基于深度学习的乒乓球击球人姿态识别及球轨迹预测研究

谈小峰，李伟健，郭润乔，王直杰*
（东华大学 信息科学与技术学院，上海，201620）

摘要 乒乓球机器人视觉系统需要识别和分析击球人的姿态动作，预判乒乓球的初始运动方向及速度，从而控制乒乓球机器人的预运动。且需在球离开球拍瞬间实时捕捉乒乓球在空间中的位置来控制机器人的关节移动实现人机对打。针对上述场景，本文提出了一种基于深度学习的算法实现人体检测、姿态识别及球轨迹预测。该算法基于深度卷积神经网络，首先对一阶段目标检测模型 YOLOv4 进行网络剪枝和压缩，为姿态识别提取目标框。然后基于目标框使用金字塔网络预测人体关键点形成击球人的姿态。再根据从相机坐标系转换到实际空间中的姿态行为判断球初始方向。最后通过对乒乓球的实时位置检测实现轨迹预测和击打点确定。理论研究和实验结果表明，本文算法提升了人体检测的准确度，增强了击球人姿态识别的速度和鲁棒性，对后续乒乓球机器人视觉系统的应用具有很好的指导意义。
关键词 姿态识别；人体检测；轨迹预测；深度卷积神经网络

Research on the player’s posture recognition and ball trajectory prediction of table tennis based on deep learning

TAN Xiao-feng, Li Wei-jian, GUO Run-qiao, WANG Zhi-jie
(School of information science and technology, Donghua University, Shanghai, 201620)

Abstract *The vision system of the table tennis robot needs to recognize and analyze the posture of the player, predict the initial movement direction and speed of the table tennis, so as to control the pre-movement of the table tennis robot. And it is necessary to capture the position of the ping-pong ball in the space in real time at the moment the ball leaves the racket to control the joints’ movement of the robot to realize the human-machine sparring. Aiming at the above scenarios, this paper proposes an algorithm based on deep learning to realize the functions of human body detection, gesture recognition and ball trajectory prediction. The algorithm is based on a deep convolutional neural network. First, the one-stage target detection model YOLOv4 is network pruned and compressed to extract the target frame for gesture recognition. Then we use the pyramid network to predict the key points of the human body based on the target frame to form the posture of the hitter. Then we judge the initial direction of the ball according to the posture behavior transformed from the camera coordinate system to the real space. Finally, the trajectory prediction and hitting point determination are realized by real-time position detection of the table tennis ball. Theoretical research and experimental results show that the algorithm in this paper improves the accuracy of human detection, enhances the speed and robustness of the gesture recognition of the hitter, and has a good guiding significance for the application of the subsequent table tennis robot vision system.*
keywords *gesture recognition; human detection; trajectory prediction; deep convolutional neural network*

引言

乒乓球运动机器人集视觉系统、决策系统、控制系统于一体，能实现不同场景下人机对打，对于辅助球员训练、大众乒乓球运动及智能机器人的研究都具有重要意义。乒乓球运动机器人的视觉系统相当于人的眼睛，是整个系统的核心，用来理解和分析对

收稿日期:
基金项目: 国家自然科学基金，神经网络系统中刺激诱发的可调高频同步振荡及其信息处理的动力学研究（11972115）
作者简介: 谈小峰，男，1997，硕士在读，湖北荆州人，研究方向为目标检测和姿态识别，tansily@163.com； 李伟健，男，1995，硕士在读，安徽阜阳人，研究方向为系统建模、目标检测；郭润乔，男，本科在读，辽宁人，研究方向为计算机控制系统；(通讯作者）王直杰，男，1969，博士，教授，浙江临海人，研究认知神经动力学、智能优化和深度学习，wangzj@dhu.edu.cn。

面击球者的行为动作、球在运动过程中的实时位置。这就需要视觉系统具备目标检测和姿态识别的能力。图 1 是乒乓球机器人视觉系统的功能图。其中①和②分别指的是姿态识别和目标检测。

