计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(10):22−33 [doi: 10.15888/j.cnki.csa.009291]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多视角学习策略的手部姿态估计①

徐梓雄, 郭　璠, 王宗雨, 唐　琎

(中南大学 自动化学院, 长沙 410083)
通信作者: 郭　璠, E-mail: fanguo@csu.edu.cn

摘　要: 手部姿态估计在人机交互、手功能评估、虚拟现实和增强现实等应用中发挥着重要作用, 为此本文提出了一种新的手部姿态估计方法, 以解决手部区域在大多数图像中占比较小和已有单视图关键点检测算法无法应对遮挡情况的问题. 所提方法首先通过引入 Bayesian 卷积网络的语义分割模型提取手部目标区域, 在此基础上针对手部定位结果, 利用所提基于注意力机制和级联引导策略的新模型以获得较为准确的手部二维关键点检测结果.然后提出了一种利用立体视觉算法计算关键点深度信息的深度网络, 并在深度估计中提供视角自学习的功能. 该方式以三角测量为基础, 利用 RANSAC 算法对测量结果进行校准. 最后经过多任务学习和重投影训练对手部关键点的 3D 检测结果进行优化, 最终提取手部关键点的三维姿态信息. 实验结果表明: 相比于已有的一些代表性人手区域检测算法, 本文方法在人手区域检测上的平均检测精度和运算时间上有一定的改善. 此外, 从本文所提姿态估计方法与已有其他方法的平均端点误差 (*EPE_mean*) 和 *PCK* 曲线下方面积 (AUC) 这些指标的对比结果来看, 本文方法的关键点检测性能更优, 因而能获得更好的手部姿态估计结果.

关键词: 人手区域提取; 关键点检测; 多视角学习; 手部姿态估计

引用格式: 徐梓雄,郭璠,王宗雨,唐琎.基于多视角学习策略的手部姿态估计.计算机系统应用,2023,32(10):22–33. http://www.c-s-a.org.cn/1003-3254/9291.html

Hand Pose Estimation Based on Multi-view Learning Strategy

XU Zi-Xiong, GUO Fan, WANG Zong-Yu, TANG Jin

(School of Automation, Central South University, Changsha 410083, China)

Abstract: Hand pose estimation plays an important role in human-computer interaction, hand function assessment, virtual reality, and augmented reality. Therefore, a new hand pose estimation method is proposed to handle the relatively small proportion of hand region in most images and the occlusion problem of single-view keypoint detection algorithms. The proposed method first extracts the hand target region by using a semantic segmentation model which introduces the Bayesian convolutional neural networks. According to the hand localization result, the proposed method adopts a new model based on the attention mechanism and cascade guidance strategy to obtain accurate 2D hand keypoint detection results. Then, the proposed method uses a deep network based on a stereo vision algorithm to calculate the depth information of the keypoints, and the view self-learning function is provided in depth estimation. The algorithm uses triangulation as the foundation, and the RANSAC algorithm is used to correct the measurement results. Finally, the 3D hand keypoint detection results can be optimized by using multi-task learning and reprojection training, and the 3D pose of the hand keypoints can be obtained. Experimental results show that compared with some representative hand region detection algorithms, the proposed method has a significant improvement in the average detection precision and running time for hand regions. In addition, in terms of the end-point-error mean (*EPE_mean*) and the area under *PCK* curve (AUC) of different pose estimation methods, it can be seen that the keypoint detection performance of the proposed

① 基金项目: 长沙市自然科学基金 (kq2208286); 湖南省自然科学基金 (2023JJ30697)
收稿时间: 2023-03-17; 修改时间: 2023-04-20, 2023-05-23; 采用时间: 2023-06-06; csa 在线出版时间: 2023-08-22
CNKI 网络首发时间: 2023-08-23

method is better. Thus, a better hand pose estimation result can be obtained.

Key words: hand region extraction; keypoint detection; multi-view learning; hand pose estimation

手势作为人类与外界传递信息的主要方式, 它的自由性和复杂性包含了大量有用的信息. 手部姿态估计目前已成为人机交互技术中十分热门的研究方向之一, 特别是在虚拟现实和增强现实等实际技术中, 手部姿态估计技术正发挥着重要作用. 此外, 在医学领域的手功能评估方面, Swanson 算法[1] 作为一种代表性的手功能评估算法, 其主要通过对手部的三维关键点进行检测以获取手指关节角度, 进而依据基于手指关节角的手部姿态评分来对手功能进行评估. 由此可见, 研究手部姿态估计方法具有重要的理论意义与实际应用价值.

为此, 本文提出了一种新的基于多视角学习策略的 RGB 图像手部姿态估计方法. 该方法的主要创新点在于: 1) 通过引入 Bayesian 卷积网络的语义分割模型, 可在图像人手区域占比小的情况下也能有效提取到手部目标区域; 2) 提出了基于手部定位的二维关键点检测方法, 可较为准确地获取各个单视角图像下的手部二维关键点; 3) 针对单视角图像有可能存在手部姿态的遮挡问题, 本文从多摄像头的角度出发, 提出了基于多视角学习策略的三维坐标求取方法, 有效地实现了对手部关键点三维姿态信息的提取.

1 相关工作

1.1 目标检测的研究现状

目标检测方法可分为传统检测算法和基于深度学习的方法两类. 其中传统的目标检测算法主要分为 3 个步骤, 即: 区域建议、特征提取和分类回归. 传统检测方法虽然在某些场景下具有较好的效果, 但需要大量的人工设计和参数调整, 且难以适应不同的场景和目标. 因此, 这些方法已经逐渐被深度学习方法所取代.

基于深度学习的目标检测方法归纳起来可分为两阶段检测、一阶段检测、介于上述两者之间的检测方法 3 类. 两阶段目标检测方法首先生成候选框, 并计算每一个候选框内是否有目标. 这类方法涉及的两个阶段为: 第 1 阶段生成候选框, 第 2 阶段对候选框进行分类. 其中最著名的方法是 R-CNN 系列[2]. 其优点是精度较高, 缺点是速度较慢, 计算资源要求高. 为此有学者提出了一阶段目标检测方法. 该方法在不生成候选框的情况下, 直接对图片进行分类和定位. 其中最著名的是 YOLO 系列[3], 目前 YOLO 已升级到 V8 版本[4]. 这些方法具有速度快的优点, 但相应地精度有所降低. 正是由于两阶段检测方法可以获得较高的检测精度, 而一阶段检测方法具有较高的检测效率, 又有学者将两者的优势互补, 提出介于两者之间的目标检测方法. RefineDet 方法[5] 就是其中的典型代表, 该方法由锚点精调模块、目标检测模块、转换连接模块组成. 实验结果表明该方法比两阶段方法更精准, 同时保持了一阶段方法的效率. 随着目标检测技术不断变化升级, 未来必将有更多高效、准确和实用的目标检测方法出现.

