

文章编号:1007-2780(2022)06-0736-10

基于级联特征和图卷积的三维手部姿态估计算法

林依林^{1,2}, 林珊玲^{2,3}, 林志贤^{1,2,3*}

(1. 福州大学 物理与信息工程学院, 福建 福州 350116;

2. 中国福建光电信息科学与技术创新实验室, 福建 福州 350116;

3. 福州大学 先进制造学院, 福建 泉州 362200)

摘要:针对手部的高自由度问题和结构相似问题引起的三维关键点姿态估计误差,本文提出了一套联合识别、检测以及姿态估计的三维手部骨架姿态回归网络。采用基于YOLOv3的预处理网络,提出基于级联多特征热度图的二维和三维关键点检测网络,并在特征提取网络架构中引入人体骨架手部约束,利用渐进的图卷积神经网络特征增强模块对骨架关键点结果进行进一步精细化修正,完成姿态由粗到细的调整。本文与现有多种算法在不同公共数据集下进行PCK指标和AUC指标比较,本文算法在不同测试集上的AUC指标均达到最高,平均AUC精度达到92.9%。实验表明本文方法可以通过单张二维数据准确、细致地估计三维手部姿态,并且在测试集与自然场景下均有较好表现。

关键词:三维姿态估计;目标检测;手势识别;特征增强;卷积神经网络;图卷积神经网络

中图分类号:TP391.4 **文献标识码:**A **doi:**10.37188/CJLCD.2021-0307

3D hand pose estimation algorithm based on cascaded features and graph convolution

LIN Yi-lin^{1,2}, LIN Shan-ling^{2,3}, LIN Zhi-xian^{1,2,3*}

(1. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China;

2. Fujian Science & Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou 350116, China;

3. School of Advanced Manufacturing, Fuzhou University, Quanzhou 362200, China)

Abstract: For the 3D key point pose estimation error caused by the high degree of freedom problem and structural similarity problem of the hand, this paper proposes a novel 3D hand skeleton pose regression framework for joint identification, detection, and pose estimation. The framework firstly adopts a YOLOv3-based detector to obtain the position of hands, then a cascade pose estimation network is designed to get initial hand poses with 2D and 3D pose supervisions. Finally, considering the natural constrains in hand graph connection, we present progressive GCN module to further refine the initial hand pose from coarse to fine. This paper compares PCK metrics and AUC metrics with the state-of-the-art approaches under

收稿日期:2021-11-29;修订日期:2021-12-27.

基金项目:国家重点研发计划(No. 2021YFB3600603);福建省自然科学基金(No. 2020J01468)

Supported by National Key R&D Program of China (No. 2021YFB3600603); Natural Science Foundation of Fujian Province(No. 2020J01468)

*通信联系人, E-mail: lzx2005000@163.com

different public benchmarks, and the proposed method achieves the highest AUC metrics on different test sets, with an average AUC accuracy of 92.9%. The experiments illustrate that the proposed method is able to effectively and robustly predict 3D hand pose from monocular image, performing well in both test sets and in the wild.

Key words: 3D pose estimation; target detection; gesture recognition; feature enhancement; convolutional neural network; graph convolutional neural network

1 引言

姿态估计是计算机视觉中的热门研究领域,是对人体姿态的位置估计。姿态估计一般可以分为单人姿态估计(如 Open Pose^[1])、多人姿态估计(如 AlphaPose^[2])、人体姿态跟踪、三维人体姿态估计。在姿态估计的研究中,基于手部的姿态估计研究备受青睐。在人所有的姿态中,手势占据了90%,是最主要的人机交互姿态。未来的生活场景朝着越来越智能化的方向发展,智能家居、自动驾驶、智慧医疗及第一视角沉浸式交互等应用场景,都离不开手势交互的身影。

近年来,随着深度学习的发展,针对手部的三维姿态估计研究突飞猛进。Cai等人^[3]提出了一种弱监督网络,使用深度正则化器,将从彩色图像估计的三维手势转换成深度图,将三维坐标估计损失转化为深度图损失,有效地解决了三维关键点标记获取困难这一问题。Ge等人^[4]将手部表面网格估计加入到网络中,将彩色图像估计的二维手部的热度图通过图卷积网络估计手表面网格,再通过手表面网格回归三维手势。该方法识别精度较高,但是手表面网格真实标记缺乏,制作合成数据也较为困难,数据获取代价较大。对于RGB图像的三维手部姿态估计任务,手部独有的严重的自遮挡性和自相似性以及复杂的背景处理,在缺少深度信息的任务里并不容易。无约束的自然场景往往包含复杂的背景和多变的光照条件,要准确地从第一视角RGB图像中检测出指尖的位置依然是一个具有挑战性的问题。本文在 Minimal-Hand^[5]的基础上结合图卷积神经网络^[6]来解决这种天然的遮挡问题,通过级联的卷积网络从粗到细优化关键点位置从而解决不自然的骨架估计。

