Computer Science and Application 计算机科学与应用, 2022, 12(12), 2675-2683
Published Online December 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1212271

Hans汉斯

基于增强高分辨率网络的乒乓球击球者姿态估计

申雨钰，王直杰

东华大学，信息科学与技术学院，上海

收稿日期：2022年11月1日；录用日期：2022年11月30日；发布日期：2022年12月5日

摘　要

乒乓球旋转球的识别是乒乓球机器人视觉系统设计中亟待解决的难题，而对乒乓球旋转变化的识别离不开对击球者挥拍姿态的研究，因此本文针对乒乓球击球者姿态分析提出了一种基于增强高分辨率网络的姿态估计算法。该算法融入Ghost模块的YOLOv5人体检测模型结合高分辨率网络姿态估计模型，减少模型的参数量，提升模型运算速度。最后在本文自制的数据集PP-Person上的实验结果表明，本文提出的姿态估计算法有效降低了网络参数量，在保持一定预测精度的情况下，响应速度较HRNet提高了55.76%。

关键词

人体姿态估计，高分辨率网络，YOLOv5，Ghost模块

Human Pose Estimation of Table Tennis Player Based on Enhanced High Resolution Network

Yuyu Shen, Zhijie Wang

College of Information Science and Technology, Donghua University, Shanghai

Received: Nov. 1st, 2022; accepted: Nov. 30th, 2022; published: Dec. 5th, 2022

Abstract

The recognition of table tennis rotation ball is a difficult problem to be solved in the design of table tennis robot vision system, and the recognition of table tennis rotation changes can not be separated from the research on the player's swing posture. Therefore, this paper proposes a pose

文章引用: 申雨钰, 王直杰. 基于增强高分辨率网络的乒乓球击球者姿态估计[J]. 计算机科学与应用, 2022, 12(12): 2675-2683. DOI: 10.12677/csa.2022.1212271

estimation algorithm based on enhanced high resolution network for the analysis of table tennis player's posture. The algorithm integrates YOLOv5 human detection model of Ghost module with high resolution network human pose estimation model, reducing the model parameters and improving the model operation speed. Finally, the experimental results on the self-made data set PP-Person in this paper show that the estimation algorithm proposed in this paper effectively reduces the number of network parameters, and with a certain prediction accuracy, the response speed is increased by 55.76% compared with HRNet.

Keywords

Human Pose Estimation, HRNet, YOLOv5, Ghost Module

1. 引言

随着机器人技术的快速发展，乒乓球机器人等运动类的机器人也逐渐受到广泛关注。乒乓球运动机器人集视觉系统、决策系统、控制系统于一体，机器人系统设计好坏与机器人控制系统反应时间、击打到来球的概率相关。现有研究工作大多关注于乒乓球本身，对乒乓球进行旋转测量[1] [2]或者预测乒乓球的飞行轨迹[3] [4] [5]。而在实际乒乓球运动场景中，运动员无法像机器一样准确地计算乒乓球的运动情况，基本上都是根据对手击球的姿态动作变化进行判断的，因此对乒乓球击球者进行姿态估计，研究其姿态变化有助于乒乓球机器人视觉系统的设计。文献[6]提出了基于轻量化特征融合网络的击球者姿态估计方法，对识别到的对手的姿态序列建立分类模型，判断击球者的挥拍动作。此外，文献[7]利用双目视觉系统获取球拍的三位位置和姿态，采用卡尔曼滤波器跟踪球拍姿态，并采用神经网络对姿态进行分类预判球的旋转类型和速度。但目前来说，由于击球者动作变化速度快且幅度较大，且相似动作太多，所以对于击球者姿态估计的发展还处于初级阶段，相关算法还相对较少，并且目前存在的算法在准确性和实时性上还有待提高[8]。

