E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)01-0254-14

论文引用格式:Dong J, Zhang H R, Fang X Y, Zhou D S, Yang X, Zhang Q and Wei X P. 2025. Combining multi-view controlled fusion and joint correlation for 3D human pose estimation. Journal of Image and Graphics, 30(01):0254-0267(董婧, 张鸿儒, 方小勇, 周东生, 杨鑫, 张强, 魏小鹏. 2025. 联合多视图可控融合和关节相关性的三维人体姿态估计. 中国图象图形学报, 30(01):0254-0267)[DOI:10.11834/jig.230908]

联合多视图可控融合和关节相关性的三维人体姿态估计

董婧[1],张鸿儒[1],方小勇[2*],周东生[1,3],杨鑫[3],张强[1,3],魏小鹏[3]

1. 大连大学先进设计与智能计算教育部重点实验室,大连　116622; 2. 湖南工学院安全与管理工程学院,衡阳　421002;
3. 大连理工大学计算机科学与技术学院,大连　116024

摘　要:**目的**　多视图三维人体姿态估计能够从多方位的二维图像中估计出各个关节点的深度信息,克服单目三维人体姿态估计中因遮挡和深度模糊导致的不适定性问题,但如果系统性能被二维姿态估计结果的有效性所约束,则难以实现最终三维估计精度的进一步提升。为此,提出了一种联合多视图可控融合和关节相关性的三维人体姿态估计算法CFJCNet(controlled fusion and joint correlation network),包括多视图融合优化模块、二维姿态细化模块和结构化三角剖分模块3部分。**方法**　首先,基于极线几何框架的多视图可控融合优化模块有选择地利用极线几何原理提高二维热图的估计质量,并减少噪声引入;然后,基于图卷积与注意力机制联合学习的二维姿态细化方法以单视图中关节点之间的联系性为约束,更好地学习人体的整体和局部信息,优化二维姿态估计;最后,引入结构化三角剖分以获取人体骨长先验知识,嵌入三维重建过程,改进三维人体姿态的估计性能。**结果**　该算法在两个公共数据集Human3.6M、Total Capture和一个合成数据集Occlusion-Person上进行了评估实验,平均关节误差为17.1 mm、18.7 mm和10.2 mm,明显优于现有的多视图三维人体姿态估计算法。**结论**　本文提出了一个能够构建多视图间人体关节一致性联系以及各自视图中人体骨架内在拓扑约束的多视图三维人体姿态估计算法,优化二维估计结果,修正错误姿态,有效地提高了三维人体姿态估计的精确度,取得了最佳的估计结果。

关键词:多视图;三维人体姿态估计;关节相关性;图卷积网络(GCN);注意力机制;三角剖分

Combining multi-view controlled fusion and joint correlation for 3D human pose estimation

Dong Jing[1], Zhang Hongru[1], Fang Xiaoyong[2*], Zhou Dongsheng[1,3], Yang Xin[3], Zhang Qiang[1,3], Wei Xiaopeng[3]

1. Key Laboratory of Advanced Design and Intelligent Computing (Ministry of Education), Dalian University, Dalian 116622, China;
2. School of Safety and Management Engineering, Hunan Institute of Technology, Hengyang 421002, China;
3. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

收稿日期:2024-01-15;**修回日期**:2024-05-13;**预印本日期**:2024-05-20

*通信作者:方小勇　xyfang@hnit.edu.cn

基金项目:111计划项目(D23006);辽宁省高校创新团队支持计划项目(LT2020015);大连市重点领域创新团队支持计划项目(2021RT06);大连市重大基础研究项目(2023JJ11CG002);辽宁省教育厅科研计划项目(LJKMZ20221839);大连大学创新团队支持计划项目(XLJ202010);大连大学学科交叉项目(DLUXK-2024-YB007);湖南省自然科学基金项目(2022JJ30017)