2 网络结构

2.1 整体网络框架

在姿态估计中,常见的手部模型有21关键点、16关键点、36关键点。本文采用 Open Pose^[1]提出的标准手部21关键点模型。其中编号为0的关键点是手腕,其余每根手指分别有4个关键点:指关节、近端指关节、远端指关节、指尖。本文所有对二维和三维的关键点估计都是建立在该手部模型之上。

用深度学习的方法进行手部姿态估计,一般是分阶段进行的。本文提出的网络框架按照处理目的分为4个部分:手部的识别、手部二维关键点的检测、手部三维关键点的检测、手部三维关键点的精细化调整。

手部的识别采用轻量级网络回归手部边框(Bounding Box)作为后续主体网络的预处理操作,使得后续关键点的定位更加准确,同时处理后图片像素的减少也使得后续计算量减小。手部二维和三维关键点的检测通过搭建卷积神经网络提取图像特征,依据区域或者特征的重要程度对权重进行调配,引导级联特征提取模块获取更加丰富的基础提取特征,监督网络主动输出越来越精确的热度置信图。手部三维关键点的精细化调整是将三维关键点粗结果基于图卷积神经网络算法进行优化后处理,拟合出更加精确的手部三维关键点坐标。

2.2 基于YOLOv3的手部识别预处理网络

常见的手部预处理方法分为基于数字图像处理方法和基于深度学习方法两大类。前者一般采取分割算法得到手部掩模定位手部区域,例如将RGB图像转化为灰度图像再转换成二值图像并选取二值图像的前景部分作为手势提取区域^[7]。

本文采取基于深度学习方法的手部预处理

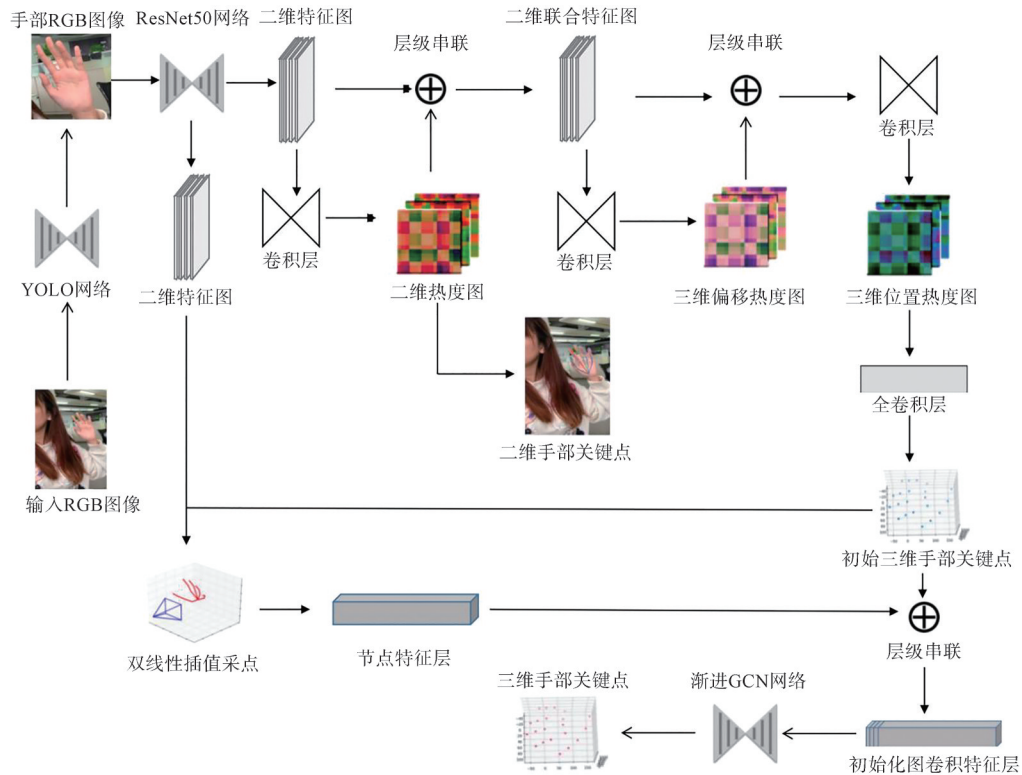


图1 网络流程图

Fig. 1 Network flow chart

网络,如图1所示。在进入主体网络流程前,先采用基于YOLOv3^[8]的预处理网络用于将输入图片中的手部和混杂的背景剥离。YOLOv3相比其他深度学习检测网络的优势在于引入Darknet-53作为骨干网络,采用K-means聚类法^[5]回归出9种大小不同的先验框,并且根据金字塔特征图思想,小尺寸的先验框用于耦合大尺寸的特征图,大尺寸的先验框用于耦合小尺寸的特征图,可以很好地整合不同尺度的感受野的特征,识别输入图片中不同占比的手部。我们将自然场景下的图片输入预处理网络,通过调整Darknet-53网络内卷积核大小可以控制输出的特征图大小,因此对任意尺寸的输入数据经过这个预处理网络都可以输出固定256×256×3尺寸的、剥离背景单独手部的肩部边框图片传输给接下来的网络。