人体姿态估计在计算机视觉领域发展迅速，姿态估计的任务中包括自顶向下(top-down)和自底向上(bottom-up)两种方式。自底向上的方式是对输入图像的中人体关键点直接进行估计，然后再通过组合算法将各个关键点进行定位，如 OpenPose [9]等。自顶向下的方式就是先对人体进行目标检测，再将人体检测的结果输入到关键点检测的网络中，实现人体姿态估计。因此要准确对击球者进行姿态估计的前提是要能快速且准确定位场景中击球人所处的位置，再通过关键点检测去估计人体所处的击球姿态。人体检测框的误检和冗余都会直接对后续单人姿态估计产生影响，因此在自顶向下的姿态估计算法研究中侧重于对检测器研究，常用的人体检测器有 FasterR-CNN、MaskR-CNN、特征金字塔网络等。

传统人体姿态估计是依赖骨架模型的建立重构人体姿态，随着卷积神经网络(Convolutional Neural Network, CNN)的快速发展，2014 年 Toshev 等提出 DeepPose [10]首次用 CNN 来解决姿态估计问题。由于神经网络层数太深时易出现梯度消失的问题，Wei 等[11]针对这个问题提出了一种多阶段估计方法——卷积姿态机(Convolutional Pose Machine, CPM)。对于一张包含人体姿态的图像，目前普遍采用的检测方法是“从高分辨率处理至低分辨率”，或者“从低分辨率处理至高分辨率”，前者由细节至整体，后者由整体至细节。两种方法同样有效，且适用于不同的场景，大部分研究均选用其中之一，Stacked Hourglass

Network [12]的提出协调地支持了这两种处理方法，多个 Hourglass 模块的叠放可以获取更加丰富的多尺度特征，残差结构的运用又避免了 Hourglass 模块中可能出现的特征信息丢失的情况。Sun K 等[13]认为在这样一个过程中，可能会导致高分辨信息的丢失，因此针对这一问题提出了 HRNet (High-Resolution Network)。HRNet 并行地进行高分辨率的特征提取，始终保持高分辨率表示，确保了高分辨率信息不在低分辨率向高分辨率恢复的过程中丢失，从而提高了人体关键点的预测的准确度。2020 年 Cheng 等人[14]在 HRNet 的基础上提出 Higher-HRNet，加入多尺度监督和反卷积模块来进一步提高分辨率，从而获得更好的预测结果。上述姿态估计算法研究在提升检测精度的同时也增加了参数量和运算复杂度，如何保证姿态估计精度的同时降低运算量是目前人体姿态估计模型改进要研究的问题[15]。

综合上述研究，针对本文要解决的实际问题——乒乓球机器人视觉系统识别对手挥拍动作姿态，提出了基于增强高分辨率的姿态估计算法：1) YOLOv5 作为人体检测器提高检测精度和速度；2) YOLOv5 融入 Ghost 模块减少模型参数量，进一步提高模型检测速度；3) 结合 HRNet 网络提高关键点检测精度，在人体姿态估计的第一步人体检测的过程中降低运算量提高运算速度，同时也保证了关键点检测结果的准确性。

2. 研究方法(Approach)

本文提出的乒乓球击球者姿态估计框架是基于 Top-down 形式的姿态估计方法，主要分为两个阶段。第一阶段，以改进的 YOLOv5 模型快速准确地定位并获取场景中击球人的具体位置，并将人体检测的位置信息传递给第二阶段——关键点检测模块，从而提高姿态估计的识别精度。输入图像提取特征到检测获得击球者所在的位置，然后通过学习击球者的关键点特征信息，并将学习到的信息作为关键点检测器回归出各个关键点的位置及类别的信息，从而得到此时乒乓球运动场景下击球者的姿态信息，整体姿态估计流程如图 1 所示。