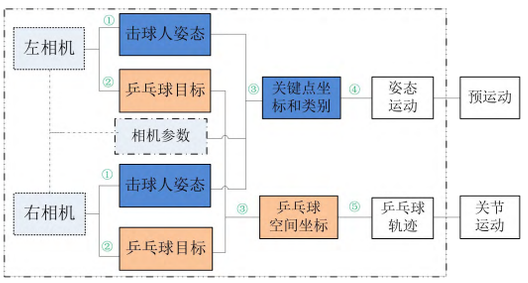

图 1 乒乓球机器人视觉系统功能

传统的目标检测工作大多是基于颜色分割、帧间差以及支持向量机的方法。这些方法严重依赖目标与背景的颜色差，且都属于手工提取特征，存在鲁棒性低、易受光线和背景等外界环境干扰等问题。同样地，早期的姿态识别工作也是基于图模型或者人工推导来推断场景中人体的骨骼序列，这不仅需要耗费大量时间对骨架进行建模，而且准确率也不尽人意。直到卷积神经网络在图像识别方面取得突破性成功，研究者们才开始着力于用深度学习的方法来解决目标检测和姿态识别的问题。

近年来，基于卷积神经网络的目标检测算法主要依托两种思路，一种是基于目标候选框思想的 two-stage，另一种是基于回归思想的 one-stage。Two-stage 先提取目标候选框，再在其基础上训练检测模型，如 Fast R-CNN[1]，Faster R-CNN[2]等。而 one-stage 则没有了目标候选框提取操作，直接利用检测网络产生目标的类别和位置信息，拥有了更高的检测速度，如 SSD[3]，YOLOv3[4]，YOLOv4[5]等。基于深度学习的人体姿态识别也有两种思路，一种是基于自上而下(Top-Down)，先对人体进行目标检测，然后将人体拆分为各个部分对骨骼关键点进行检测比较经典的模型有 RMPE[6]、CPN[7]等。另一种思路是基于自下而上（Bottom-Up），先把人的每类关键点当作一类目标进行检测并分类，再对属于同一人的关键点按照模式组合。其中属 Openpose[8]提出的部分亲和力字段（PAFS）最为经典，后续的大多以此为模板进行改进。Top-Down 算法结果依赖人体检测框的准确度，更适用于高精度人流少的场景。而 Bottom-Up 算法并行检测，速度快，但精度不高，适用于实时性强人数多的场景。

本文在上述研究成果的基础上，通过对模型的改进和重构，提出一种基于深度学习的乒乓球击球人姿态识别及球轨迹预测的方法。首先对论文研究工作的背景和现状进行了介绍，引出本文的主要内容；接着从击球人体检测、姿态识别、球初始运动方向判别和轨迹预测四个方面详细展开叙述；最后给出了实验结果与分析。

1 基于 YOLOv4 模型的人体检测

考虑到场景中乒乓球击球人大多是单人，且希望姿态识别的结果越准确越好。因此采用自上而下的思路，首先基于 YOLOv4 改进模型检测人体目标，再将结果作为姿态识别的输入。

YOLOv4 算法在原有 YOLOv3 目标检测架构基础上，从数据处理、主干网络、网络训练、激活函数、损失函数等各个方面引入一些优化方法，使得模型在检测速度和精度上达到了目前为止的最优匹配。其思想仍是将整张图片作为神经网络的输入，经过网络传递，在输出层直接回归出目标检测框的位置和类别信息。和 YOLOv3 一样，首先通过特征提取网络（也称主干网络）提取输入图像的特征，然后将输入图像划分成 S*S 的网格，目标中心所在的网格负责完成对目标的检测。为了完成对 C 类目标的检测，每个网格需要预测 B 个边界框以及分别属于 C 个类别的条件概率，并输出表征边界框中是否包含目标及输出边界框准确度的置信度信息 $Conf(Object)$

$$IoU = \frac{area(box(P) \cap box(T))}{area(box(P) \cup box(T))} \quad (1)$$

$$Conf(Object) = \Pr(Object) \times IoU \quad (2)$$

$\Pr(Object)$ 用来表示是否有目标物落入该候选网格中，如果有则为 1，没有则为 0。式（1）代表预测框与真实框之间的交并比，其中 $box(P)$ 代表预测框，$box(T)$ 代表真实框。每个预测边界框包含五个参数，分别是 x, y, w, h, $Conf(Object)$，其中(x,y)代表预测框中心与真实框中心的偏移，（w,h）代表预测框的宽和高。模型最终输出的参数为 $S \times S \times (5 \times 5 \times B + C)$。