2.3 基于级联特征提取的手部关键点检测网络

如图1所示,二维特征提取模块使用经典的ResNet50^[9]网络,输入256×256×3的RGB图像输出32×32×256的二维特征图。相比于直接回归关节点坐标,基于热度图的方法具有渐变连续

可微分的特点,可以提高坐标估计的精细程度^[10],因此我们在二维和三维检测模块融合多特征热度图。

二维检测模块是一个两层的全连接卷积层。输入32×32×256的二维特征图,输出二维热度图(Heat Maps)。二维热度图包含21个手部关键点的关节预测置信图,通过二维高斯函数编码每个像素点被每个关键点覆盖的置信度,其公式如式(1)所示:

$$H_l(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{((x-u)^2 + (y-v)^2)}{2\sigma^2}}, \quad (1)$$

其中 l 代表第 l 个关键点, σ 为函数的宽度参数, (x, y) 代表该点像素坐标, (u, v) 是中心点坐标,即该关键点二维真值(Ground truth)坐标。

三维检测模块从多热度图和特征图回归三维手部姿态。如图1所示,将二维特征图和二维热度图(2D Heat Maps)进行层级串联,得到二维联合特征图,对二维联合特征图进行卷积操作,得到三维偏移热度图(3D Delta Maps)。三维偏移热度图是子节点相对于根节点的三维方向向量,可以很好地反应父子节点之间的位置关系,

将三维偏移热度图作为中间热度图作为三维检测模块预测结果添加运动学约束,使网络架构嵌入类似手部结构的物理限制。将二维联合特征图和三维偏移热度图进行层级串联和卷积层操作,分别从XYZ坐标轴表示的图中选择置信度最大的点所对应的值为坐标轴的数值,对XYZ轴都执行以上操作,将得到的值保存为三维坐标点^[11],就得到了三维位置热度图(3D Location Maps)。三维位置热度图和二维热度图一样,反

映了每个像素被每个手部关键点的三维坐标覆盖的预测置信度。特征提取网络具体级联结构如图2所示,为了能更好地预测三维位置偏移量,我们先预测一个二维热度图,将其作为三维热度图的一个条件来提升对三维位置预测的准确性。之后我们再将二维热度图和三维偏移图作为共同条件和特征结合在一起去预测最后的三维位置,通过这样多层级联的条件来得到更加准确鲁棒的位置信息。

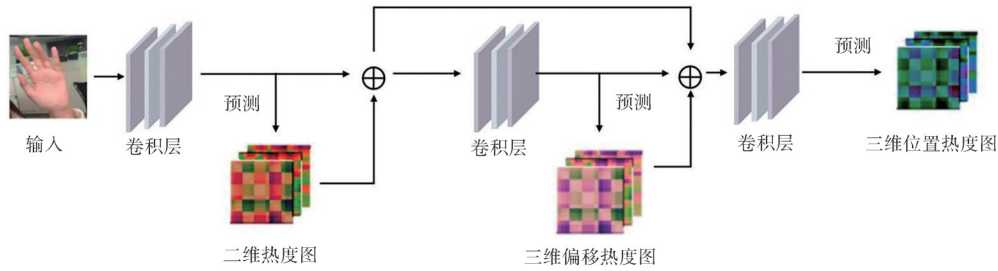


图2 级联特征提取网络结构

Fig. 2 Cascade feature extraction network structure

2.4 基于GCN特征增强的手部骨架回归网络

对于手部骨架而言,它本身是一个天然的图结构。因此我们很自然地想到基于图卷积神经网络(Graph Convolutional Network, GCN)方法来获取它内部的隐式关系^[12]。GCN的计算过程与信号处理过程相同,先将卷积核和图数据通过傅里叶变换转换到频域空间,再对频域空间的系数进行数值运算,最后进行逆傅里叶变换得到卷积后的结果。

利用上述模块生成热图后,采用积分回归方法^[13]将热图表示转化为坐标表示,作为GCN特征增加网络的初始输入姿态。对热度图的初始化姿态进行softed-argmax操作,将热图传播到Softmax层中,该层将热度图像素值标准化为似然值(0~1)之后,再对似然图层进行积分运算求和操作,从而估算关节位置:

$$\hat{J}_i^k = \int_{p \in A} p \cdot H_k(p), \quad (2)$$

其中, $\hat{J}_i^k$ 表示第 k 个关节的位置估计, A 表示似然区域, $H_k(p)$ 表示 p 点上的似然值。因此,每个热图矩阵都包含生成初始姿势的信息。