1.2 姿态估计的研究现状

目前已有的姿态估计方法主要可以分为基于深度图像的方法和基于彩色 RGB 图像的方法两大类. 前者的工作主要是从由深度相机获取的深度数据中回归手姿势. 由于此类方法过于依赖所采集的深度数据, 大部分情况下深度传感器不能在较强光线下工作, 导致该方法在应用场景上有很大的局限性. 同时深度相机高昂的费用也导致该方法难以得到普及. 其典型代表性方法包括: PoseNet[6]、3D-CNN[7]、PointNet[8,9].

后者由于成本低廉、使用方便, 因而应用范围更为广泛. 例如, Zimmermann 等人[10] 提出的深度网络通过预先学习 3D 关节, 并结合图像中检测到的关键点, 实现了从单目 RGB 图像中有效估计 3D 手姿势的功能. Iqbal 等人[11] 提出了一种 2.5D 姿态表示, 以便从单目图像中估计三维手姿态. 此外, 针对多视角下的三维姿态估计问题也有一些相关研究. 例如 Kadkhodamohammadi 等人[12]提出将所有单视图中的关节二维坐标连接成单个批处理, 作为全连接网络的输入进而通过训练来预测全局三维关节坐标, 然而该方法容易出现强过拟合. Iskakov 等人[13] 提出利用三角测量方法实现多视图下的人体姿态估计, 为姿态估计方向提供了新的思路. Simon 等人[14] 提出了一种多视图引导的方法, 该方法通过多个摄像机采集的图像数据来优化不完整的检测器, 从而获取手关节等姿态估计的关键点. 但基于单视角 RGB 图像的姿态估计方法由于缺乏深度信息, 导致三维姿态估计往往不够准确, 同时单视角图像

还可能会导致关键点之间出现遮挡问题.

总体而言, 已有姿态估计方法在多任务学习方面考虑较少, 且面对图像中人手占比较小、分辨率较低等挑战也往往效果不佳.

2 系统整体实现流程

本文所提手部姿态估计方法的整体实现流程如图 1 所示. 由图 1 可知, 该流程所涉及的工作主要包括以下两个方面.

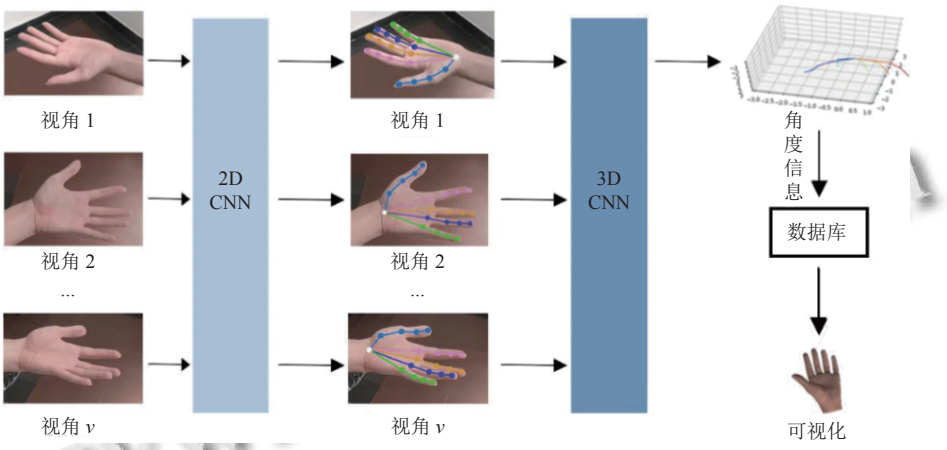


图 1 手部姿态估计方法的实现流程图

1) 基于手部定位的二维关键点检测方法. 手部定位主要是采用所提语义分割方法首先找到手部区域, 然后将定位得到的人手目标区域作为 2D 关键点检测器的输入, 以检测得到各个单视角图像下的二维关键点.

2) 基于多视角学习的三维关键点坐标求取. 由于手部姿态有可能会出现各种情况的遮挡问题, 单个视角采集的图像从根本上无法解决这样的难题. 因此, 本工作主要是基于多视角图像来求取手部关键点的三维坐标, 所涉及的研究内容包括: 特征点的描述与匹配、多任务网络学习、三角测量计算、重投影训练等. 最后, 依据所求取的三维关键点即可实现对手部姿态的有效估计.

3 手部姿态估计的关键技术

3.1 人手区域提取与二维关键点检测

手部姿态估计工作首先从人手检测和人手二维姿态估计两个方面进行考虑. 为此, 本文提出了一种基于手部定位的二维关键点检测方法. 该方法主要可以分成手部定位和二维关键点检测两个部分. 该工作一方面以 RGB 图像作为人手检测模块的输入, 经过人手分割网络定位出手的目标区域; 另一方面以人手目标区域作为 2D 姿态检测器的输入, 经过级联网络得出不同分辨率的热点图, 再利用热图聚合策略计算出较为准确的二维关键点位置. 图 2 即给出了手部定位的二维关键点检测方法流程图.

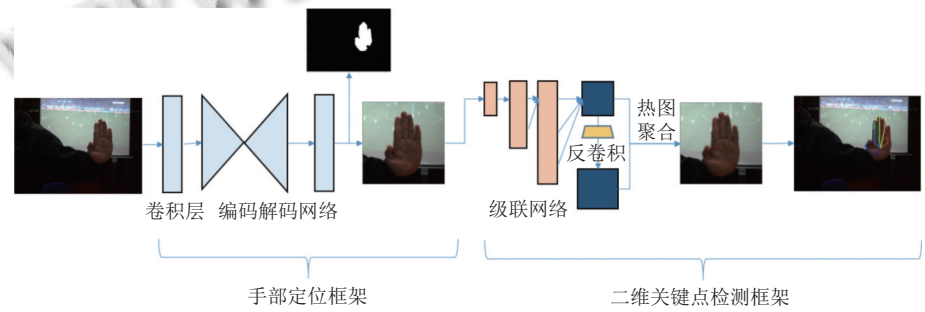


图 2 基于手部定位的二维关键点检测方法流程图

(1) 手部区域的提取

由图 2 可知, 手部定位主要通过所提分割网络实现. 由于 Bayesian CNN 通过考虑模型参数的分布来提供 CNN 模型的概率解释, 可提供较为可靠的估计模型

不确定性的方法[15]. 为此, 本文提出使用 Bayesian CNN 通过估计模型的不确定性来预测像素类标签, 从而进行手部分割, 图 3 即为所提 Bayesian IC-Net 人手分割方法流程图.