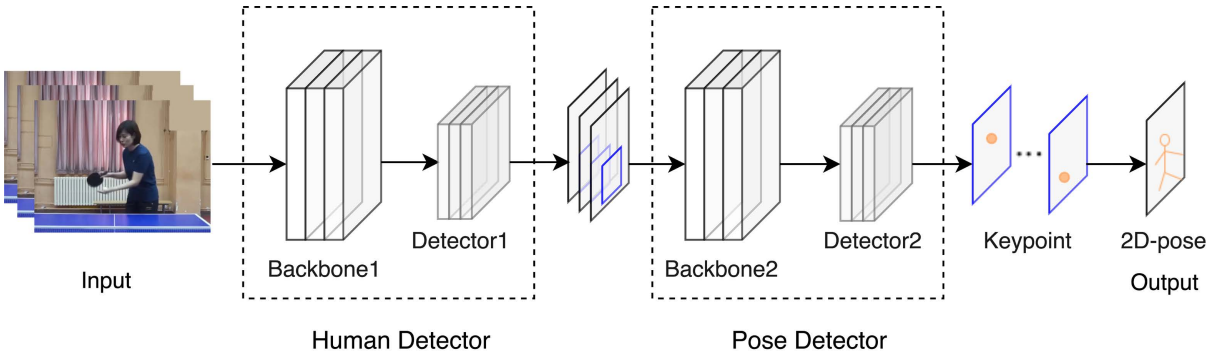


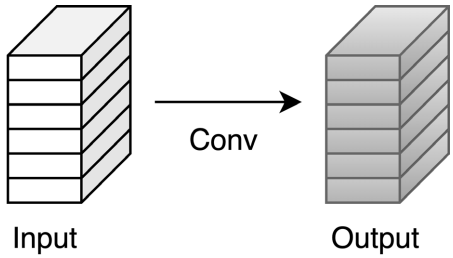
Figure 1. The process of pose estimation
图 1. 姿态估计流程

2.1. 人体检测模块(Human Detector)

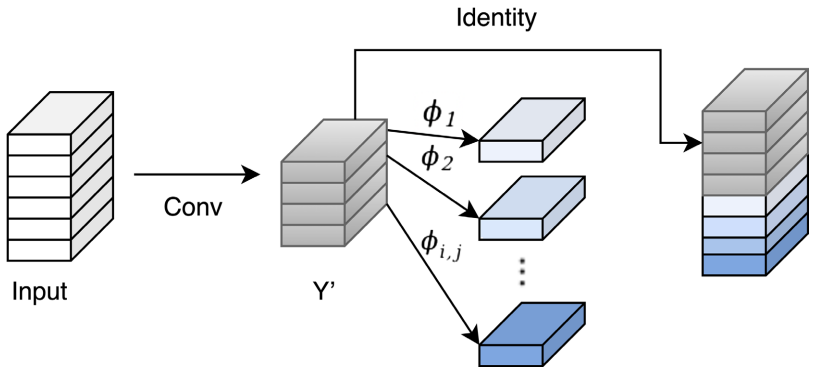
由于乒乓球运动是一项动作以及位置变换都很迅速的运动，需要快速且准确的定位此时击球者所处的位置，尽可能做到实时的人机交互，因此能快速且准确地定位到击球者的位置对于提高后续的姿态估计准确度至关重要。

2.1.1. Ghost 模块

Ghost 模块(图 2)是 GhostNet 模型中针对卷积神经网络中存在大量特征图冗余而提出的轻量化解决方法[15]。它将标准卷积分为以下两部分，由少量计算生成大量的特征图，从而有效减少特征图的冗余问题。



(a) The convlutional layer



(b) The Ghost module

Figure 2. The Ghost module
图 2. Ghost 模块

1) 将输入的特征图 $X \in R^{h*w*c}$，其中 c 为输入通道数，h 和 w 分别为长和宽；卷积核 $f \in R^{c*k*k*n}$，k 和 n 分别为卷积核尺寸和数量，由经典卷积公式(式 1)得到特征图。用少量卷积运算生成 m 个通道的本征特征图 Y' (intrinsic feature maps)；