热图模块和坐标转换相互耦合,使得GCN特征增强网络可以获得更准确的初始化姿态,有助于在进行校正之前获得更精确的局部上下文

理解。此外,由于尺寸的限制,基于热图的表示在一定程度上导致了关键点的量化误差,转化为坐标后可以解决这个问题。

由于手部姿态估计涉及的节点数量较多,我们使用切比雪夫多项式进行逼近。当有 n 个节点时,得到GCN层与层之间传播公式如式(3)所示:

$$Z = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} XW, \quad (3)$$

其中, $\tilde{D} = D + I$, $\tilde{A} = A + I$, I 是单位矩阵, A 是代表各个节点之间位置关系的 $n \times n$ 维的邻接矩阵(Adjacency matrix), $\tilde{D}$ 是 $\tilde{A}$ 的度矩阵(Degree matrix)。 X 是输入层的特征, Z 是输出层的特征, $X \in R^{n \times m}$, m 是特征向量的维度, W 是网络需要学习的权重, $W \in R^{m \times d}$, d 是输出向量的维度。

考虑到特征图之间感受野的由粗到细,我们在模块中设计了一种从粗到精的学习过程,用于增强局部特征学习,纠正部分遮挡的手部关键点的坐标。由于基于坐标的模块缺少图像的上下文信息,我们为每个关节位置挖掘了相关的图像特征,并融合到模块中。如图1所示,通过双线性插值采点,将从图像特征中挖掘出的初始关键点坐标 (x, y) 上的节点特征输入到渐进图卷积层中来改善姿态估计结果。渐进GCN特征增强模块网络结构如图3所示,对抽

取的每个图卷积的节点特征,我们用3个密集连接的GCN模块来抽取特征,并通过层级1、2的预测在每个层级中进行监督,在最后一层输出

预测的三维手部关键点坐标。该机制建立了渐进的GCN架构,并通过逐步融合多尺度图像特征来优化关键点输出。

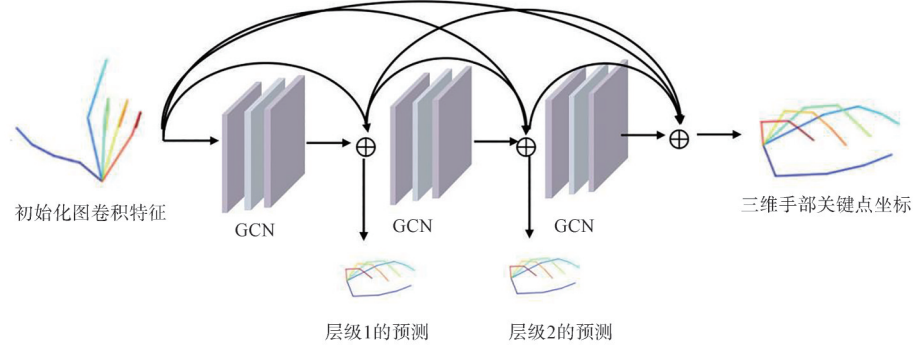


图3 渐进GCN特征增强网络结构

Fig. 3 Progressive GCN feature enhancement network structure

2.5 损失函数

整体网络是基于端到端的学习,通过多任务学习的策略使得网络结果更好地收敛。损失函数的定义如式(4)所示:

$$L_{\text{all}} = \alpha_1 (L_{\text{heat}} + L_{\text{delta}} + L_{\text{loc}}) + L_{\text{GCN}}, \quad (4)$$

其中 L_{heat} 为二维热度图损失函数,表示为 $L_{\text{heat}} = \|H^{\text{GT}} - \hat{H}\|^2$, 这里 H^{GT} 为热度图的真值, $\hat{H}$ 为网络估计的热度图; L_{delta} 为三维偏移热度图损失函数,表示为 $L_{\text{delta}} = \|H^{\text{GT}} \odot (D^{\text{GT}} - \hat{D})\|^2$, 这里 H^{GT} 作为掩模, D^{GT} 为热度图的真值, $\hat{D}$ 为网络估计的热度图, $\odot$ 表示的是逐像素积。同理,三维位置热度图损失函数 $L_{\text{loc}} = \|H^{\text{GT}} \odot (L^{\text{GT}} - \hat{L})\|^2$ 。 L_{GCN} 是图卷积神经网络的损失函数,其定义如式(5)所示:

$$L_{\text{GCN}} = \sum_{i \in \Omega} \sum_{j=1}^n \lambda_j \|(\hat{P}_{ij} - P_i^{\text{GT}}) \odot M\|, \quad (5)$$

其中 Ω 为目标点集,对于我们的手部重建的目标关键点其个数为21个; j 表示的是第 j 层图卷积网络的输出结果。通过这种多层级的三维骨架估计监督可以让网络表现出更好的结果。 M 是一维度的向量表示关键点掩模,如果关键点存在标注数据, $M[i] = 1$, 否则为0。