YOLOv4 的主干网络 CSPDarknet53[9]是算法的核心，用来提取目标特征。为了能

够在轻量化的同时保持准确性、降低计算瓶颈、降低内存成本，在 Darknet53 的每个大残差块上加上 CSP，将基础层的特征映射划分为两部分，再通过跨阶段层次结构合并，在减少计算量的同时保证准确率。图 2 是主干网络结构。

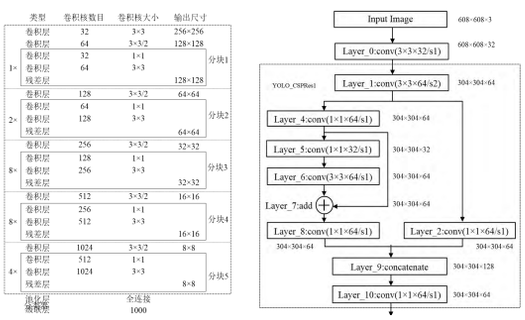

图 2 人体检测模型的主干网络结构

为了后续姿态识别的结果更加精准，需要对人体检测模型进行改进。针对人体目标的特点，对网络的卷积层进行压缩，去掉了第五层和第七层。且将用来检测大中小三个分支进行结构化剪枝，去掉经过多次下采样融合的小分支。这样做的目的是为了在不太影响精度的前提下，降低模型的冗余，减少模型参数量，进而大大加快模型的推理速度。尤其是对于一些富有特别意义的场景完全不需要一个庞大的模型来支撑，这样既浪费空间，又无法达到场景的实时性要求。经过结构化剪枝的模型结构如图 3 所示。可以看到，改进后的模型复杂度大大降低。通过对比实验，也证实了改进模型的效果。实验结果将在第 5 节展示。

模型训练过程中迁移使用 MPII 和 MSCOO 数据集保存参数作为初参。

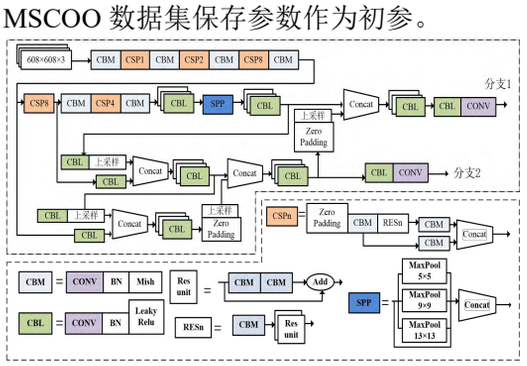

图 3 改进后的 YOLOv4 网络结构

通过爬虫获得少量与乒乓球击球场景相似的数据集再训练。由于爬取的数据是有限的，所以使用 Mosaic 数据增强方式，极大地扩充训练数据集，让模型能够更好地学习到特征。通过实验发现 Mosaic 数据增强的效果要远远好于常规的随机数据增强。

2 基于金字塔网络的姿态识别

在人体目标框确定后，就可以基于每个框进行姿态识别。姿态识别本质是对人体的每一个骨骼关键点进行估计。与一般的目标识别任务不同的是，需要关注更多的高层语义信息。CNN 为了学习到更多特征，通常增加卷积层或者多次上采样，这会造成丰富的低分辨率表示。而金字塔模型通过并行连接高分辨率到低分辨率卷积以保持高层语义信息，并通过重复跨并行执行多尺度融合来增强高分辨率表示。

本文采用金字塔模型 HRNet[10]作为姿态识别部分的模型。将预测得到的人体框结果扩展到固定的长宽比 4：3，然后从图像中裁剪出来，调整为 384*288 的统一尺寸，经过对输入图像的上采样，并聚合低维信息和高维信息，再通过监督学习，输出每个关键点的热图估计。热图越靠近中心置信度越高。经过处理，取每个关键点热图的中心点来表示人体姿态的 2D 识别。2D 的关键点信息局限于平面，而要想将关键点应用到实际的乒乓球场景中，需要转换为 3D 关键点。

