和推理成本。为了降低自上而下方法的庞大计算量与学习难度，提出一种空间依赖卷积(spatially-aware convolution，SA Conv)，结构如图 2 所示。

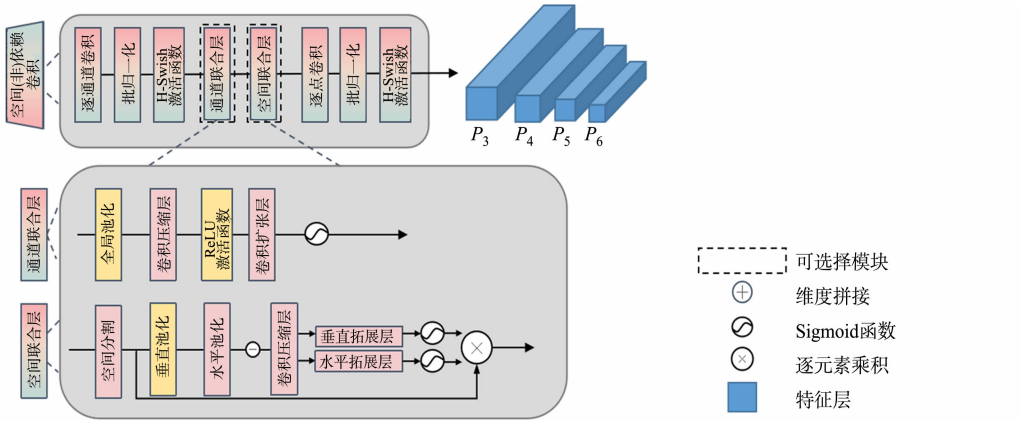

图 2　空间依赖卷积的结构

Fig. 2　Structure of spatially-aware convolution

空间依赖卷积通过逐深度卷积和逐点卷积减少计算消耗。通道联合层对通道特征进行重新缩放，缩小通道，将网络的决策能力集中在最具有区分性和重要性的特征上。放大通道可以增强有用特征的

表示,减少冗余和噪声特征的影响,提高网络的表达能力和泛化性能。空间联合层考虑特征图中垂直和水平特征之间的关系,学习两个不同方向上的权重,并通过分裂和聚合重新对特征加权,提高重要特征的表示能力,以此来弥补轻量化带来的精度损失。两层设计的注意力层进一步降低优化难度。

通道联合层的作用是将每个通道的多任务耦合特征信息编码为单个描述符,以实现通道级别的特征融合。通道联合层中通过全局平均池化(GAP)将每个通道的多任务耦合特征 $\boldsymbol{T}$ 信息编码为单个描述符。它的输出是一个维度为 $\boldsymbol{Z}\in\boldsymbol{R}^{1\times1\times C}$ 的张量。压缩层按照缩放率 r 缩小卷积核挤压多任务耦合特征 $\boldsymbol{T}$ 获得 $\boldsymbol{W}_1$,扩张层重新以缩放率 $\frac{1}{r}$ 放大卷积核获得 $\boldsymbol{W}_2$,σ 激活两层缩放层的前馈网络执行通道联合,激活后的特征以元素乘积方式与初始多任务耦合特征联合得到 $\boldsymbol{F}_1$,其表达为

$$\boldsymbol{F}_1=\boldsymbol{T}\otimes\sigma(\boldsymbol{W}_2(R_{\mathrm{L}}(\boldsymbol{W}_1(\boldsymbol{Z}))。\tag{3}$$

式(3)中:$\boldsymbol{W}_1\in\boldsymbol{R}^{(C,C/r)}$ 为压缩层卷积变换函数;$\boldsymbol{W}_2\in\boldsymbol{R}^{(C/r,C)}$ 为扩张层卷积变换函数;R_{L} 为激活函数 ReLU;σ 为 Sigmoid 激活函数;$\otimes$ 为逐元素乘积符号。