3 实验结果

3.1 实验指标

对于手部的骨架姿态估计而言,其本质是一

个回归问题,回归问题的指标很难通过上述二分类的指标来进行度量。为了验证模型的优越性,这里使用3D平均准确性(PCK)和2D的关节点相识性(Object keypoint similarity, OKS)进行评价。3D PCK的计算公式如式(6)所示:

$$\text{PCK} = \frac{\sum_i \delta(\frac{d_i}{d} \leq T)}{\sum_i 1}, \quad (6)$$

其中 i 表示关节点的编号, d_i 表示第 i 个关节点的预测值和表达的欧氏距离。 d 为人体骨架的尺度因子, T 为关键点的阈值。在本次实验中 d 为手掌中心岛中指末端的欧式距离, T 的阈值设为0.1。

同理2D OKS的度量公式如式(7)所示:

$$\text{OKS} = \frac{\sum_i \exp\{-d_i^2 / 2s^2 \sigma_i^2\} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \quad (7)$$

其中 s 通常指人体骨架的尺度因子,本次实验设置为手部最小外接框的面积开根; σ_i 是一个归一化因子,与关键点标注的难易有关,这主要是根据数据集的标注分布决定的; v_i 表达的是某些关键点 i 是否存在标答(数据集中该关键点是否标注了真值)。

对于3D的PCK值,由于需要设定一定的阈值,因此在数值结果受到了阈值 T 的影响。在机器学习中通常使用一种度量方式来计算这类数字的分数,我们称之为曲线下面积值的平均(Area Under Curve, AUC)。AUC曲线的计算公式如

式(8)所示:

$$AUC = \frac{\sum_i PCK(T_i)}{\sum_i l}, \quad (8)$$

这里我们设 $T_i = \frac{a_i}{d}$, a_i 设为 20~50 mm, 中间取 10 个间隔。

3.2 实验数据集

本文在以下 5 个数据集上展开模型的训练和测试,多维度验证我们设计的人体手部姿态估计器的性能。

CMU Hand Keypoints Detection Dataset^[14]是由卡内基梅隆大学发布的手部骨架估计的数据集。Rendered Handpose Dataset (RHD)^[15]

是由弗莱堡大学在 2017 发布的手部姿态渲染数据集。Dexter+Object^[16]是 2016 年由德国的马普所发布的手部重建和手部对象跟踪的数据集。Ego Dexter datasets^[17]是由 MPI 于 2017 年发布于 ICCV2017 的数据集。GANeratedDataset^[18]是由 MPI 在 CVPR2018 年推出的数据集,该数据集主要是由 GAN 生产的合成数据集。

实验所用数据的分布如表 1 所示,记录了数据集所含有的标注及被本次实验用作训练集和测试集的数据量。此外,原始的 Dexter^[16]数据集有 1 912 训练集和 846 验证集,但是考虑到更加充分的验证算法,我们将 1 912 个训练集也作为本次实验的测试集。

表 1 实验所用数据分布

Tab. 1 Distribution of data used in the experiment

	手部 2D 标注	手部 3D 标注	训练集	测试集
CMU ^[14]	✓	—	16 720	—
RHD ^[15]	✓	✓	41 258	2 728
Dexter+Object ^[16]	✓	✓	—	3 145
EgoDexter ^[17]	—	✓	—	3 190
GANeratedDataset ^[18]	—	✓	331 499	—
合计	—	—	389 477	9 063

3.3 结果与分析

实验过程中的数值曲线如图 4 所示,其中图 4(a)是训练过程中的损失曲线,lossH 是二维热度图损失值,lossD 是三维偏移热度图的损失值,lossL 是在经过渐进 GCN 模块后的三维关键点损失值。从图 4(a)可以看出,输出的粗结果在经过多热度图耦合的三维关键点检测器和渐进的 GCN 模块的精细化调整后收敛出更低的损失值。图 4(b)、(c)、(d)分别表示在 RHD^[15]、DO^[16]、ED^[17]测试集下每个训练周期下的 AUC 值。

本次实验主要和 Xiong Zhang 等人提出的 Mesh2HAND^[19]、Donglai Xiang 等人提出的 Mon-Cap^[20]、Adnane Boukhayma 等人提出的 3D pose in the wild^[21]以及 Y X Zhou 等人提出的 Minimal-hand^[5]进行对比。

定量的实验结果与上述 4 种方法比较对照如表 2 所示,其中 20 mm 和 30 mm 分别指当阈值取相应值时的 PCK 值,AUC 是当阈值取 20~50 mm 时的 PCK 曲线面积值。由表 2 可见,我们所改进

的方法在 3D 骨架回归任务上的结果在各个数据集上的 AUC 曲线都优于其他方法。具体来说,我们所提出的方法相较于 Minimal-hand^[5]在 DO 数据集上 AUC 大约高了 0.8%,在 ED^[17]数据集上比 Minimal-hand^[5] AUC 大约高了 0.7%。对于 RHD^[15]数据集,我们所设计的方法比 Mesh2 Hand^[19] AUC 高了大约 3%。就单纯的 PCK 值而言,我们所提出的方法更加接近于真实值,在阈值设为 20 mm 处,我们提出的方法在 DO^[16]数据集上比最好的方法高了 0.9%,在 RHD^[15]数据集上比最好的方法高了 0.8%。虽然在 ED^[17]数据集上略低于最好的方法,但是在阈值 30 mm 处,我们提出的方法远高于最好的方法(相较于 Minimal-hand^[5]高了大约 3.7%)。