Supported by: 111 Project(D23006); Program for Innovative Research Team in University of Liaoning Province(LT2020015); Support Plan for Key Field Innovation Team of Dalian(2021RT06); Dalian Major Projects of Basic Research(2023JJ11CG002); Scientific Research Foundation of Education Department of Liaoning Province(LJKMZ20221839); Natural Science Foundation of Hunan Province, China(2022JJ30017)

Abstract: **Objective** 3D human pose estimation is fundamental to understanding human behavior and aims to estimate 3D joint points from images or videos. It is widely used in downstream tasks such as human-computer interaction, virtual fitting, autonomous driving, and pose tracking. According to the number of cameras, 3D human pose estimation can be divided into monocular 3D human pose estimation and multi-view 3D human pose estimation. The ill-posed problem caused by occlusion and depth ambiguity means that estimating the 3D human joint points by monocular 3D human pose estimation is difficult. However multi-view 3D human pose estimation can obtain the depth of each joint from multiple images, which can overcome this problem. In most recent methods, the triangulation module is used to estimate the 3D joint positions by leveraging their 2D counterparts measured in multiple images to 3D space. This module is usually used in a two-stage procedure: First, the 2D joint coordinates of the human on each view are estimated separately by using a 2D pose detector, and then the 3D pose from multi-view 2D poses by applying triangulation. On this basis, some methods work with epipolar geometry to fuse the human joint features to establish the correlation among multiple views, which can improve the accuracy of 3D estimation. However, when the system performance is constrained by the effectiveness of the 2D estimation results, improving the final 3D estimation accuracy further is difficult. Therefore, to extract human contextual information for more effective 2D features, we construct a novel 3D pose estimation network to explore the correlation of the same joint among multiple views and the correlation between neighbor joints in the single view. **Method** In this paper, we propose a 3D human pose estimation method based on multi-view controllable fusion and joint correlation (CFJCNet), which includes three parts: a controllable multi-view fusion optimization module, a 2D pose refinement module, and a structural triangulation module. First, a set of RGB images captured from multiple views are fed into the 2D detector to obtain the 2D heatmaps, and then the adaptive weights of each heatmap are learned by a weight learning network with appearance information and geometric information branches. On this basis, we construct a multi-view controlled fusion optimization module based on epipolar geometry framework, which can analyze the estimation quality of joints in each camera view to influence the fuse process. Specifically, it selectively utilizes the principles of epipolar geometry to fuse all views according to the weights, thus ensuring that the low-quality estimation can benefit from auxiliary views while avoiding the introduction of noise in high-quality heatmaps. Subsequently, a 2D pose refinement module composed of attention mechanisms and graph convolution is applied. The attention mechanism enables the model to capture the global content by weight assignment, while the graph convolutional network (GCN) can exploit local information by aggregating the features of the neighbor nodes and instruct the topological structure information of the human skeleton. The network combining the attention and GCN can not only learn human information better but also construct the interdependence between joint points in the single view to refine 2D pose estimation results. Finally, structural triangulation is introduced with structural constraints of the human body and human skeleton length in the process of 2D-to-3D inference to improve the accuracy of 3D pose estimation. This paper adopts the pre-trained 2D backbone called simple baseline as the 2D detector to extract 2D heatmaps. The threshold $\varepsilon = 0.99$ is used to determine the joint estimation quality, and the number of layers $N = 3$ is designed for 2D pose refinement. **Result** We compare the performance of CFJCNet with that of state-of-the-art models on two public datasets, namely, Human3.6M and Total Capture, and a synthetic dataset called Occlusion-Person. The mean per joint position error (MPJPE) is used as the evaluation metric, which measures the Euclidean distance between the estimated 3D joint positions and the ground truth. MPJPE can reflect the quality of the estimated 3D human poses, providing a more intuitive representation of the performance of different methods. On the Human3.6M dataset, the proposed method achieves an additional error reduction of 2.4 mm compared with the baseline Adafuse. Moreover, because our network introduces rich priori knowledge and effectively constructs the connectivity of human joints, CFJCNet achieves at least a 10% improvement compared with most methods that do not use the skinned multi-person linear (SMPL) model. Compared with learnable human mesh triangulation (LMT) incorporating the SMPL model and volumetric triangulation, our method still achieves a 0.5 mm error reduction. On the Total Capture dataset, compared with the excellent baseline Adafuse, our method exhibits a performance improvement of 2.6%. On the Occlusion-Person dataset, the CFJCNet achieves optimal estimation for the vast majority of joints, which improves performance by 19%. Furthermore, we compare the visualization results of 3D human pose estimation between our method and the baseline Adafuse on the Human3.6M dataset and the Total Capture dataset to provide a more intuitive demonstration of the estimation performance. The qualitative experimental results on