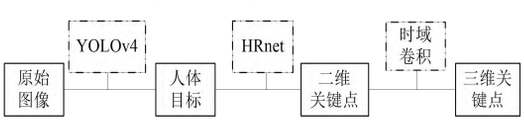

图 4 击球人姿态识别流程

整个流程如图 4 所示。从图像的 2D 人体姿态推理出 3D 姿态可以参考时域空洞卷积[11]，本文直接使用模型得到了不错的效果，这里不做详细介绍。

3 乒乓球的初始运动方向判别

经过前面两节的人体检测和姿态识别，得到了击球者可以检测到的各个 3D 关键点。这些关键点与采样时间结合，就形成了关键点的运动过程。对于一般的回归问题而言，给定一定数量的样本数据，总是希望可以学习得到一个函数 f(x,y,z)=0,使得这些样本最吻合，这个函数就是拟合函数。支持向量机回归就是一个不错的方法，通过计算实

际取值与函数值的差值评价拟合程度。只要差值在场景可以接受的范围，则可以认为该拟合函数是符合要求的。其原理如图所示。

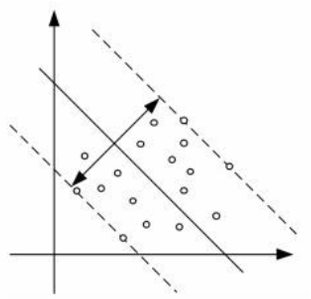

图 5　支持向量回归原理图

为了简化模型拟合难度，选取最关键的手腕关键点和手肘关键点拟合。由于乒乓球击球者发球时的动作随机，手肘和手腕运动没有固定的物理模型，所以利用 Matlab 的 polyfit 拟合函数，经过多次插样取值的方式求出对应的运动函数，再根据运动函数在采样点的斜率就可以计算乒乓球的初始方向和速度。

4 乒乓球的轨迹预测

4.1 乒乓球位置跟踪

乒乓球的位置跟踪调用第 1 节的人体检测模块，但是需要做一些处理。考虑到乒乓球运动过程中速度快目标小的特点，之前经过剪枝的模型不再适用。所以在对模型的预测分支进行剪枝时，设置了两个模块，一个是针对人体目标检测，一个是针对运动过程中乒乓球位置检测。模型可以分时复用，节约整个算法的冗余和浪费。但是针对乒乓球这类小目标检测，根据已有的经验，需要借助 k-means 聚类获取靠谱的先验框。球检测的训练数据集一部分通过双目相机在实验室拍摄采集，另一部分来源于网络乒乓球比赛视频，共组成 5102 张图片，同样经过 Mosaic 数据增强扩充数据集，还能扩展目标的分辨率大小，增强模型的适应性。实验时，对数据集中的标签分布情况进行统计，从 1 开始不断增加聚类中心的个数，经过迭代，分别获得聚类中心。由于使用欧式距离衡量标签相似性会导致大尺寸标签产生更大的误差，影响聚类结果，所以将聚类中心与标签的交并比 $IoU_{(l,c)}$ 作为 K-means 聚类的相似度参数。距离计算公式为 $d=1-IoU_{(l,c)}$。

实验发现，随着聚类中心的个数 k 逐渐增加，聚类距离 d 越小。

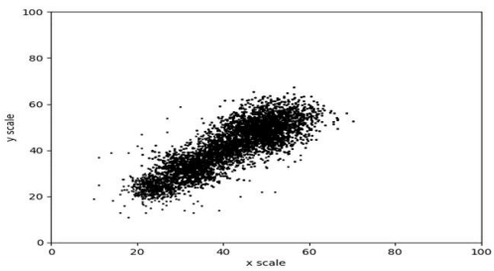

图 6　训练数据集标签分布

但当 k 大于 4 时，d 减小的程度非常缓慢。为了避免冗余，选择 k=4 时的四个聚类中心（15，11），（30，25）（48，35），（58，47）。考虑到 Mosaic 数据增强过程中，不排除有一些乒乓球尺寸被放大，所以加入一组稍大的（62，70）来匹配。在训练迭代次数达到 30000 次基本可以停止迭代，过程中每 100 次迭代进行依次验证。检测结果对比如下表。

表 1 模型检测结果对比

算法	时间 ms	速度 f/s	AP(%)
YOLOv4	47.4	36.16	91.86
改进的 YOLOv4	**41.3**	**39.34**	**94.12**