本次实验的操作系统为 Ubuntu18.04,CPU 核为 i5-6500,GPU 的配置为 GTX-2080 11 GB。在图像分辨率为 256×256 的情况下,对算法处理时间和所需功耗进行分析。如表 3 所示,算法所需的推理时间为 52 ms,算法所需的每秒浮点计

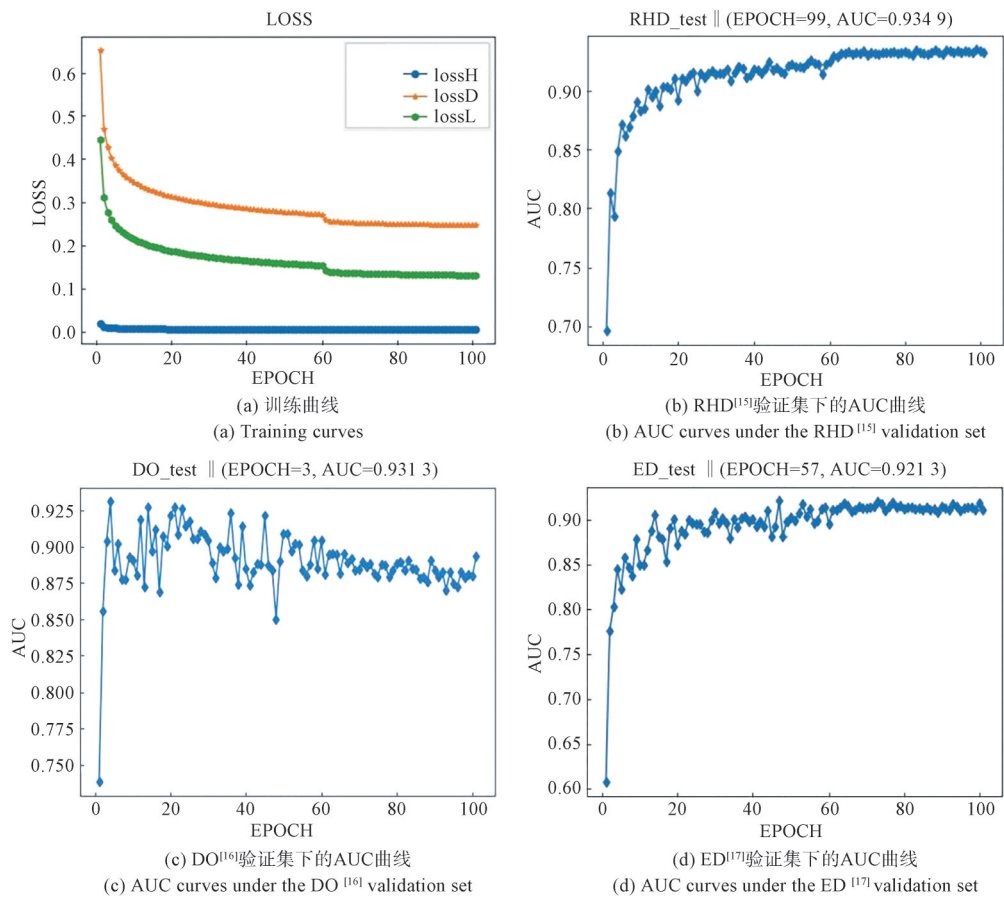


图4 网络训练时的损失函数曲线和在验证集下的AUC精度曲线

Fig. 4 Loss function curves during network training and AUC accuracy curves under the validation set

表2 本文方法与其他方法实验结果对比

Tab. 2 Comparison of experimental results between this method and other methods									
数据 方法	Dexter+Object ^[16]			EgoDexter ^[17]			RHD ^[15]		
	20 mm	30 mm	AUC	20 mm	30 mm	AUC	20 mm	30 mm	AUC
Mesh2HAND ^[19]	0.732	0.856	0.825	0.717	0.744	0.746	0.800	0.915	0.901
Mon-Cap ^[20]	0.795	0.933	0.912	0.732	0.811	0.779	0.816	0.923	0.911
Adane ^[21]	0.653	0.819	0.763	0.636	0.701	0.674	0.798	0.902	0.887
Minimal-hand ^[5]	0.821	0.953	0.923	0.838	0.917	0.914	0.799	0.883	0.856
本文方法	0.830	0.967	0.931 3	0.821	0.954	0.921 3	0.824	0.961	0.934 9

算量(FLOPs)为 9.3×10^8 次。结合算法对照实验分析可知,我们提出的算法在处理时间上和模型复杂度上处于较优水平,不仅推理时间和计算量近似于 Minimal-hand^[5],而且姿态估计精度在多个数据集上都超过了 Minimal-hand^[5]的效果。