both datasets demonstrate that CFJCNet can use the prior constraints of skeleton length to correct unreasonable erroneous poses. **Conclusion** We propose a multi-view 3D human pose estimation method CFJCNet, which is capable of constructing human joint consistency between multiple views as well as intrinsic topological constraints on the human skeleton in the respective views. The method achieves excellent 3D human pose estimation performance. Experimental results on the public datasets show that CFJCNet has significant advantages in evaluation metrics over other advanced methods, demonstrating its superiority and generalization.

Key words: multi-view; 3D human pose estimation; joint point correlation; graph convolutional network(GCN); attention mechanism; triangulation

0 引 言

三维人体姿态估计能够从图像或者视频中估计出人体关节点的三维坐标,精确展示人体信息,广泛应用于人机交互(Pascual等,2022)、虚拟试衣(Minar和Ahn,2021)、自动驾驶(Du等,2019)、姿态追踪(马淼 等,2020)、情感识别(闫静杰 等,2013)等下游任务。然而,现有的单目三维人体姿态估计方法对遮挡问题十分敏感且难以获取精确的深度估计信息。与之不同的是,多视图三维人体姿态估计能够从多个角度视图的测量结果中受益,从而获取缺失的深度信息以解决这种不适定性问题。

多视图人体姿态估计通常采用二阶段的方式来获取三维人体姿态。该类方法最常见的框架主要分为两个步骤:1)从同一个对象在不同视角下的图像中提取其各自视图下的所有二维关节点坐标;2)从获得的多个视图下的二维姿态中重建出三维人体姿态。

现有的三维人体姿态估计算法主要围绕第1个步骤展开研究,Kadkhodamohammadi和Padoy(2021)提出一种可归纳的三维人体姿态回归方法,将在所有视图中提取的二维关节点坐标输入到全连接网络以预测全局三维关节点坐标;可学习三角剖分(Iskakov等,2019)通过二维检测器提取多视图二维关节点坐标,并与置信度结合估计三维人体姿态。然而,这些算法通常没有考虑多视图人体关节一致性和人体骨架内在的拓扑约束。一方面,在检测二维姿态后忽视了多个视图之间的信息交互,从而导致估计结果失去了视图之间的一致性。为此,极线Transformer(He等,2020)、自适应多视图融合(Zhang等,2021)、跨视图自监督融合(Kim等,2023)等方法基于极线几何的原理,将关节点特征与其在其他视图极线上的特征进行融合,以此构建多视图间的联系。另一方面,仅有部分研究者关注了人体的上下文信息,在跨视图融合Transformer(Ma等,2021)和人体标签剪枝Transformer(Ma 等,2022)方法中,利用Transformer对二维特征图进行处理,以捕捉人体的全局信息。虽然这种方法能够学习到人体的内部联系,但网络的计算成本过于昂贵。

因此,为了进一步取得更好的三维人体姿态估计效果,需要探索人体关节间的相关性,包括多视图间同一关节的相关性和单视图中不同关节之间的相关性。首先,同一个关节在不同视图间的信息一致性是多视图融合的关键,但现有方法在构建的过程中忽略了对单视图估计质量的评价,导致融合时可能引入部分噪声,反而降低了二维热图估计的有效性。同时,人体关节是一个有机的整体,同一视图中不同关节之间的联系常被忽视,而人体骨架间的内在约束可以有效保证三维姿态估计的合理性。基于此,本文构建了人体关节点的相关性模型(controlled fusion and joint correlation network,CFJCNet),通过视图权重融合和人体骨骼先验的引入,取得了先进的估计结果。