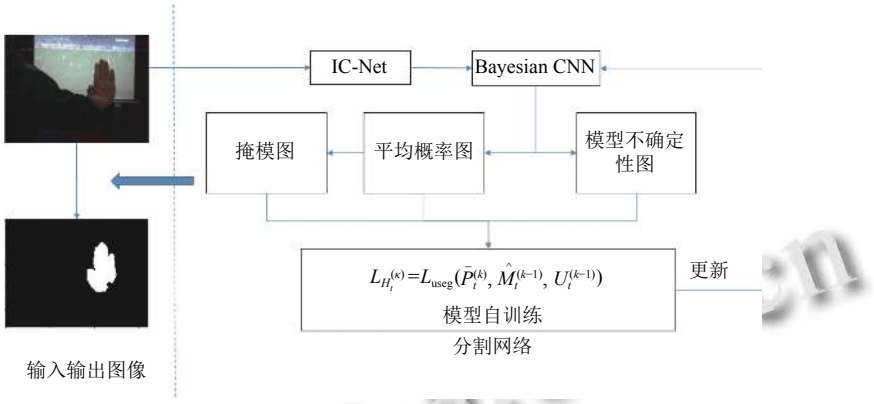


图 3 Bayesian IC-Net 人手分割方法流程图

由该图可知, 所提方法以 IC-Net 网络[16] 为基础, 在卷积层中添加正则化层, 然后对输出数据使用 Bayesian 决策进行后处理. 经过多次随机正向传递以后, 可在多个结果概率图中的每个像素点位置上进行均值处理以得到新的人手分割图, 在此基础上求取方差后记为模型的不确定性图. 该过程的关键步骤有: Bayesian 决策、模型不确定性估计、损失函数设计.

Bayesian 决策过程主要是将深度学习中的 *dropout* 操作作为 Bayesian 神经网络的近似推理并对测试时权值的后验分布进行采样. 这种近似的好处是通过 *dropout* 训练的现有 CNN 模型可以被转换为 Bayesian 模型.

在模型不确定性估计方面, 所提网络主要利用 Bayesian 模型的不确定性来构造更可靠的伪标签. 假设训练的手部分割模型为 $H(I, w)$, 该模型在给定输入彩色图像 I 的情况下输出手部概率图 P. 通过 P 和 H, 利用式 (1) 可计算得到平均概率图 $\overline{P}$ 和不确定性图 U.

$$\begin{cases} \overline{P}=\dfrac{1}{T}\displaystyle\sum_{i=1}^{T}H(I,w_i), \quad w_i\sim dropout(w) \\ U=\dfrac{1}{T}\displaystyle\sum_{i=1}^{T}P_i^2-\overline{P}^2 \end{cases} \tag{1}$$

其中, $P_i=H(I,w_i)$ 表示经过一次随机正向传递后得到的手概率图. $\overline{P}$ 和 U 与输入图像 I 具有相同的尺寸, 其中不确定性图 U 实质上等同于计算每个像素上的手概率的方差. 由此通过阈值化处理 $\overline{P}$ 后, 即可得到预测的手部分割掩模 $\hat{M}$. 此手部分割掩模图即为最终的手部区域提取图, 如图 3 的输出图像所示.

此外, 为了实现模型自适应, 所提方法将其视为一个迭代自训练的过程, 即利用前一次迭代得到的手概率图和不确定性图对当前模型进行训练. 由此, 设计了如式 (2) 所示的模型不确定性引导下的手部分割损失 $L_{\text{uma}}(P,\hat{M},U)$. 当平均不确定性得分的减少小于 10% 时, 终止迭代, 最后得到人手区域提取图.

$$L_{\text{uma}}(P,\hat{M},U)=-\frac{1}{M}\sum_{m=1}^{M}(1-U_m) \cdot\left(\hat{M}_m\log P_m+\left(1-\hat{M}_m\right)\log(1-P_m)\right) \tag{2}$$

(2) 基于手部定位的二维关键点检测

手部关键点检测旨在从一个尺寸为 $W\times H\times 3$ 的图像中检测出手指关节或者部位 (如指尖) 的位置, 为此可将该问题转化成生成 P 个大小为 $W\times H$ 的热点图, 其中热点图 k 表示的是第 p ($p\in\{1,\cdots,P\}$) 个关键点的位置置信度. 热点图的分辨率对目标对象的预测效果很重要. 现有方法大多通过串行连接高分辨率和低分辨率特征图, 在热图预测中使用输入图像的 1/4 分辨率. 这种方法在目标对象体积较小的情况下高斯核会在关键点的定位中混淆, 导致准确度降低. 为此, 本文以级联网络模型为基础提出了一种基于注意力机制和级联引导策略的新模型, 该模型网络结构如图 4 所示.

由图 4 可知, 彩色输入图像经过一个卷积网络将分辨率降低到 1/4, 该卷积网络是由两个大小为 3×3 的跨步卷积构成. 所提方法以卷积网络输出的特征图作

为输入, 经过一个高分辨率和低分辨率并行的网络作为级联网络. 此级联网络一直保持着图像高分辨率特征并且高分辨率和低分辨率特征图之间会重复交换信息, 由此可利用低分辨率提高高分辨率特征图的表示. 同时, 将不同分辨率特征图经过通道拼接后输入至反卷积模块并进行热图聚合, 将输出的热图称为最终预测结果. 此外, 在通道拼接之前该方法还添加了坐标注意力机制, 从而有效地改善了网络模型性能. 在此基础上, 可计算级联网络中各级输出热点图产生的级联引导损失来获得较好的二维关键点检测结果.

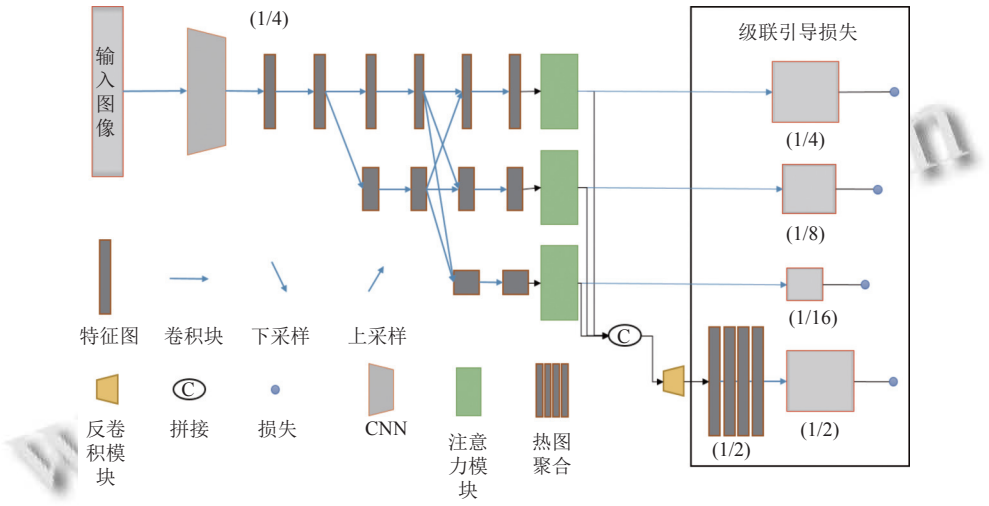


图 4 基于注意力机制的级联网络模型

具体来说, 所提方法先将不同分辨率的热图进行图像拼接, 然后使用反卷积模块将拼接后的特征图作为输入. 在反卷积模块中, 先使用尺寸为 4×4 的反卷积网络, 然后使用批归一化和线性修正单元 ReLU 不断学习输入图的特征同时进行上采样. 对于采样后的特征图, 使用 3 个基本残差块来进行细化, 由此可得到更为精确的关键点热图.