该结果表明在不影响算法速度的情况下,我们提出的算法在效率上得到了较大的提升,识别推理的帧率(Frames Per Second,FPS)达到了

表3 推理速度与浮点计算量比较

Tab. 3 Comparison of the time of inference and FLOPs		
	推理时间/ms	计算量(FLOPs)
Mesh2Hand ^[19]	1 173	11. 3E10
Mon-Cap ^[20]	564	7. 9E9
Adanet ^[21]	642	8. 5E9
Minimal-hand ^[5]	49	8E8
本文方法	52	9. 3E8

19.23,因此该算法可以使用在视频流的实时手部骨架检测中。与此同时,较小的计算量满足了对模型低功耗、轻量化的需求。

我们选择在上述定量分析中4个对照方法中精度指标表现最好的 Minimal-hand^[5]作为定性分析对象。定性的可视化结果如图5所示(在RHD^[15]

测试集上的检测结果)。从图中高亮部分的细节可以看出,本文算法在测试集上的结果明显优于 Minimal-hand^[5]算法。在引入渐进 GCN 模块后,手部骨架耦合了图结构的约束,使得其在一些自遮挡比较严重的场景下,也能够检测出合理的结果。

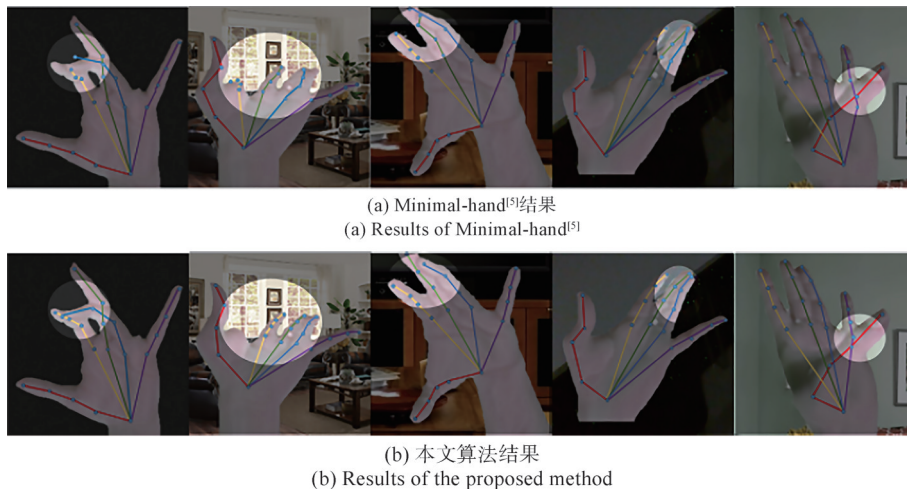


图5 可视化实验结果对比

Fig. 5 Comparison of visualization experimental results

在自然场景下,涉及手物交互时骨架回归结果如图6所示。由于人手在抓取物体过程中出现了大面积的遮挡,导致骨架回归结果在尺度的还原上有些许偏差。除了部分关键点遮挡因素外,我们分析误差的产生还有以下原因:由于当前手部3D训练数据集真实数据不足,因此参与实验的训练数据大多是CG合成的虚拟3D数据集,或者是通过实验室设备所收集的特定

场景下标定好的真实数据集,并且单目RGB相机在投影时失去了深度信息。因此将该算法用于自然场景下时,存在着较大的状态空间误差(Domain Gap),从而导致了些许计算误差。此外,最终输出的3D坐标的回归在一定程度上依赖于初始化的3D坐标位置,在初始坐标位置偏差较大时,会导致系统估计的姿态坐标产生回归误差。

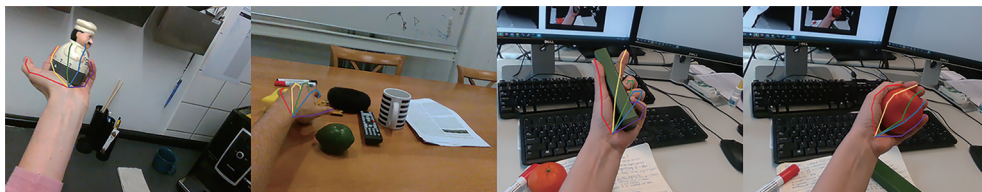


图6 自然场景下的可视化结果

Fig. 6 Visualization results in the wild

尽管如此,本文模型在精度指标上和大致的外形姿态上还是表现得足够精准,总体上算法可以在自然场景下给出合理的估计结果,这体现了我们算法整体的鲁棒性。

4 结 论

本文所提出的三维手部姿态估计算法通过结合人体关节结构之间的基本约束信息以及级联

神经网络和渐进图卷积神经网络挖掘出的特征图中包含的被遮挡关键点的相关数据,能够精确地调整被遮挡关键点的位置,对于人体手部骨架的检测有较高的正确率。我们在 3 个公开数据集上验证了网络性能,与现有的 4 种算法进行实验对比,本文算法在 30 mm 阈值下的 PCK 精度表现非常出色,最低精度达到 95.4%,高出次优算法 3.7% 精度值。并且在 3 个测试集上的 AUC 曲线指标均达到最高,平均 AUC 精度达到 92.9%。经过与其他方法的定量、定性类比,可以看出本文