本文的贡献如下:1)提出了一个基于极线几何框架的多视图可控融合优化模块,利用极线几何原理,根据各视图中同一关节的估计权重融合不同质量的热图矩阵,从而保证了低质量关节热图从其他视图收益的同时减少了高质量关节热图中的噪声引入,进而提高二维姿态估计的可信度。2)提出了由注意力机制和图卷积构成的二维姿态细化模块,该模块通过注意力机制和图卷积分别学习人体关节点之间的整体信息和局部信息,并通过汇聚两种不同信息以帮助网络全面学习人体关节点的内在联系,实现二维姿态的进一步优化。3)引入了基于人体结构和骨长先验的结构化三角剖分方法,以相邻关节

间的骨长为约束，逐步优化人体结构的树型骨向量模型，实现精确的三维姿态估计。

1 相关工作

1.1 多视图中的二维特征优化

为了实现更为准确的三维人体姿态估计，许多研究者从三维重建前的特征入手，获取更为良好的二维人体姿态，进而实现精准重建。

He 等人（2020）构建了极线 Transformer 网络对极线上的像素进行采样并通过加权平均和来近似参考点的特征，从而实现了多视图特征融合，获得三维感知信息以改进二维姿态估计。Zhang 等人（2021）提出了自适应多视图融合方法，应用热图外观和多视图几何信息学习融合权重，并通过极线几何实现多视图信息融合，提高了二维热图的质量，在 human3. 6M 上取得了 19. 5 mm 的平均关节误差（mean per joint position error，MPJPE）指标结果。跨视图融合 Transformer 方法（Ma 等，2021）将 Transformer 直接应用于多个视图特征的融合过程，利用辅助视图中的全局信息修正主特征图中每个通道所对应的像素值，增强其特征表示。人体标签剪枝 Transformer（Ma 等，2022）则对此进行了改进，通过剪枝令牌使得 Transformer 能够关注特征图中人体所在区域而非整幅特征图的全局信息，降低计算成本。此外，部分学者从多个视角的特征图中聚集出体素特征，利用三维卷积获得三维人体姿态。Remelli 等人（2020）提出了一个轻量级多视图三维人体姿态估计方法，通过将特征图重塑为点集并根据投影矩阵进行条件变换，从而将特征图映射到同一规范表征空间，融合出姿态潜在表征，实现了精准且快速的三维人体重建。体素 Transformer 网络（Chen 等，2023b）通过聚合来自所有视图的二维关节点特征，学习三维体素空间中的相对关系，并引入残差结构和稀疏 Sinkhorn 注意力机制，在降低内存消耗的同时提高了人体姿态估计性能。体素姿态估计方法（Tu 等，2020）遵循体素三角剖分的设计思路，由二维特征图聚合为体素空间，以此提取三维表征，重建人体姿态。

1.2 关节点联系

对人体关节点的相关性探索主要聚焦于单目人体姿态估计中二维向三维提升的研究，由于深度信息的缺失，这类方法通常采用图卷积或者注意力机制分别构建人体的局部联系与整体联系，帮助网络实现更好的姿态估计。

由于图卷积网络对图表征具有良好的学习能力，可将其用于人体拓扑结构的学习。然而，图卷积只能聚集相邻节点的信息以构建关节点间的局部联系，这也导致了基于图卷积的方法缺乏对全局关系的学习。研究者通过高阶图卷积滤波器（Lu 等，2023）或构建多跳关节点之间的关系（Hassan 和 Ben Hamza，2023；Wu 等，2022）尽可能地引入更长距离的联系来解决这种问题。