对于最终得到的预测热图, 本文提出了一种聚合多分辨率热图的策略. 在生成多种分辨率 (1/16、1/8、1/4、1/2 等) 热图后, 采用双线性插值方法以这些多尺度图像作为输入, 向上采样恢复到原始输入图像的大小, 并将所有尺度的热图平均处理, 由此得到的输出作为最终的预测. 同时通过设计级联引导损失来合理地学习每个分辨率的热图特征, 从而利用不同的尺度来引导低、中以及高分辨率热图的特征学习过程.

上述过程的具体操作为: 给定 T 个分支和 N 个类别. 在分支 t 中, 预测的特征图F^t空间大小为$Y_t \times X_t$. 位置(n,x,y)处的值为$F^t_{n,x,y}$. 2D 位置(x,y)对应的真值标签为$\hat{n}$. 为了训练级联网络模型, 所提方法在每个分支中附加加权最大交叉熵损失和相关损失权重λ_t, 由此最小化损失函数可定义为:

$$L_{\mathrm{CLG}} = -\sum_{t=1}^{T}\lambda_t \frac{1}{X_t Y_t}\sum_{x=1}^{X_t}\sum_{y=1}^{Y_t}\log\frac{\mathrm{e}^{F^t_{\hat{n},x,y}}}{\sum_{n=1}^{N}\mathrm{e}^{F^t_{n,x,y}}} \tag{3}$$

其中, 每个热图损失的权重λ_t均设为 1. 相比于现有方法只使用高分辨率提取的特征, 所提方法综合热图聚合方法和级联引导损失设计的级联标签引导策略, 使得网络模型的训练过程效果更优. 同时由于引入了多分辨率特征图, 这样就使最终预测的结果不会受某个单一热图所左右, 由此即可获得较为准确的手部二维关键点检测结果.

3.2 基于多视角学习策略的三维坐标求取

为了重建手部 3D 姿态, 所提方法将采集到的所有图像进行人手区域检测和二维姿态检测的逐帧处理. 在此基础上, 本文提出了一种基于立体视觉算法计算关键点深度信息的深度网络, 并且在深度估计中附加视角自学习的功能. 该立体视觉算法以三角测量为基础, 利用 RANSAC 算法[17] 对三角测量结果进行校准, 最后经过多任务学习和重投影训练对结果进行优化. 图 5 即给出了基于多视角学习策略的三维关键点检测方法的主要步骤.

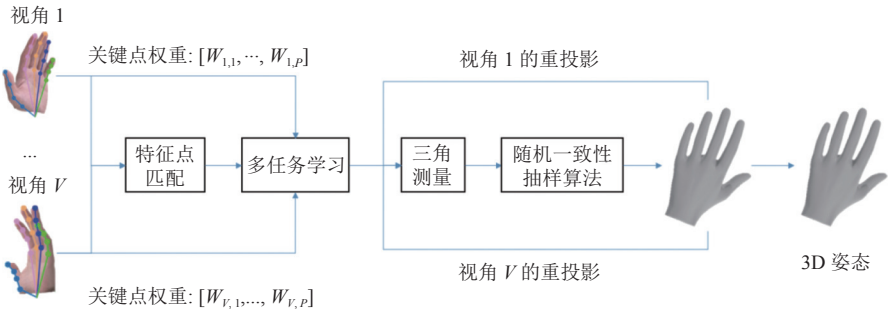


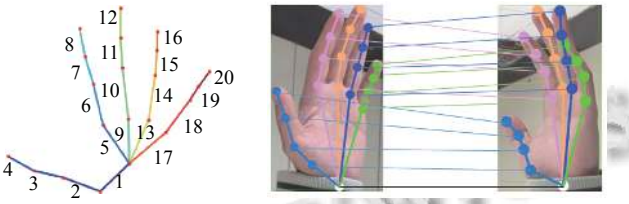
图 5 基于多视角学习策略的三维关键点检测方法示意图

由图 5 可知, 所提基于多视角学习策略的三维关键点检测方法的关键步骤主要包括特征点匹配、多任务学习、三角测量计算、重投影训练等. 各关键步骤的主要工作如下.

1) 特征点匹配: 该项工作主要是匹配不同视图之间相同的关键点, 以解决多角度视图带来的数据多义性问题.

将由本文方法所得到的人手感兴趣域输入 2D 关键点检测器可计算得到 P 个 2D 关键点坐标如式 (4) 所示. 该式中$x_p \in R^2$, 表示关键点 p 的二维坐标. w_p表示关键点带有的相关检测置信度, P 为当前视角检测出的总关键点个数. 每个关键点 p 在手部图像都指向如图 6(a) 所示的不同关节位置.

$$d_0(I) \to \{(x_p, w_p) \text{ for } p \in [1,\cdots,P]\} \tag{4}$$



(a) 人手关键检测点示意图 (b) 关键点匹配示意图

图 6 人手关键点及其匹配示意图

如图 6(b) 所示, 本文使用快速近似最近邻搜索 FLANN 算法[18] 一次对两个视角图像进行匹配. FLANN 算法在匹配之前为特征关键点建立索引树, 并在匹配过程中使用最邻近搜索算法进行优化. 此外, 还选用 KNN-matching 算法[19] 通过设置恰当的阈值来进一步消除匹配误差. 从最终的基于多视角学习的三维关键点检测结果来看, 该特征点匹配精度符合应用需求.

2) 多任务学习: 本文采用热图方法从姿态估计器中获取每个关键点的二维坐标信息. 在此热图中, 每个关键点区域都可以计算出其置信度大小. 所提方法以关键点真实位置为中心绘制带有置信度的关键点热图, 不同颜色深度表示不同的置信度大小. 图 7 即为带有关节置信度的可视化热图.