算法在三维关键点预测精度上比现有的方法有明显提高,并且可视化实验结果表明,本文算法在图像细节的捕捉上也具有优势。此外,本文方法推理的帧率达到了 19.23,可以满足实时视频流检测的需求。模型所需 FLOPs 为 9.3×10^8 次,复杂度较低,满足对模型轻量化的要求。综上,本文基于多任务学习的方法提出了一种端到端的训练方式,加速了网络的收敛,减少了特征的过拟合,使得网络在三维手部姿态估计任务上准确性和鲁棒性相比于现有技术有较为显著的改进。

参 考 文 献:

- [1] CAO Z, SIMON T, WEI S E, *et al.* Realtime multi-person 2D pose estimation using part affinity fields [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 1302-1310.
- [2] FANG H S, XIE S Q, TAI Y W, *et al.* RMPE: regional multi-person pose estimation [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 2353-2362.
- [3] CAI Y J, GE L H, CAI J F, *et al.* Weakly-supervised 3D hand pose estimation from monocular RGB images [C]//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018: 678-694.
- [4] GE L H, REN Z, LI Y C, *et al.* 3D hand shape and pose estimation from a single RGB image [C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 10825-10834.
- [5] ZHOU Y X, HABERMANN M, XU W P, *et al.* Monocular real-time hand shape and motion capture using multi-modal data [C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 5345-5354.
- [6] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [C]//*Proceedings of the 5th International Conference on Learning Representations*. Toulon: OpenReview.net, 2017.
- [7] 负卫国,史其琦,王民. 基于深度卷积神经网络的多特征融合的手势识别[J]. 液晶与显示, 2019, 34(4): 417-422.
YUN W G, SHI Q Q, WANG M. Multi-feature fusion gesture recognition based on deep convolutional neural network [J]. *Chinese Journal of Liquid Crystals and Displays*, 2019, 34(4): 417-422. (in Chinese)
- [8] REDMON J, FARHADI A. YOLOV3: an incremental improvement [J]. *arXiv*: 1804.02767, 2018.
- [9] HE K M, ZHANG X Y, REN S Q, *et al.* Deep residual learning for image recognition [C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 770-778.
- [10] PFISTER T, CHARLES J, ZISSERMAN A. Flowing convnets for human pose estimation in videos [C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015: 1913-1921.
- [11] MEHTA D, SRIDHAR S, SOTNYCHENKO O, *et al.* VNect: real-time 3D human pose estimation with a single RGB camera [J]. *ACM Transactions on Graphics*, 2017, 36(4): 44.
- [12] ZHAO L, PENG X, TIAN Y, *et al.* Semantic graph convolutional networks for 3D human pose regression [C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 3420-3430.
- [13] SUN X, XIAO B, WEI F Y, *et al.* Integral human pose regression [C]//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018: 536-553.
- [14] SIMON T, JOO H, MATTHEWS I, *et al.* Hand keypoint detection in single images using multiview bootstrapping [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 4645-4653.

- [15] ZIMMERMANN C, BROX T. Learning to estimate 3D hand pose from single RGB images [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 4913-4921.
- [16] SRIDHAR S, MUELLER F, ZOLLHÖFER M, *et al.* Real-time joint tracking of a hand manipulating an object from RGB-D input [C]//*Proceedings of the 14th European Conference on Computer Vision*. Amsterdam: Springer, 2016: 294-310.
- [17] MUELLER F, MEHTA D, SOTNYCHENKO O, *et al.* Real-time hand tracking under occlusion from an egocentric RGB-D sensor [C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice: IEEE, 2017: 1163-1172.
- [18] MUELLER F, BERNARD F, SOTNYCHENKO O, *et al.* Generated hands for real-time 3D hand tracking from monocular RGB [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 49-59.
- [19] ZHANG X, LI Q, MO H, *et al.* End-to-end hand mesh recovery from a monocular RGB image [C]//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 2354-2364.
- [20] XIANG D L, JOO H, SHEIKH Y. Monocular total capture: posing face, body, and hands in the wild [C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 10957-10966.
- [21] BOUKHAYMA A, DE BEM R, TORR P H S. 3D hand shape and pose from images in the wild [C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 10835-10844.

作者简介:



林依林(1997—),女,福建南平人,硕士研究生,2019年于福州大学获得学士学位,主要从事基于计算机视觉的三维人体姿态估计研究。E-mail: 370300890@qq.com



林志贤(1975—),男,福建泉州人,博士,教授,2010年于福州大学获得博士学位,主要从事图像处理技术以及平板显示器件驱动技术方面的研究。E-mail: lzx2005000@163.com