空间联合层作用是通过沿水平和垂直方向的自适应池化进行操作,保持特征图上每组关键点之间的空间关系,并利用这些关系构建方向感知特征图。

$$\boldsymbol{F}_2=\boldsymbol{W}_3\left(\left[\sum_{0\leqslant i\leqslant W}\frac{1}{W}\sum_{0\leqslant i\leqslant H}\boldsymbol{F}_1(H,i),\sum_{0\leqslant i\leqslant H}\frac{1}{H}\sum_{0\leqslant j\leqslant W}\boldsymbol{F}_1(W,j)\right]\right),\tag{4}$$

$$\boldsymbol{F}_3=\partial(\boldsymbol{F}_2)。\tag{5}$$

式(4),(5)中:$[\cdot,\cdot]$ 为特征图沿空间方向上的连接操作;$\boldsymbol{F}_2\in\boldsymbol{R}^{W\times H\times C\times N_{\mathrm{kpt}}}$ 是在垂直与水平两个方向上编码姿态关键点的中间特征图;∂ 为非线性激活函数;$\boldsymbol{W}_3\in\boldsymbol{R}^{(C,C/r)}$。

然后,$\boldsymbol{F}_3$ 沿垂直和水平方向拆分为 $\boldsymbol{F}_3^{\mathrm{h}}\in\boldsymbol{R}^{C/r\times H\times N_{\mathrm{kpt}}(y)}$ 和 $\boldsymbol{F}_3^{\mathrm{w}}\in\boldsymbol{R}^{C/r\times W\times N_{\mathrm{kpt}}(x)}$ 两个独立的张量。2个卷积核大小为 1×1 的空间非依赖卷积变换 W_{h} 和 W_{w},用于将 $\boldsymbol{F}_3^{\mathrm{h}}$ 和 $\boldsymbol{F}_3^{\mathrm{w}}$ 分别变换为与输入 $\boldsymbol{T}$ 具有相同通道数的张量,产生为

$$\boldsymbol{g}^{\mathrm{h}}=\sigma(W_{\mathrm{h}}(\boldsymbol{F}_3^{\mathrm{h}})),\tag{6}$$

$$\boldsymbol{g}^{\mathrm{w}}=\sigma(W_{\mathrm{w}}(\boldsymbol{F}_3^{\mathrm{w}})),\tag{7}$$

$$\boldsymbol{T}^{*}=\boldsymbol{F}_1\otimes\boldsymbol{g}^{\mathrm{h}}\otimes\boldsymbol{g}^{\mathrm{w}}。\tag{8}$$

式(6)~(8)中:$\boldsymbol{g}^{\mathrm{w}}$ 和 $\boldsymbol{g}^{\mathrm{h}}$ 分别为垂直因子和水平因子;$\boldsymbol{T}^{*}$ 为空间依赖卷积输出结果。

最后,通道联合层特征 $\boldsymbol{F}_1$ 与垂直因子 $\boldsymbol{g}^{\mathrm{w}}$ 和水平因子 $\boldsymbol{g}^{\mathrm{h}}$ 共同以元素乘积方法作用,获得空间依赖卷积的结果 $\boldsymbol{T}^{*}$。

2 实验结果与分析

2.1 实验环境

训练时预热阶段的迭代设置为3个轮次,预热期间动量设置为0.8,偏置大小初始化为0.1;初始学习率设置为0.01,优化器使用随机梯度下降法,初始动量为0.937,交并比的阈值设置为0.2,锚框阈值为4.0。数据增强方面,考虑到小目标的识别,通过马赛克法拼接并进行随机随选、翻转、平移等几何操作,提高模型的泛化能力,混合增强使用概率为0.1,图像复制使用概率为0.1。实验在2台 NVIDIA GeForce GTX TITAN Xp GPU 上进行,使用 Python3.8 和 Pytorch 深度学习框架。