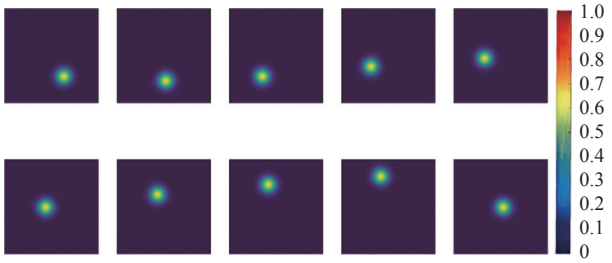


图 7 带有关节置信度的可视化热图

由此可将多视图人手看作是多任务网络体系, 借助每个任务损失率的变化对该任务中的关键点进行权重自学习, 避免某些图像数据在计算中占主导地位. 同时, 依据不同视角拍摄的同一关键点具有一致性, 本文提出了一种多视图一致性损失 L_{MC} 来优化网络. L_{MC} 定义为刚性对准后不同视图之间三维关节位置差的加权和.

$$L_{\text{MC}} = \sum_{\substack{v,v' \in V \\ v \neq v'}} \sum_{p \in P} w_{v,p} w_{v',p} \cdot d(\hat{p}_{v,p}, R_v^{v'} \hat{p}_{v',p}) \tag{5}$$

其中, $w_{v,p}$、$w_{v',p}$分别为视角I_v和$I_{v'}$下关键点 p 的置信度; $\hat{p}_{v,p}$、$\hat{p}_{v',p}$分别为视角I_v和$I_{v'}$下关键点 p 的尺度归一化三维坐标. $R_v^{v'} \in R^{3\times4}$是两个视图之间由旋转矩阵 R 和平移向量 t 构成的变换矩阵, 用于对齐两个 3D 位姿. d 是对齐位姿之差的距离度量, 如式 (6) 所示.

$$d(\hat{p}_{v,p}, R_v^{v'} \hat{p}_{v',p}) = d\left(z_{v,p} K_v^{-1} \begin{bmatrix} x_{v,p} \\ y_{v,p} \\ 1 \end{bmatrix}, R_v^{v'} z_{v',p} K_{v'}^{-1} \begin{bmatrix} x_{v',p} \\ y_{v',p} \\ 1 \end{bmatrix}\right) \tag{6}$$

其中, $(x_{v,p},y_{v,p},1)^{\mathrm{T}}$和$(x_{v',p},y_{v',p},1)^{\mathrm{T}}$是通过 2D 姿态检测器得到的视角$I_v$和$I_{v'}$下关键点 p 的二维坐标. K_v和$K_{v'}$是视角. I_v和$I_{v'}$下相机的内参, 同时使用 L_1 范数作为距离度量 d.

3) 三角测量计算: 该工作主要从特征点匹配和三角测量结果两个方面进行优化. 一方面通过 RANSAC 算法抽样出匹配效果较好的点来训练模型, 由此即可优化匹配算法和消除置信度较小的图像所带来的误差. 另一方面, 利用三角测量法初步得到的结果在标签数据较差的视图进行投影, 并与该标签标注图像进行比较分析, 以得到一个鲁棒的三维估计结果.

三角测量是利用不同视角的二维平面信息, 进行特征匹配, 结合相机参数估计出相机之间相对位置姿态的一种方法[20]. 当使用两个相机在不同位置同时拍摄时, 同时出现在不同视角的成像必定出现一些差异, 而三角测量正是利用相机的参数和这些差异对相机位姿关系进行估计. 利用 RANSAC 算法最终得到的内点集 S_{in} 对三角测量的结果进行优化, 从而得到一个鲁棒的位置结果:

$$X_p^f = \arg\min_X \sum_{v\in S_p^j} \left\|\rho_v(X) - x_p^v\right\|_2^2 \tag{7}$$

其中, S_p^j为内点集, $\rho_v(X)\in R^2$是 3D 点 X 在视角 v 的投影点, X_p^v为视角 v 中关键点 p 的 2D 位置, $X_p^f\in R^3$是第 f 帧图像中关键点 p 的三维位置.

4) 重投影训练: 姿态估计器生成的 2D 标注图需要校正, 为此将三角测量得到的 3D 坐标投影到各个视角的平面上, 以实现一个多视角自学习的 3D 坐标检测方法. 具体来说, 本文所提重投影训练方法是在已有的关键点检测器 d_0 基础上, 添加由外在信息监督优化得到新检测器 d_1, 并确保 d_1 包含 d_0 中不存在的信息. 定义 T_0 为初始训练集, 首先利用 T_0 训练初始检测器 d_0 得到各个视角关键点位置, 然后进行三角测量估计其深度信息. 该过程可表示如下, 其中 F 是图像 I 的帧数个数, X_p^f是关键点标签注释.

$$T_0 := \left\{\left(I^f, \left\{X_p^f\right\}\right) \text{ for } f\in[1,\cdots,F]\right\} \tag{8}$$

重投影训练是将 3D 关键点投影到每个视图中, 生成新的标注图像 T_1. T_1 可以为检测失败的视图提供新的二维注释. 因此, 可定义优化后的检测器 d_1 为:

$$d_1 \leftarrow \text{Reproject}(T_0 \text{ and } T_1) \tag{9}$$

以下算法伪代码描述了本文所提基于多视角学习的三维检测过程. 其中$d_0(I)\rightarrow\{(x_p, w_p) \text{ for } p\in[1,\cdots,P]\}$表示初始关键点检测器$d_0$, 其主要用来检测二维关键点位置并生成位置权重信息. $I_v^f\rightarrow\{v\in[1,\cdots,V], f\in[1,\cdots,F]\}$表示未标记的多个视角多帧图像的集合. T_0 为初始训练集.

图 8 即为利用上述方法所完成的一组握拳实验及其可视化结果, 其中每行代表一种手部握拳姿态, 前 5 列分别对应了运动过程的 5 个不同视角图像, 最后一列为手部关键点 3D 检测结果图.

算法 1. 多视角三位检测过程

输入: 多视角图像 I_v; 关键点检测器 d_0; (3) 标注的训练数据 T_0
输出: 改进的检测器 d_1 和识别出的三维关键点

For v from 0 to V
1. 输入图像对关键点检测结果进行权重的自学习再使用三角测量计算.
For 每一帧图像:
  a) 对所有视角 进行二维关键点检测.
  b) 对关键点权重 进行自学习.
  c) 利用 RANSAC 算法对关键点进行鲁棒的三角测量.
  d) 利用多视图一致性损失 优化 3D 结果.
2. 利用重投影训练的方法优化姿态检测器.

4 实验结果与分析

4.1 人手区域提取实验结果分析

本方法所采用的研究数据集为 RHD 数据集[10] 和 UTG 数据集[21]. 其中, RHD 数据集包含 41 258 个训练样本和 2 728 个测试样本, 每个样本均提供像素为 320×320 的 RGB 图像、深度图像, 以及每只手的 21 个关键点标注. 此关键点标注图包括有图像坐标系的 u-v 坐标和世界坐标系的 x-y-z 坐标. 而 UTG 数据库则刻画了 5 名试验者进行 17 种不同类型的握拳动作, 适合用于检验手部姿态估计. 下面主要针对所提算法的 3 个关键步骤进行分析.