2) 将 Y'经过线性变换运算 ϕ 生成 Ghost 特征图 $Y_{i,j}$，并将两者进行拼接输出特征图 Y。

$$Y = X * f + b, \tag{1}$$

式中，b 为偏置项，Y 为输出特征图，$*$ 表示卷积运算。

$$y_{i,j} = \phi_{i,j}\left(y_i'\right), \forall i, j = 1, 2, \cdots, m, \tag{2}$$

式中，y_i' 代表 Y'中第 i 个本征特征图，$\phi_{i,j}$ 代表第 j 次线性变换生成的阴影特征 $y_{i,j}$。

最终 Ghost 模块输出 $n = m \times s$ 个特征图，输出的特征图大小为 $h' * w' * n$。

普通卷积和 Ghost 卷积所需运算量分别如式(3)和(4)所示：

$$l_c = n * h * w * c * k * k, \tag{3}$$

$$l_G = n/s * h' * w' * c * k * k + (s-1) * n/s * h' * w' * d * d, \tag{4}$$

其中 $d * d$ 为线性运算核尺寸，s 为线性运算的数量，且 $s \ll c$。

因此，由式(5)可以看出，Ghost 模块在理论上计算量约为普通卷积的 $\frac{1}{s}$，参数量的计算也同样约为其 $\frac{1}{s}$，这说明模型中将 Ghost 模块替换原有的普通卷积能够有效减少模型参数量，加快模型运算速度。

$$r_s = \frac{l_G}{l_C} = \frac{\frac{1}{s} * c * k * k + \frac{s-1}{s} * d * d}{c * k * k} \approx \frac{s + c - 1}{s * c} \approx \frac{1}{s} \tag{5}$$

2.1.2. 融合 Ghost 的 YOLOv5 算法

相较于诸如 CascadeR-CNN、FasterR-CNN 等 Two-stage 类的目标检测模型，如 YOLO、SSD 等 One-stage 模型虽然在检测精度上不如 Two-stage 模型，但在检测速度上更具有优势，在保证一定检测精度的情况下，减少了模型检测的时间。YOLO 系列算法是 One-Stage 目标检测模型中典型的算法，且 YOLOv5 模型是 YOLO 系列算法中具有更高检测精度且推理速度也更快的最新算法，它采用 CSPDarknet 作为 Backbone，并将 FPN (Feature Pyramid Networks)和 PAN (Path Aggregation Networks)相结合做网络的特征融合和加强提取，加强网络的特征融合能力和定位信息，易于在乒乓球运动场景下保证击球者检测精度的同时也能减少检测时间。

YOLOv5 在 YOLOv4 的基础上根据不同通道的尺度缩放，构建了 YOLOv5-N/S/M/L/X 5 种模型，本文选用 YOLOv5s 模型，并在此基础上将网络中颈部层中普通卷积模块替换为 Ghost 卷积模块，降低模型运算量，加快模型检测速度，改进后的 YOLOv5 模型结构如图 3 所示。