2.2 实验数据集

实验数据集来自两个商场的监控视频,通过监控视频及手持相机采集了不同角度下行人在扶梯场景的危险行为,共6 553张图片。通过 Labelme 软件标注每个人类乘客的目标检测标签框,扶梯危险行为关键点数据集部分场景,如图3所示。为每个人类乘客标注17个人体关键点信息,包括鼻子、左眼、右眼、左耳、右耳、左肩、右肩、左肘、右肘、左手腕、右手腕、左臀、右臀、左膝、右膝、左脚踝、右脚踝。每个关键点的坐标信息包括(x,y,v),其中,(x,y)表示该关键点归一化后所在的图像坐标;v 表示该关键点在图像中的可见度,$v\in\{1,2,3\}$,其中,1为完全可见,2为遮挡可见,3为完全不可见。为方便模型训练,扶梯危险行为关键点的可见度皆为1。

数据集划分训练集、验证集和测试集,其中,5 306张乘客状态图片为训练集,657张乘客状态图片

为测试集，590 张乘客状态图片为验证集。图像像素大小为 1 080 px×1 080 px，训练时将图片尺寸统一缩放为 640 px×640 px。

(a) 携带宠物　(b) 携带包裹　(c) 翻越扶手　(d) 跌倒　(e) 逆行

图 3　扶梯危险行为关键点数据集部分场景

Fig. 3　Partial scenarios from escalator dangerous behavior key point dataset

2.3　评价指标

准确率 η_{P} 是指在识别出来的图片中正确识别样本的数量与样本总数量的比例，即

$$\eta_{\mathrm{P}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}。\tag{9}$$

式(9)中：TP 和 FP 分别为正确和错误识别样本数量。

召回率 η_{R} 是测试集中正确识别的样本被分类器正确识别为正的比例，即

$$\eta_{\mathrm{R}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}。\tag{10}$$

式(10)中：FN 为未正确识别的样本数量。

计算每张图片中真实关键点与预测的关键点之间的相似度 S_{OK}，通过所有图片的 S_{OK} 计算平均准确率 η_{AP}，即

$$\eta_{\mathrm{AP}}=\sum\nolimits_{p}\sigma(S_{\mathrm{OK}}>T)\Big/\sum\nolimits_{p}1。\tag{11}$$

式(11)中：p 为当前组的预测值；σ 为标准差；T 为阈值，表示识别的困难程度。

所有关键点类别的 η_{AP} 的平均值(η_{mAP})表示模型在所有关键点类别上的平均性能，其表达式为

$$\eta_{\mathrm{mAP}}=\sum\nolimits_{i=1}^{k}\eta_{\mathrm{AP},i}/k。\tag{12}$$

$\eta_{\mathrm{mAP0.5}}$ 表示 IoU 阈值设置为 0.5 的 η_{mAP}，$\eta_{\mathrm{mAP0.95}}$ 表示 IoU 阈值从 0.50 到 0.95 的 η_{mAP}。η_{mAP} 得分越高，表明模型在测试集上的拟合程度越高。因此，选取 η_{P}，η_{R}，η_{AP}，$\eta_{\mathrm{mAP0.5}}$ 和 $\eta_{\mathrm{mAP0.95}}$ 对姿态估计模型的有效性进行评估。

2.4　空间依赖卷积消融实验

对不同轻量化方法搭建的多任务解耦姿态网络的计算成本进行对比，如表 1 所示。表 1 中：参数量(N_{P})和浮点计算量(N_{F})为网络大小指标，这 2 个指标越小，表明网络占用资源越小；η_{P}，η_{R} 和 $\eta_{\mathrm{mAP0.5}}$ 为精度指标，这 3 个指标越高，说明网络的准确性能越好；每种网络完成 590 张图片推理的时间(t)和每张图片检测时间(t_{p})为速度指标，这 2 个指标越小，说明网络推理图片的速度越快。将图像统一缩放至像素为 640 px×640 px，在相同 GPU 设备上推理相同的 590 张扶梯危险行为关键点测试集。通过上述指标评估空间依赖卷积的轻量化效果、优化性能，以及在自动扶梯中对危险行为检测的及时性。