为了验证所提基于语义分割的人手区域检测方法的有效性, 在上述 RHD 和 UTG 数据集上使用平均检测精度指标对所提方法与 YOLOv4[4,22], YOLOv8[4] 和 RefineDet[5] 等已有代表性目标检测方法进行了比较. 其中 YOLOv4 和 YOLOv8 是 one-stage 的设计思想, 而 RefineDet 是两步回归的目标检测方法, 实验结果如表 1 所示.

本文所提方法虽然在运行时间上比 YOLO 系列方法稍长, 但是在人手区域的平均检测精度上有小幅提

升. 此外, 相比于 RefineDet 方法, 本文方法在运行时间少于 RefineDet 方法的前提下, 平均检测精度仍有 1% 左右的提升. 由此可见, 本文方法能在检测精度与运行时间方面取得相对较好的均衡.

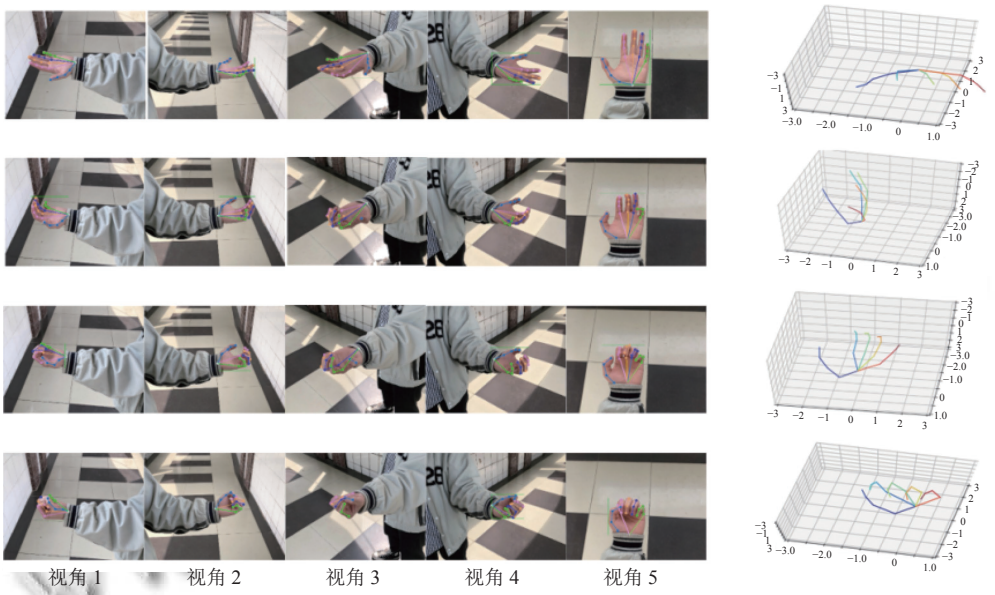


图 8 多视角三维关键点检测实验结果图

表 1 所提人手检测算法与现有方法的平均检测精度对比

方法	RHD (%)	UTG (%)	运行时间 (s)
YOLOv4[4,22]	87.43	88.31	0.018
YOLOv8[4]	89.92	91.04	0.012
RefineDet[5]	89.67	90.96	0.036
本文方法	91.22	91.74	0.023

图 9 即为本文方法与 YOLOv4、RefineDet 模型以及 YOLOv8 模型的手部区域检测效果对比示例. 由该图可知, 已有模型对于有遮挡且面积较小的手部区域往往不能较好检出. 而本文方法则不存在上述问题, 究其原因主要是因为手部小目标所占像素少, 输入图像经过卷积后用于目标检测和分类的特征就少了, 因此目标被漏检的可能性较大. 另一方面, 从模型本身来说, 随着网络层数的增加和感受野的变大, 微观的信息将会丢失, 由此造成聚合后的特征减少, 影响检测的最终效果. 而本文方法通过在 IC-Net 语义分割网络中加入 Bayesian 决策来估计模型不确定性, 同时采用一个迭代自训练的过程实现了模型的自适应, 实验结果表明该方法能获得相对较优的人手检测效果.

4.2 人手关键点检测实验结果分析

在提取人手感兴趣域后, 即可利用热点图检测人手关键点. 如图 10 即为所提关键点检测方法对上述两个数据集图像的关键点检测结果示例. 其中, 左图为原图, 右图为人手关键点检测结果.

图 9 人手区域检测算法效果示例

图 10 二维关键点检测结果示例

实验采用端点误差 (end-point-error, *EPE*)[23] 对所提方法和已有二维姿态估计网络 CPM (convolutional pose machines) 方法[24] 和 HR-Net (high-resolution net) 方法进行了性能对比. 这里的 *EPE* 指标是衡量关键点检测错误率的指标, 表示所有关键点的标签点和预测点之间的欧氏距离 (mm). 实验中采用了平均端点误差 (*EPE_mean*) 和 *EPE* 中位数 (*EPE_median*) 两个指标.

EPE_mean 的计算方式如式 (10) 所示, 其中 P 表示关键点的个数, 是 ground truth 坐标, 是预测点坐标. *EPE_mean* 值越低表示预测越准.

$$EPE\_mean = \frac{1}{P}\sum_{p=1}^{P}\sqrt{(x-x_p)^2+(y-y_p)^2} \quad (10)$$

另一方面从网络结构来说, 本文方法采用了级联网络 CN (cascade net)、卷积块注意力机制 CBAM (convolutional block attention module)、坐标注意力机制 CA (coordinate attention) 和级联标签引导策略 CLG (cascade label guidance) 来提高级联网络的关键点检测性能. 本文对上述各模块的作用也进行了测试, 表 2 即给出了二维关键点检测方法的指标值对比结果. 由表 2 可知, 相比于经典卷积姿态网络模型 CPM, 级联网络 CN 在特征提取方面保持着高分辨率特征, 可使预测结果更精准. 而在级联网络末端添加的坐标注意力机制模块, 相比于卷积注意力机制模块, 因其保留了通道的位置信息, 可使网络结构具有更好的鲁棒性. 同时级联标签引导策略通过最小化级联引导损失, 也可使最终的预测热图更为准确.

表 2 二维关键点检测方法的指标值对比

方法	RHD		UTG	
	EPE_median	*EPE_mean*	*EPE_median*	*EPE_mean*
CPM[24]	6.34	7.21	7.491	9.386
HR-Net[25]	5.92	7.63	6.936	8.779
CN+CBAM	7.91	8.67	9.536	10.256
CN+CA	6.87	7.65	9.469	10.237
CN+CA+CLG	5.84	6.45	7.164	8.284

本文还对 Google 最新推出的机器学习应用框架 MediaPipe[26] 的手部关键点检测功能进行了测试. 实验结果表明: MediaPipe 在逆光环境下或处理手部阴影时容易将手指的阴影误检为人手, 导致手指阴影部分也被检测出关键点. 本文所提方法由于首先通过语义分割方式定位出手部区域, 因而可较好排除人手阴影的干扰, 为后续的关键点检测奠定了良好基础.