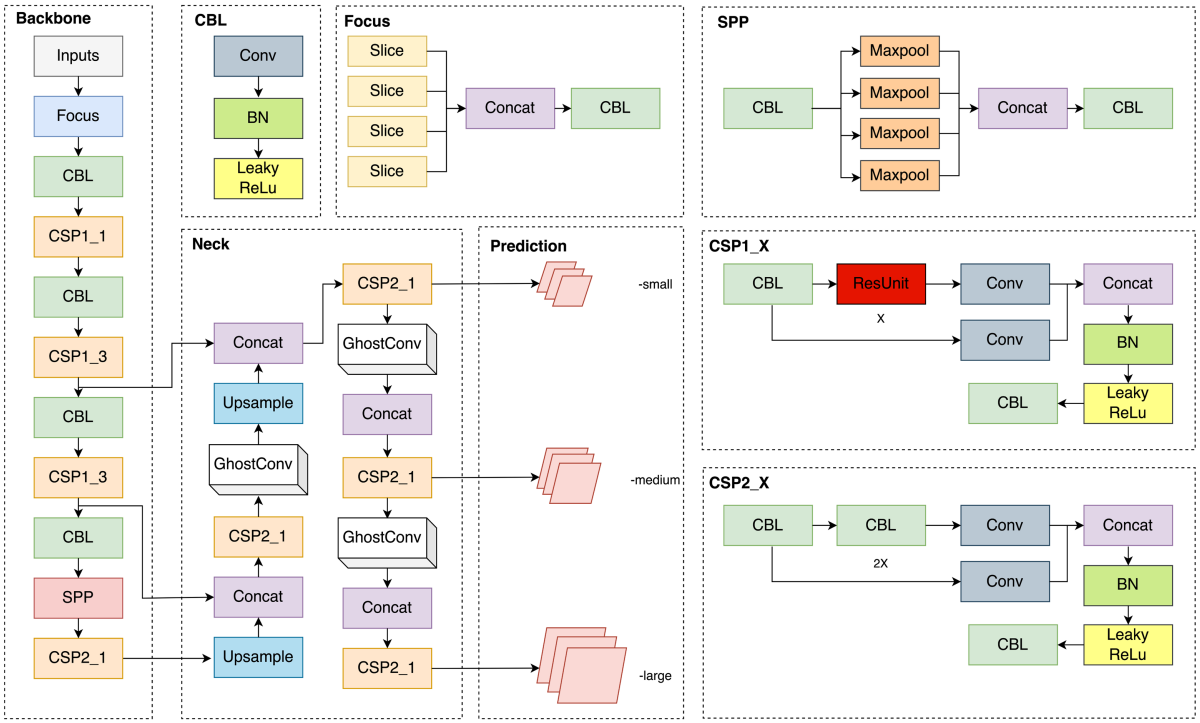


Figure 3. Structure diagram of YOLOv5 model integrating Ghost module
图 3. 融合 Ghost 模块的 YOLOv5 模型结构示意图

2.2. 关键点检测模块(PoseDetector)

将高分辨率网络[13] (High-Resolution Network, HRNet)作为乒乓球运动场景下击球者关键点检测网络的主干网络可以有效提高姿态检测的精度，通过在第一阶段的高分辨率子网络中逐步并行加入低分辨率子网络，同时进行重复的多尺度表征融合，始终保持高分辨率表征，采用的并联结构避免了串联结构方案中可能存在的细节信息丢失。

本文提出的姿态估计模型将基于 Ghost 模块的 YOLOv5 网络作为人体检测器，将人体检测框的结果输入关键点检测网络中对各类关键点进行检测与分类，实现乒乓球运动场景下击球者的姿态估计(如图 4)。