表 1　不同轻量化方法搭建的多任务解耦姿态网络的计算成本

Tab. 1　Calculation cost of multi-task decoupled pose networks built by different lightweight methods

网络	$N_{\mathrm{P}}/\times10^{6}$	$N_{\mathrm{F}}/\times10^{9}$	$\eta_{\mathrm{P}}/\%$	$\eta_{\mathrm{R}}/\%$	$\eta_{\mathrm{mAP0.5}}/\%$	t/s	t_{p}/ms
全卷积	26.2	61.5	98.7	98.7	99.5	63.5	103.4
空间依赖卷积	16.4	24.9	98.6	98.7	99.4	43.2	73.3
ShuffleNet V2[27]	12.3	18.6	98.2	97.8	97.9	52.0	89.5
EfficientNetV2[28]	29.2	31.2	94.3	94.3	97.9	54.3	92.0

由表 1 可知：全卷积使用最复杂的卷积计算提取图像特征，其精度指标 $\eta_{\mathrm{mAP0.5}}$ 为 99.5%，体现了最好的准确性能，对应产生最大的参数量 26.2×10^{6} 和浮点计算量 61.5×10^{9}，庞大的计算量增加了推理

成本。使用全卷积的计算成本最高,原因可能是在大尺度特征图,如在 $F_0 \in (80,80)$ 分辨率下进行特征提取时,会产生指数式增长的计算成本。空间依赖卷积使用逐点卷积和逐深度卷积替代全卷积,在大分辨率的每个像素点上进行分组卷积,参数量减少了48%,浮点计算量减少了59%,完成590张图片的推理时间最少,仅为43.2 s,每张照片检测时间仅为73.3 ms。同时,空间依赖卷积增强了不同任务的表达能力和学习效果,相比全卷积,其 $\eta_{\text{mAP0.5}}$ 仅降低0.1%。

将目前最新的轻量化网络 ShuffleNet V2[27] 与 EfficientNetV2[28] 进行对比。ShuffleNet V2 将参数量和浮点计算量分别压缩至 12.3×10^6,18.6×10^9,获得了最小的参数量;但由于设计通道重排,每张图片的检测时间相比空间依赖卷积增加了16.2 ms,其 $\eta_{\text{mAP0.5}}$ 较空间依赖卷积降低了1.5%。相比空间依赖卷积,EfficientNetV2 的参数量和浮点计算量分别增加了 12.8×10^6,6.3×10^9,每张图片的检测时间增加了18.7 ms,并且由于缺少对不同任务的强化表达,其 $\eta_{\text{mAP0.5}}$ 相对空间依赖卷积降低了1.5%。

综合表2结果可知,空间依赖卷积在精度指标和速度指标的平衡中取得最优。

2.5 多任务解耦姿态网络消融实验

在扶梯危险行为关键点数据集上评估多任务解耦姿态网络的性能并进行比较,结果如表2所示。由表2可知:在扶梯危险行为关键点数据集中,相比 YOLOPOSE[22] 网络,MTDPN 的 $\eta_{\text{mAP0.5}}$ 和 $\eta_{\text{mAP0.95}}$ 分别提升了0.3%和4.4%,η_{P} 和 η_{R} 分别提高了1.6%和1.8%,这得益于任务解耦架构对姿态估计方法作用;YOLOv7-POSE[29] 具有更高的准确率,这是因为自上而下方法是个复杂的多任务网络,受到目标检测精度的影响,而 YOLOv7-POSE 为不同目标动态分配最佳候选对象,提高了其在目标检测上的准确性,并采用了更加复杂的卷积提取模块,其参数量较 MTDPN 增加了 9.9×10^6,因此,YOLOv7-POSE 网络姿态估计的准确率略高。

表2 多任务解耦姿态网络在扶梯危险行为关键点数据集上的性能比较

Tab.2 Performance comparison of multi-task decoupled pose network on escalator dangerous behavior key point dataset