4.2 坐标变换

通过目标检测方法得到乒乓球在左右相机的不同预测框，取框的中心作为预测像素点位置，再通过视差法得到乒乓球在三维空间中的实际位置。视差法需要通过相机标定，得到左右相机的焦距、主点等内部参数，再通过双目相机校正，使得图像在同一个平面，如图 7 所示。通过坐标转换的方式得到了一系列乒乓球运动过程中的空间坐标，保存相机每次采样的时间。

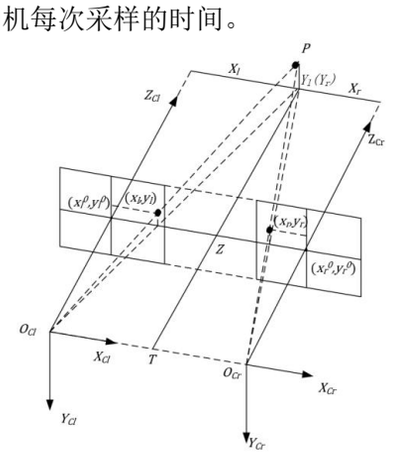

图 7 双目相机校正原理

通过对乒乓球进行受力分析，建立轨迹模型，再进行迭代运算，将在 4.3 轨迹预测部分详细介绍。

4.3 轨迹预测

由于实验条件限制，暂时不考虑乒乓球的反弹，只考虑球离开球拍运行这一段轨迹，期间的运动方程为：

$$m\mathbf{a} = -mg - \frac{1}{2}\rho S C_D \| v \| v \tag{3}$$

其中 m 是乒乓球的质量，**a** 是加速度，g 是重力加速度。S 是乒乓球的横截面积，C_D 是空气阻力系数，v 是乒乓球的瞬时速度矢量。为了方便分析，将乒乓球加速度度分解为 x,y,z 三个方向，有：

$$\mathrm{a}_x = -\frac{1}{2m}\rho S C_D \| v \| v_x \tag{4}$$

$$\mathrm{a}_y = -\frac{1}{2m}\rho S C_D \| v \| v_y \tag{5}$$

$$\mathrm{a}_z = -g - \frac{1}{2m}\rho S C_D \| v \| v_z \tag{6}$$

将公式中不变的量的乘积表示为常数 k，从而得到迭代关系式：

$$d\begin{bmatrix} x \\ y \\ z \\ V_x \\ V_y \\ V_z \end{bmatrix} / dt = \begin{bmatrix} V_x \\ V_y \\ V_z \\ -kVV_x \\ -kVV_y \\ -kVV_z - g \end{bmatrix} \tag{7}$$

将坐标变换得到的乒乓球位置采样点经过噪点消除之后，同样基于支持向量机回归的方式来拟合最佳的运动轨迹。

5 实验结果与分析

（1）关键点检测效果

为了验证本文算法的有效性，实验采用经典数据集 MPII 和 MSCOCO 预训练模型，将参数保存。通过网络爬虫和实验室采集了一批与本文针对的场景高度吻合的数据，共 7038 张，图像分片率大小不一致，在模型训练之前的数据预处理部分可以归一化。并针对统一数据集用 CPN 模型和 RMPE 模型进行对比实验，表 2 是选取的比较关心的四个关键点的结果。关键点的识别准确度仍使用比较流行的 PCK 标准。

表 2 基于 PCK 标准部分关键点准确度对比

算法	左手腕	右手腕	左手肘	右手肘
HRnet	**90.12**	**89.78**	**93.31**	**94.19**
CPN	84.17	85.44	89.72	88.23
RMPE	89.35	88.68	92.15	92.76

图 8 是部分人体检测和姿态识别结果。可以看到，尽管场景中人体的一些手部动作比较夸张，但是本文的算法还是能够较好地把人体可视部分给检测出来。只有少部分手指出现截断情况，考虑后续再进行优化。

图 8 部分人体检测和姿态识别结果

（2）乒乓球运动过程的检测效果

乒乓球运动过程的检测效果对比已经在表 1 中，图 9 是选取的连续帧之间的部分检测效果。可以看到模型对球的尺寸适应性比较好，尽管在距离很远的也能检测到（需要设置一个较低置信度）。