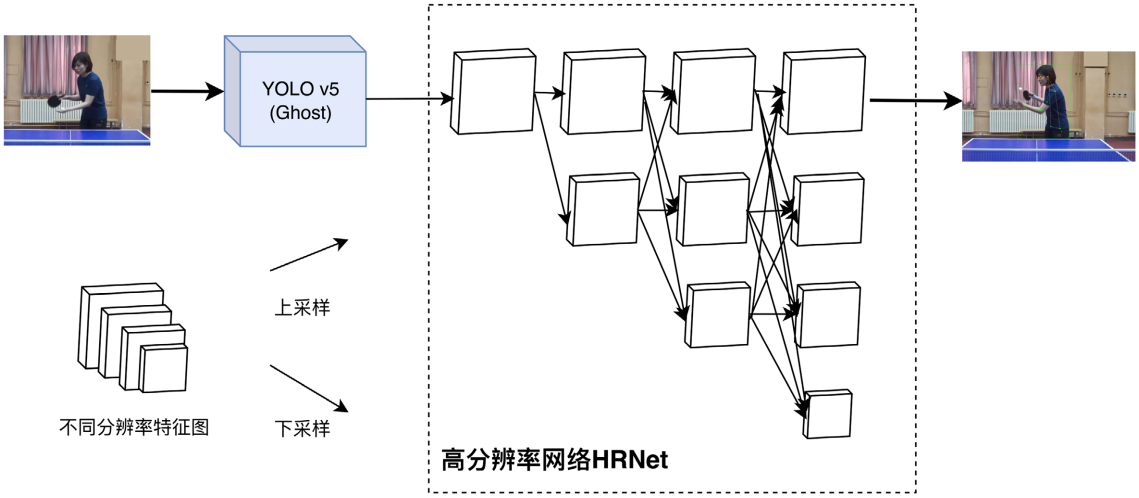


Figure 4. Structure diagram of key point detection module
图 4. 关键点检测模块结构示意图

3. 实验与分析

3.1. 数据集处理

本文针对乒乓球机器人的视觉场景制作了一个特定的数据集 PP-Person，其中数据集主要来源于从网络爬取的乒乓球比赛视频，实验室采集以及公开数据集中抽取，将球桌对面机位的包含不同光照、轻微遮挡程度、姿态的视频作为可用的数据样例(图 5(a)~(d))，而场景中包含大量除击球者以外的无关人员，遮挡程度过高、摄像头在侧面等视频数据作为不可用样例(图 5(e)~(f))。本文共收集了不同场地、不同运动员的乒乓球运动场景视频共 117 段，将每段视频均剪辑至 9 s 的视频片段，并以 30 fps 的帧率将各段视频进行分帧，并抽取每 10 帧中的 1 帧作为实验数据，一共获得乒乓球运动场景下图像 6107 帧。为了降低数据间的关联性，本文将同一视频采集的分帧图像统一放入训练集或者测试集当中，具体来说，本文将上述乒乓球运动员姿态估计数据集中 100 段视频采集的共 5091 帧图像用于训练，将另外剩余的 17 段视频中共 1016 帧图像用于测试，保证两者之间没有交叉。

Figure 5. Table tennis video framing image ((a)~(d): Examples of available data; (e), (f): Unavailable data sample)
图 5. 乒乓球视频分帧图像((a)~(d)：可用数据样例；(e)，(f)：不可用数据样例)

本文研究的是乒乓球机器人视觉场景下的运动员姿态估计，需要检测的目标是机器人视觉系统捕捉采集到的对手，需要检测的关键点主要基于运动员的上半身，而下半身膝盖和脚踝的关键点不是本文研究的范围。因此，本文以 COCO 数据集的标签格式作为范式，在此基础上对关键点的标签进行了修改(如图 6)，以达到更加契合我们研究场景的目的。