网络	$N_{\text{P}}/\times10^6$	$N_{\text{F}}/\times10^9$	$\eta_{\text{P}}/\%$	$\eta_{\text{R}}/\%$	$\eta_{\text{mAP0.5}}/\%$	$\eta_{\text{mAP0.95}}/\%$
YOLOPOSE	15.1	20.5	97.2	96.9	99.1	77.2
YOLOv7-POSE	26.3	20.7	99.6	99.5	99.6	87.3
MTDPN	16.4	24.9	98.6	98.7	99.4	81.6

为了进一步评估多任务解耦姿态网络的有效性,将 MTDPN 与自上而下和自底向上的姿态估计网络在 COCO 关键点数据集上中进行性能比较,结果如表3所示。表3中:输入尺寸为输入网络的分辨率指标,输入尺寸越大,网络的准确率越高;N_{P} 和每秒 10^9 次的乘法-加法运算次数(N_{GMACS})为网络大小指标,这两个指标越小,网络占用资源越小;η_{AP},IoU 阈值为0.5的 η_{AP}($\eta_{\text{AP0.5}}$),IoU 阈值为0.75的 η_{AP}($\eta_{\text{AP0.75}}$),检测物体面积大于像素 96 px×96 px 的 η_{AP}($\eta_{\text{AP}^{\text{L}}}$),IoU 阈值范围在[0.5,1.0]的最大召回率的平均值(η_{AR})为精度指标。

表3 多任务解耦姿态网络在COCO关键点数据集上的性能比较

Tab.3 Performance comparison of multi-task decoupled pose network on COCO key point dataset

网络	输入尺寸/px×px	$N_{\text{P}}/\times10^6$	N_{GMACS}	$\eta_{\text{AP}}/\%$	$\eta_{\text{AP0.5}}/\%$	$\eta_{\text{AP0.75}}/\%$	$\eta_{\text{AP}^{\text{L}}}/\%$	$\eta_{\text{AR}}/\%$
Hourglass[30]	512×512	277.8	413.8	56.6	81.8	61.8	67.0	—
PifPaf[31]	—	—	—	66.7	—	—	72.9	—
OpenPose[9]	—	—	—	61.8	84.9	67.5	68.2	66.5
EfficientHRNet-H_0[11]	512×512	23.3	268.8	67.1	—	—	—	—
HigherHRNet[10]	640×640	63.8	308.6	68.4	88.2	75.1	74.2	—
DEKR[32]	512×512	29.6	90.8	67.3	87.9	74.1	76.1	72.4
YOLOPOSE[22]	960×960	15.1	22.8	51.7	80.5	56.1	49.2	56.0
YOLOv7-POSE[29]	640×640	26.3	20.7	58.7	84.5	63.7	72.1	65.3
MTDPN	640×640	16.4	24.9	57.9	83.3	63.6	51.9	66.7

自底向上的方法 Hourglass、HigherHRNet、PifPaf 为每个关键点独立估计高斯分布热图,再通过关节配对方法一次性组合所有的关键点,具有实时性快的优点;自上而下的方法 EfficientHRNet-H_0 通过保持高分辨组合不同特征尺度中的关键点,实现精确的人体姿态估计;DEKR 通过解开每个关键点

独立回归，在检测物体面积大于像素 96 px×96 px 的指标中取得了最佳。

由表 2，3 可知：MTDPN 通过调整不同视觉任务的特征关注方向，有效提升了姿态估计方法的准确率；空间依赖卷积能够增强不同任务的表达能力和学习效果，对乘客特征的关注具有正向作用。