注意力机制则关注于人体实现各种姿态时的关键节点信息，通过 Transformer 网络（Li 等，2022）或门自控系统对输入的姿态特征进行计算，分配相应的权重，能够提取出对姿态估计最重要部分的特征，因此注意力机制常用于构建人体关节之间的全局信息。为了构建更为细粒度的联系，自适应多视图与时间融合 Transformer（Shuai 等，2023）根据人体关节之间的相对独立和运动联系，将输入的人体关节特征分为 5 个部分，从而起到了类似局部联系的作用。而 HDFormer（high-order directed transformer）（Chen 等，2023a）引入了细粒度人体运动先验知识，提出了高阶 Transformer 以构建人体的高阶交互关系。门自控系统则利用计算量较小的优势，对输入的特征信息进行计算，获得门控权重以起到注意力的作用。SENet（squeeze-and-excitation networks）（Hu 等，2018）通过显式建模通道之间的相互依赖性，计算门控权重，自适应地重新校准通道特征。MLP-JCG（multi-layer perceptron with joint-coordinate gating）（Tang 等，2023）利用门控单元和选择性门控单元分别聚集沿关节通道和空间通道上的人体运动学信息，从而提取出人体的全局结构信息。

1.3 三角剖分优化

受益于可学习三角剖分（Iskakov 等，2019）中提出的代数三角剖分和体素三角剖分，多视图人体姿态估计得到了进一步的发展。受此启发，部分学者从优化三角剖分的角度开展工作。整体三角剖分（Wan 等，2023）通过主成分分析提取解剖学先验知识，并将人体结构特征引进三角剖分过程，从而考虑了人体的全局上下文关系与逐关节之间的联系。结构化三角剖分（Chen 等，2022）则通过引入人体骨长约束丰富先验知识，并采用步长约束算法递推求解以减小错误估计，实现了更为合理的三维重建。可

学习的网格三角剖分(Chun等,2023)通过基于可见性的子顶点解决了内存使用率过大问题,利用体素三角剖分估计出人体表面顶点的同时基于蒙皮多人线性(skinned multi-person linear,SMPL)人体模型拟合表面,实现了更为精准的三维人体姿态估计。

2 方 法

本文探索了多视图下人体关节的相关性,构建了多视图的三维姿态估计网络,包含了一个基于极线几何框架的多视图可控融合优化模块、基于图卷积与注意力机制共同学习人体关节联系性的二维姿态细化模块,以及引入人体骨架结构与长度先验知识的结构化三角剖分模块。

图1展现了CFJCNet网络的整体框架。网络的输入是一组多个视图下的RGB图像$\boldsymbol{I}_c \in \mathbf{R}^{C\times H\times W\times 3}$,其中$c \in \{1,2,\cdots,C\}$是视图编号,$H$和$W$分别为图像的高和宽。二维检测器从中提取出多个视图下的原始二维热图$\boldsymbol{M}_c \in \mathbf{R}^{C\times \frac{H}{4}\times \frac{W}{4}\times J}$,再经由权重$w \in \mathbf{R}^{C\times J}$进行融合,获得二维姿态$\boldsymbol{P} \in \mathbf{R}^{J\times 2}$,其中$J$为人体的关节数量。基于联合图卷积和注意力机制的细化模块修正姿态估计中的错误,并通过结构化三角剖分将其重建为精确的三维人体姿态$\boldsymbol{Y} \in \mathbf{R}^{J\times 3}$。