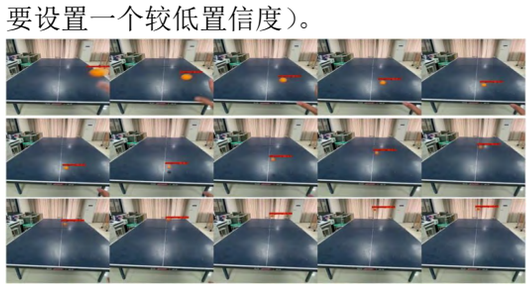

图 9 乒乓球运动过程中部分检测结果

（3）轨迹预测效果

图 10 显示了乒乓球在 x 方向的轨迹预测结果。如图，在刚开始的一小段时间，只有识别轨迹没有预测轨迹是因为已知数据少，在拟合过程中缺少可信度。可以看出，

虽然通过轨迹预测得到的数据与实际测量得到的数据存在一定误差，但是考虑到乒乓球拍直径大约 0.15m，误差在可接受范围内，证实了该算法在此部分的有效性和可行性。

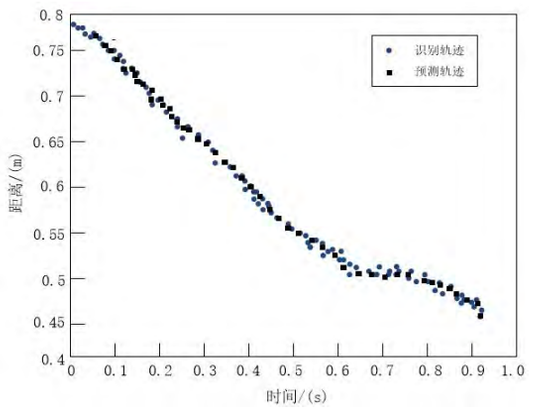

图 10　乒乓球 x 方向轨迹预测结果

6 结论

针对乒乓球击球人姿态识别及球轨迹预测，本文基于目标检测模型 YOLOv4 进行改进，用于对人体目标框检测和运动过程中的乒乓球识别。针对采集数据集特点，采用 K-means 方法对数据标签聚类。针对人体检测和乒乓球识别各自目标物的特性和速度要求，通过压缩网络卷积层并分别对分支进行结构化剪枝，提高网络实时性和识别精度，达到最优效果。球的初始方向和轨迹预测使用支持向量机回归的方式拟合函数。实验结果表明，本文提出的算法在精度和速度性能方面都有一定的提升，可以应用于乒乓球机器人视觉系统，也可以应用于其他涉及目标检测、目标跟踪和轨迹预测的球类运动场景中。

参考文献

[1] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE international conference on computer vision(ICCV), December 7-13, 2015, Santiago, Chile.New York: IEEE, 2015: 1440-1448.

[2] Ren R S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J].Transactions on Pattern Analysis & Machine Intelligence, IEEE, 2016, 39(6):1137-1149.

[3] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[M]//Leibe B, Matas J, SebeN, et al. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37

[4] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement[J], 2018..arXiv preprint arXiv:1804.02767,20.

[5] Bochkovskiy A , Wang C Y , Liao H Y M . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

[6] Fang H S, Xie S, Tai Y W, et al. Rmpe: Regional multi-person pose estimation[C] //Proceedings of the IEEE International Conference on Computer Vision. 2017: 2334-2343.

[7] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7103-7112.

[8] Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields[J]. arXiv preprint arXiv:1812.08008, 2018.

[9] Wang C, Liao H M, Yeh I, et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN.[J]. arXiv: Computer Vision and Pattern Recognition, 2019.

[10] Huang J, Zhu Z, Huang G. Multi-Stage HRNet: Multiple Stage High-Resolution Network for Human Pose Estimation[J]. arXiv preprint arXiv:1910.05901, 2019.

[11] Pavllo D, Feichtenhofer C, Grangier D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7753-7762.

[12] Jun M R, Luo H B, Wang Z B, et al. Improved yolov3 algorithm and its application in small target detection[J]. Acta Optica Sinica, 2019, 39(12):1210002.

[13] J. Song, L. Wang, L. Van Gool, O. Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4220-4229.

[14] D.C. Luvizon, D. Picard, H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.