3　结束语

为实现自动扶梯场景下姿态估计方法的快速响应和准确估计，提出一种基于空间依赖的多任务解耦姿态网络，将检测网络解耦为分类、定位两个不共享的任务分支，以满足不同视觉任务的特征关注方向差异的需求，从而实现分类和定位任务各自的最优性能，提高人类检测器的精确度。通过设计空间依赖卷积和空间非依赖卷积网络搭建 MTDPN 的多分支结构，相比全卷积网络，其参数量减少了 48%，浮点计算量减少了 59%，每张图片检测时间仅为 73.3 ms。相比原始网络 YOLOPOSE，MTDPN 在扶梯危险行为关键点数据集的精度指标 $\eta_{\mathrm{mAP0.5}}$ 和 $\eta_{\mathrm{mAP0.95}}$ 分别提高了 0.3% 和 4.4%，在 COCO 关键点数据集的 η_{AP} 提高了 6.2%。推理速度和精度的提升保证了基于自动扶梯危险行为检测的准确估计和速度优势。然而，多分支检测架构会增加模型训练的时间消耗，因此，下一阶段的研究目标是在训练阶段并行融合检测和估计分支，以缩短多分支姿态估计网络的时间训练成本。

参考文献：

[1] 舒文华，欧阳惠卿. 自动扶梯乘客行为智能感知和自主安全管理技术标准探讨[J]. 质量与标准化，2021，354(10)：39-43. DOI：10.3969/j.issn.2095-0918.2021.10.015.

[2] 蒋儒浩. 自动扶梯综合性能检测仪研制[D]. 合肥：合肥工业大学，2019.

[3] 付春平. 自动扶梯几起安全事故的共性分析与探讨[J]. 科技与创新，2023，217(1)：82-84，89. DOI：10.15913/j.cnki.kjycx.2023.01.023.

[4] 张栓柱. 基于事故树的商场电梯事故分析[J]. 消防界(电子版)，2022，8(21)：21-23. DOI：10.16859/j.cnki.cn12-9204/tu.2022.21.040.

[5] 解云蕾. 自动扶梯安全探讨[J]. 中国科技信息，2022(3)：64-66. DOI：10.3969/j.issn.1001-8972.2022.03.021.

[6] CHEN Yucheng，TIAN Yingli，HE Mingyi. Monocular human pose estimation：A survey of deep learning-based methods[J]. Computer Vision and Image Understanding，2020，192：102897. DOI：10.1016/j.cviu.2019.102897.

[7] ZHENG Ce，WU Wenhan，CHEN Chen，*et al*. Deep learning-based human pose estimation：A survey[EB/OL]. (2023-07-03)[2023-09-19]. https：//doi.org/10.48550/arXiv.2012.13392.

[8] 汤一平，杨冠宝，胡飞虎，等. 基于计算机视觉的自动扶梯节能系统[J]. 计算机测量与控制，2011，19(7)：1659-1661，1677. DOI：10.16526/j.cnki.11-4762/tp.2011.07.052.

[9] CAO Zhe，HIDALGO G，SIMON T，*et al*. OpenPose：Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，43(1)：172-186. DOI：10.1109/TPAMI.2019.2929257.

[10] CHENG B，XIAO Bin，WANG Jingdong，*et al*. HigherHRNet：Scale-aware representation learning for bottom-up human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle：IEEE Press，2020：5386-5395. DOI：10.48550/arXiv.1908.10357.

[11] NEFF C，SHETH A，FURGURSON S，*et al*. Efficienthrnet：Efficient scaling for lightweight high-resolution multi-person pose estimation[EB/OL]. (2020-12-30)[2023-09-19]. https：//doi.org/10.48550/arXiv.2007.08090.

[12] 田联房，吴啟超，杜启亮，等. 基于人体骨架序列的手扶电梯乘客异常行为识别[J]. 华南理工大学学报(自然科学版)，2019，47(4)：10-19. DOI：10.12141/j.issn.1000-565X.180186.

[13] FANG Haoshu，XIE Shuqin，LU Cewu，*et al*. RMPE：Regional multi-person pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice：IEEE Press，2017：2334-2343. DOI：10.48550/arXiv.1612.00137.

[14] SUN Ke，XIAO Bin，LIU Dong，*et al*. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE Press，2019：5693-5703. DOI：10.1109/CVPR.2019.00584.