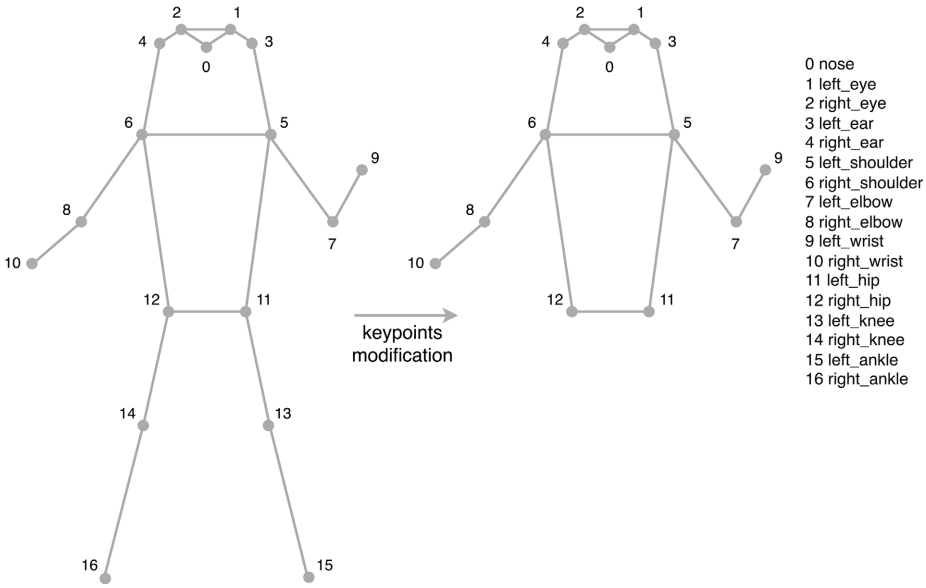


Figure 6. Schematic diagram of keypoints correction
图 6. 关键点修正示意图

3.2. 评价指标

本文采用 COCO 数据集的评测指标——基于关键点相似度(Object Keypoint Similarity, OKS)，其定义如式(6)所示：

$$\mathrm{OKS}=\frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i>0\right)}{\sum_i \delta\left(v_i>0\right)}, \tag{6}$$

其中，d_i 表示预测关键点与对应 grountruth 坐标间的欧氏距离；v_i 代表 groundtruth 是否可见；k_i 为用以区分不同关键点类别的常量；s 表示目标尺度大小。平均精度和召回分数 AP 为当 OKS = 0.50, 0.55, ..., 0.90, 0.95 时取得平均值，AP^{50} 和 AP^{75} 分别表示 OKS = 0.50 和 0.75 时预测关键点的准确率。

3.3. 实验平台与环境配置

本研究实验在 Ubuntu 18.04 系统使用 Python 语言进行程序代码的编译，软件配置为 PyTorchv1.8.0、CUDNNv7.2、CUDAv10.2；硬件配置为 Intel i7-9700K 3.60 GHz 的 CPU，12G NVIDIA 2080Ti 的 GPU。

本文将数据集 PP-Person 中的图像固定纵横比，裁剪到固定尺寸为 256 × 192，并且为了增强提出模型的鲁棒性，对本标注数据集采用 Mosaic 数据增强策略提出的姿态估计模型输入尺寸为 256 × 192 × 3，Batchsize 设置为 32，训练 50 个 epoch。

3.4. 实验验证与分析

本文基于 PP-Person 数据集，对比了 HRNet 与不同的经典人体检测模型结合的姿态估计结果，同时也将本文提出的姿态估计模型与当下热门的人体姿态估计网络模型进行了对比，对比结果如表 1 所示。

Table 1. Comparison results of different attitude estimation model algorithms
表 1. 不同姿态估计模型算法对比结果

	HumanDetector	PoseDetector	AP/%	AP^{50}/%	AP^{75}/%	Time/s
1	FasterR-CNN	Hourglass	67.2	78.3	66.4	0.208
2	FasterR-CNN	Simplebaseline	70.3	83.9	72.8	0.260
3	FasterR-CNN	HRNet	71.2	84.3	75.3	0.217
4	CascadeR-CNN	HRNet	69.1	81.3	70.0	0.284
5	SSD	HRNet	68.7	80.6	68.7	0.181
6	YOLOv5	HRNet	71.3	82.1	73.9	0.120
7 (Ours)	YOLOv5 + Ghost	HRNet	72.6	88.4	78.5	0.096

从上表中可以看出，在本文提出的姿态估计算法模型较其他热门的姿态估计算法模型在本文数据集 PP-Person 上来说关键点的预测精度有所提高，并且在检测速度上最多提高了 66.2%。结果表明，本文提出的算法在保证一定检测精度的情况下，能大大减少模型运行的时间，从而使乒乓球机器人能在较短的时间内捕捉对手的姿态变化，有充足的时间做出控制决策作出高质量回球。

3.5. 结果可视化研究及分析

本文在提出的数据集 PP-Person 上做可视化验证，从验证集中随机选择不同程度的自身遮挡状态以及不同拍摄环境下的图片，得到的检测情况如下图 7。

Figure 7. Visualization results of some models
图 7. 部分模型可视化结果图

从上图可以看出本文提出的模型在运动员姿态变化幅度较大存在自身遮挡时也能较好地检测出各个关键点的位置，在环境光照条件不足等情况下也能较为准确地检测出各个关键点的位置。