此外, 实验中我们发现若仅考虑单视角图像无法较好解决某些手部姿态遮挡问题. 图 11 即给出了两组单视角图像下二维关键点检测的失败示例. 其中, 左图为原图, 右图为人手关键点检测结果. 由图 11 可知, 因为手部其他手指或外物的遮挡, 导致被遮挡手指的关键点检测错误. 例如图 11(a) 把被遮挡的握拳姿势识别成手指半伸姿势, 导致相应的手指关键点检测错误. 而图 11(b) 大拇指由于受外物遮挡其关键点检测错误. 正是因为仅利用一幅单视角图像进行手部姿态估计存在诸多问题, 为了更好地应对手指或手掌间的遮挡等挑战, 本文提出利用多幅单视角图像, 采用多视角学习策略来求取手部关键点三维坐标, 在此基础上即可实现对手部姿态的有效估计.

(a) 示例 1 (b) 示例 2

图 11 二维关键点检测的失败示例

4.3 三维关键点检测实验结果分析

三维关键点估计实验采用 Human3.6M 数据集[27] 和 CPS 数据集[28]. Human3.6M 数据集包含 4 个同步数码相机拍摄得到的 360 万帧图像, 其中每张图像都包括 3D 姿态标注. 该数据集有 11 名人类受试者 (5 名女性和 6 名男性), 分为训练、验证和测试集. CPS 是一个由卡内基梅隆大学维护的多摄像机数据集, 该数据集提供了由 31 个同步摄像机采集的共 14 816 幅标注了全身的关键点的人体和人手图像.

(1) 不同模块组合实验

本文对多视角学习策略的不同组成模块进行了对比分析, 分析结果如表 3 所示. 表 3 中 2D 代表二维关键点姿态检测器, MTN (multi-task network) 表示使用多任务学习网络进行视图权重的学习, TL (triangulation) 表示使用三角测量的方法进行点深度计算, TLO (triangulation optimization) 表示优化后的三角测量方法, RC (reproject correction) 表示在三角测量后利用重投影校正二维姿态估计图. 由表 3 可知, 对于三维姿态估计, 仅依靠传统的三角测量方法具有很大的误差. 为此, 本文所提方法对三角测量中特征匹配、多视图置信度以及二维姿态估计这几个部分均实现了优化, 在 Human3.6M 和 CPS 数据集上的实验结果表明采用所提方法可使关键点检测精度得到很大的提升.

表 3 多视角学习网络三维姿态估计实验评估表

方法	Human3.6M		CPS	
	EPE_mean	AUC	*EPE_mean*	AUC
2D+TL	42.295	0.617	49.461	0.619
2D+TLO	37.909	0.625	42.874	0.663
2D+MTN+TLO	38.123	0.732	36.259	0.773
2D+MTN+TLO+RC	27.036	0.823	24.164	0.843

(2) 与主流三维人手姿态估计方法的实验对比

为了检验所提方法的有效性, 本文还在 Human3.6M 和 CPS 数据集上与目前主流的三维人手姿态估计方法进行了对比, 所选取的已有方法包括 Zimmermann 等人提出的单视图方法 Pose-Prior[10], Iqbal 等人提出的 2.5D 热图回归方法 2.5D Heatmap[11], 以及 Simon 等人提出的多视角方法 Bootstrapping[14]. 表 4 即给出了所提多视图方法与已有方法的 *EPE_mean* 指标对比结果. 由于 *EPE_mean* 指标值越小表示算法性能越好, 因此从表 4 可以看出, 整体而言本文方法能取得比已有方法更优的三维关键点检测结果.

表 4 本文方法与现有人手姿态估计方法的对比结果

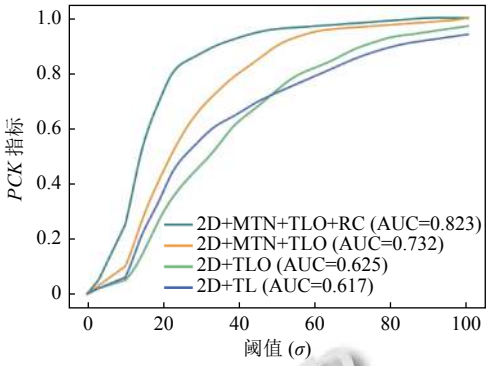
方法	Human3.6M		CPS	
	EPE_mean	AUC	*EPE_mean*	AUC
Pose-Prior[10]	37.295	0.781	35.159	0.859
2.5D Heatmap[11]	33.869	0.806	31.237	0.940
Bootstrapping[14]	28.823	0.862	25.591	0.972
本文方法	27.036	0.893	24.164	0.990

此外, 为了对进一步验证多视角自学习策略的有效性, 本文还绘制了 *PCK* (percentage of correct key-points) 曲线以衡量相关关键点的平均性能. *PCK* 指标计算被正确估计出的关键点百分比, 其计算公式为:

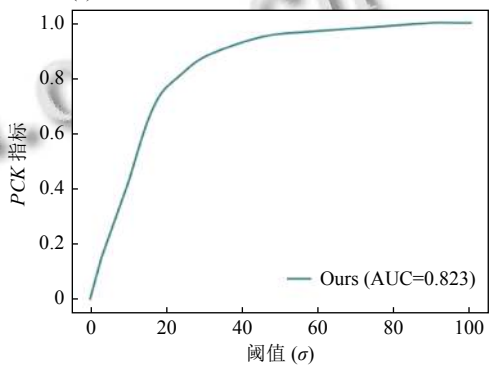
$$PCK_p^{\sigma}(d_0)=\frac{1}{|T|}\sum_{T}\delta\left(\|x_p^f-y_p^f\|_2\leqslant\sigma\right) \tag{11}$$

其中, $x_p^f\in d_0(I^f)$表示第 p 个关键点的三维坐标, y_p^f表示关键点的真实坐标, σ是人工设定与真实距离之间的阈值, 单位为 mm. T 表示评估指标时使用的数据集, $\delta(\cdot)$是指示函数, PCK_p^{σ}表示我们求出的第 p 个关键点在阈值σ上的 *PCK* 百分比. *PCK* 曲线上每个点都表示检测器在测试数据集上的表现, 为了客观的定量分析这种表现上的差异, 我们计算了 *PCK* 曲线下方的面积 (AUC) 的数值, 此 AUC 指标值越大表示检测性能越好.