[15] CHEN Yilun，WANG Zhicheng，PENG Yuxiang，*et al*. Cascaded pyramid network for multi-person pose estimation

[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 7103-7112. DOI: 10.48550/arXiv.1711.07319.

[16] ASGARY S, MOTAZEDIAN H R, PARIROKH M, *et al*. KAPAO: A MEMS-based natural guide star adaptive optics system[J]. Iranian Endodontic Journal, 2013, 8(1): 1-5. DOI: 10.1117/12.2005959.

[17] PAPANDREOU G, ZHU T, KANAZAWA N, *et al*. Towards accurate multi-person pose estimation in the wild [C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 3711-3719. DOI: 10.1109/CVPR.2017.395.

[18] 杨学存,李杰华,陈丽媛,等.基于人体骨架的扶梯乘客异常行为识别方法[J/OL].安全与环境学报,2022:1-9 [2023-06-25]. DOI: 10.13637/j.issn.1009-6094.2022.2404.

[19] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010. DOI: 10.48550/arXiv.1706.03762.

[20] LI Yanjie, ZHANG Shoukui, WANG Zhicheng, *et al*. TokenPose: Learning keypoint tokens for human pose estimation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE Press, 2021: 11313-11322. DOI: 10.48550/arXiv.2104.03516.

[21] LI Ke, WANG Shijie, ZHANG Xiang, *et al*. Pose recognition with cascade transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE Press, 2021: 1944-1953. DOI: 10.1109/CVPR46437.2021.00198.

[22] MAJI D, NAGORI S, MATHEW M, *et al*. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE Press, 2022: 2637-2646. DOI: 10.48550/arXiv.2204.06806.

[23] REDMON J, FARHADI A. YOLOv3: An incremental improvement[C]// Computer Vision and Pattern Recognition. Berlin: Springer, 2018, 1804: 1-6. DOI: 10.48550/arXiv.1804.02767.

[24] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-09-08]. https://doi.org/10.48550/arXiv.2004.10934.

[25] WU Yue, CHEN Yinpeng, YUAN Lu, *et al*. Rethinking classification and localization for object detection[EB/OL]. (2020-04-02)[2023-09-07]. https://doi.org/10.48550/arXiv.1904.06493.

[26] SONG Guanglu, LIU Yu, WANG Xiaogang. Revisiting the sibling head in object detector[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Press, 2020: 11563-11572. DOI: 10.1109/CVPR42600.2020.01158.

[27] MA Ningning, ZHANG Xiangyu, ZHENG Haitao. ShuffleNet V2: Practical guidelines for efficient CNN architecture design[C]// European Conference on Computer Vision. [S.l.]: Springer, 2018: 122-138. DOI: 10.1007/978-3-030-01264-9_8.

[28] TAN Mingxing, LE Q V. EfficientNetV2: Smaller models and faster training[C]// International Conference on Machine Learning. [S.l.]: PMLR, 2021: 10096-10106. DOI: 10.48550/arXiv.2104.00298.

[29] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE Press, 2023: 7464-7475. DOI: 10.1109/CVPR52729.2023.00721.

[30] NEWELL A, YANG Kaiyu, DENG Jia. Stacked hourglass networks for human pose estimation[EB/OL]. (2016-07-26)[2023-09-07]. https://doi.org/10.48550/arXiv.1603.06937.

[31] KREISS S, BERTONI L, ALAHI A. PifPaf: Composite fields for human pose estimation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE Press, 2019: 11969-11978. DOI: 10.1109/CVPR.2019.01225.

[32] CAO Xianshuai, SHI Yuliang, YU Han, *et al*. DEKR: Description enhanced knowledge graph for machine learning method recommendation[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information. [S.l.]: ACM, 2021: 203-212. DOI: 10.1145/3404835.3462900.

(**责任编辑**: 黄晓楠 **英文审校**: 吴逢铁)