4. 结束语

本文针对乒乓球机器人视觉系统提出的自顶向下的人体姿态估计模型将融入 Ghost 模块的 YOLOv5 网络作为姿态估计模型中的人体检测器，在乒乓球机器人视觉场景数据集 PP-Person 上，模型关键点检测精度达到 72.6%，比 Hourglass 提高了 5.4%，比 Simplebaseline 提高了 2.3%，比 HRNet 提高了 1.4%。在提高了关键点检测的精度的同时，降低了模型的运算量，加快了模型的运算速度，实时运行的帧率能达到 10.4 fps/s，速度较 HRNet 提高 55.76%，从而能够帮助乒乓球机器人在较短时间内捕捉并识别到对手的姿态，为今后设计乒乓球机器人视觉系统时分析乒乓球旋转状态提供击球者姿态估计的模型基础。

参考文献

[1] 章逸丰. 快速飞行物体的状态估计和轨迹预测[D]: [博士学位论文]. 杭州: 浙江大学, 2015: 1-7.

[2] 赵永生. 旋转飞行乒乓球的状态估计和轨迹预测[D]: [博士学位论文]. 杭州: 浙江大学, 2017: 29-56.

[3] 任艳青, 方灶军, 徐德, 谭民. 基于模糊神经网络的乒乓球旋转飞行轨迹模式分类[J]. 控制与决策, 2014, 29(2): 263-269.

[4] Wang, H., Cao, C. and Leung, H. (2006) An Improved Locally Weighted Regression for a Converter Re-Vanadium Prediction Modeling. 2006 6*th World Congress on Intelligent Control and Automation*, Dalian, 21-23 June 2006, 1515-1519.

[5] Huang, Y., Xu, D., Tan, M. and Su, H. (2011) Trajectory Prediction of Spinning Ball for Ping-Pong Player Robot. 2011 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, 25-30 September 2011, 3434-3439. https://doi.org/10.1109/IROS.2011.6095044

[6] 谈小峰. 人体检测和姿态识别在乒乓球机器人上的应用研究[D]: [硕士学位论文]. 上海: 东华大学, 2021: 17-72.

[7] Gao, Y., Tebbe, J., Krismer, J. and Zell, A. (2019) Markerless Racket Pose Detection and Stroke Classification Based on Stereo Vision for Table Tennis Robots. 2019 3*rd IEEE International Conference on Robotic Computing*, Naples, 25-27 February 2019, 189-196. https://doi.org/10.1109/IRC.2019.00036

[8] 索芳菲, 季云峰. 乒乓球机器人核心算法研究综述[J]. 电子科技, 2022: 1-8.

[9] Cao Z, Simon, T., Wei, S.-E. and Sheikh, Y. (2017) Realtime Multi-Person 2D Pose Estimation Using Part Affinity Field. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1302-1310. https://doi.org/10.1109/CVPR.2017.143

[10] Toshev, A. and Szegedy, C. (2014) DeepPose: Human Pose Estimation via Deep Neural Networks. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1653-1660. https://doi.org/10.1109/CVPR.2014.214

[11] Wei, S.-E., Ramakrishna, V., Kanade, T. and Sheikh, Y. (2016) Convolutional Pose Machines. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 4724-4732. https://doi.org/10.1109/CVPR.2016.511

[12] Newell, A., Yang, K. and Deng, J. (2016) Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, Cham, 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

[13] Sun, K.，Xiao, B., Liu, D. and Wang, J. (2019) Deep High-Resolution Representation Learning for Human Pose Estimation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 5686-5696. https://doi.org/10.1109/CVPR.2019.00584

[14] Cheng, B., Xiao, B., Wang, J., *et al*. (2020) HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 5385-5394. https://doi.org/10.1109/CVPR42600.2020.00543

[15] 钟宝荣, 吴夏灵. 基于高分辨率网络的轻量型人体姿态估计研究[J]. 计算机工程, 2022: 1-10.