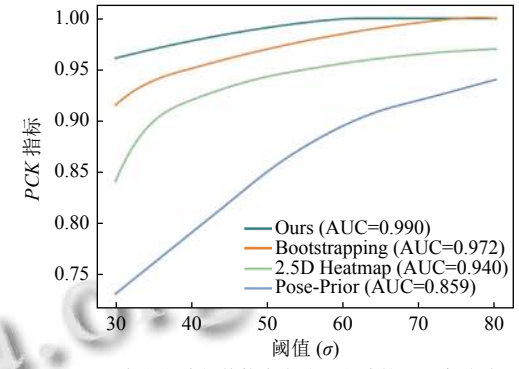
如图 12 所示, 本文利用数据集中提供的内外参数将关键点坐标转换到同一坐标系中, 以自学习策略所使用的组成模块来命名各个模型, 通过调整 *PCK* 指标中的相关阈值来绘制 *PCK* 指标曲线图. 其中, 图 12(a) 为所提多视角学习策略在数据集 Human3.6M 下的实验结果, 图 12(b) 为所提方法在 CPS 数据集下的测试结果, 图 12(c) 为将所提方法与其他三维姿态估计方法在 CPS 数据集上的比较结果.



(a) 本文方法在 Human3.6M 数据集上的实验结果



(b) 本文方法在 CPS 数据集上的实验结果



(c) 本文方法与其他姿态估计方法的对比实验结果

图 12 *PCK* 指标统计结果图

由图 12 所示的 *PCK* 指标统计结果可知, 传统三角测量方法仍需进一步改善. 同时, 单视图姿态回归的方法往往会在手部运动过程中丢失关键点的重要信息. 因此, 本文所提方法利用二维关键点检测得到多个摄像头视角关键点的位置和权重信息, 由此设计和训练了一个多视图一致性的学习网络, 以减小检测效果较差的视角图像对三维估计结果带来的干扰. 此外, 所提方法还利用 RANSAC 算法优化特征匹配过程以及三角测量估计结果. 在得到鲁棒的三维坐标后, 提出的重投影训练方法也可以对三维估计网络的初始训练集进行一定程度的优化. 由图 12 所示的统计结果可知, 上

述这些所提模块可明显提升检测结果的准确度.

5 结论

本文所提方法结合立体视觉匹配的方法, 采用多视角学习策略来获取手部三维关键点信息. 与单一视角的手部姿态估计方法相比, 其主要优势在于基于多视角学习策略的手部姿态估计方法可以更好地解决手部区域在图像中占比小, 以及已有单视图关键点检测算法无法应对的遮挡等问题. 为此, 所提算法主要围绕手部区域提取、手部关键点的二维估计及三维估计展开研究, 并且在多个公开数据集上验证了所提方法相比于其他人手姿态估计方法的优势, 从而为人机交互、手功能评估、虚拟现实与增强现实等诸多领域的应用奠定了良好基础. 将来的工作主要包括选取最为合适的相机位置和个数, 以期构建更为完整的多视角姿态数据集. 在此基础上, 实现了一种端到端的多摄像头手部姿态估计网络, 同时进一步提升所提手部姿态估计方法的实时性, 以期将该姿态估计方法在多个领域进行应用.

参考文献

1 Swanson AB, Hagert CG, deGroot Swanson G. Evaluation of impairment of hand function. The Journal of Hand Surgery, 1983, 8(5): 709–722. [doi: 10.1016/S0363-5023(83)80253-6]

2 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

3 Shao YH, Zhang D, Chu HY, *et al*. A review of YOLO object detection based on deep learning. Journal of Electronics & Information Technology, 2023, 44(10): 3697–3708.

4 Terven J, Cordova-Esparza D. A comprehensive review of YOLO: From YOLOv1 and beyond. arXiv:2304.00501, 2023.

5 Zhang SF, Wen LY, Lei Z, *et al*. RefineDet++: Single-shot refinement neural network for object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(2): 674–687. [doi: 10.1109/TCSVT.2020.2986402]

6 Chang JY, Moon G, Lee KM. V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5079–5088.

7 Ge LH, Liang H, Yuan JS, *et al*. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5679–5688.

8 Qi CR, Su H, Mo K, *et al*. PointNet: Deep learning on point sets for 3D classification and segmentation. Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 77–85.

9 Qi CR, Yi L, Su H, *et al*. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 5105–5114.

10 Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 4913–4921.

11 Iqbal U, Molchanov P, Breuel T, *et al*. Hand pose estimation via latent 2.5D heatmap regression. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 125–143.

12 Kadkhodamohammadi A, Padoy N. A generalizable approach for multi-view 3D human pose regression. Machine Vision and Applications, 2021, 32(1): 6. [doi: 10.1007/s00138-020-01120-2]

13 Iskakov K, Burkov E, Lempitsky V, *et al*. Learnable triangulation of human pose. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7717–7726.

14 Simon T, Joo H, Matthews I, *et al*. Hand keypoint detection in single images using multiview bootstrapping. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1145–1153.

15 Gal Y, Ghahramani Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv:1506.02158, 2016.

16 Zhao HS, Qi XJ, Shen XY, *et al*. ICNet for real-time semantic segmentation on high-resolution images. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 418–434.

17 Fischler MA, Bolles RC. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981, 24(6): 381–395. [doi: 10.1145/358669.358692]

18 Muja M, Lowe DG. Fast approximate nearest neighbors with automatic algorithm configuration. Proceedings of the 4th International Conference on Computer Vision Theory and Applications. Lisbon: VISAPP, 2009. 331–340.

19 Xu DW, Wang YD, Jia LM, *et al*. Real-time road traffic states estimation based on kernel-KNN matching of road traffic spatial characteristics. Journal of Central South University, 2016, 23(9): 2453–2464. [doi: 10.1007/s11771-016-3304-9]

20 Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge: Cambridge University Press, 2000.

21 Cai MJ, Kitani KM, Sato Y. An ego-vision system for hand grasp analysis. IEEE Transactions on Human-machine Systems, 2017, 47(4): 524–535. [doi: 10.1109/THMS.2017.2681423]

22 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.

23 Yao Y, Luo ZX, Li SW. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1787–1796.

24 Wei SE, Ramakrishna V, Kanade T, *et al*. Convolutional pose machines. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4724–4732.

25 Wang JD, Sun K, Cheng TH, *et al*. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349–3364. [doi: 10.1109/TPAMI.2020.2983686]

26 Zhang F, Bazarevsky V, Vakunov A, *et al*. MediaPipe hands: On-device real-time hand tracking. arXiv:2006.10214, 2020.

27 Ionescu C, Papava D, Olaru V, *et al*. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325–1339.

28 Joo H, Simon T, Li XL, *et al*. Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(1): 190–204. [doi: 10.1109/TPAMI.2017.2782743]

(校对责编: 牛欣悦)