本节在权重学习的基础上,重点阐述基于极线几何框架的多视图可控融合优化模块、二维姿态细化模块和结构化三角剖分3个部分。

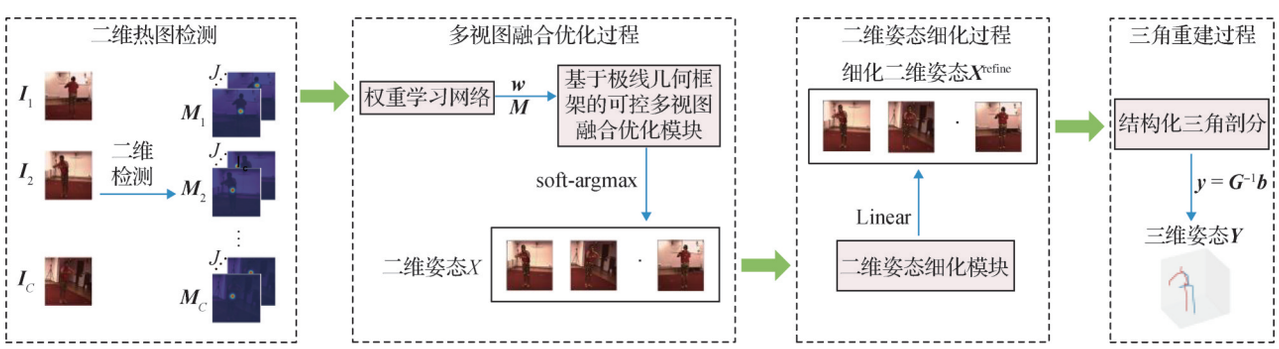


图1 CFJCNet流程

Fig. 1 Overview of CFJCNet

2.1 权重学习网络

在多视图融合中,如果将所有视图的关节特征采用相同权重进行融合,则不同质量的视图对于最终融合热图的贡献度将完全相同,这忽视了低质量估计中的不良特征对良好视图估计结果的干扰,降低了二维姿态估计效果。权重学习网络(Zhang等,2021)通过两个分支学习单个热图的外观信息和多个热图之间的几何信息,根据视图质量的高低赋予不同的权重,在融合视图中体现出相应的贡献度。

第1个分支将热图的高斯核形状作为评定关节估计质量的标准之一。具有少量高关节响应值的热图具有较好的估计质量;反之,若一个热图上具有多个低响应值则说明其估计质量不佳。第2个分支通过获取各个视图之间的几何一致性判断各视图中的关节估计质量。基于极线几何的原理,同一个关节在多个视图中的位置信息可以利用各自已知的相机参数连结起来,通过计算Samposon距离来表示两个视图下同一关节的估计一致性,距离越小说明两个关节的估计结果越接近。任意两组关节点$\boldsymbol{x}_1$和$\boldsymbol{x}_2$之间的Samposon距离$S(\cdot)$为

$$\mathrm{S}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\frac{\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{F}\boldsymbol{x}_1}{(\boldsymbol{F}\boldsymbol{x}_1)_1^2+(\boldsymbol{F}\boldsymbol{x}_1)_2^2+(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{x}_2)_1^2+(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{x}_2)_2^2} \tag{1}$$

式中,下标1或2表示向量的第1个或第2个元素,$\boldsymbol{F}$为由两个相机参数求得的点与极线之间的转换矩阵。具体而言,一个关节的估计结果与其他视图中的关节估计越一致,几何分支便会判定该估计质量越好。

最后,网络输出层汇总两个分支的信息以学习融合权重。

2.2 基于极线几何框架的多视图可控融合优化模块

得益于二维热图的稀疏性,二维热图的融合过程可以借由极线几何原理跳过精确的点对点匹配过程,将源视图上的关节点特征与获取的对应参考视图极线上的匹配点特征进行融合(Zhang等,2021)。

然而，这种多视图的权重融合方式虽然可以使得低质量关节估计受益于其他视图的高质量关节特征，但是却忽视了高质量视图融合过程中可能存在的噪声引入。

基于此，本节提出了一个基于极线几何框架的多视图可控融合优化模块，以构建视图之间的有效联系，如图2所示。

融合权重是网络学习得到的用于评定视图质量的量化标准，通过对每个关节融合权重的分析可以判定二维关节的估计质量，设计不同的关节融合策略。具体而言，同一关节在不同视图中，由于遮挡或二维估计检测性能等原因呈现出不同的估计质量，并能够通过由权重学习网络获取的权重较为直观地表征出来，从而使得模块能够通过设定评估阈值 ε 实现对关节估计质量的判定。当所评定视图 $\boldsymbol{c}$ 下的关节 j 的融合权重 $w_{j,c}$ 大于阈值 ε 时，说明该关节的热图估计较为精准，可直接作为其在相应视图下的优化热图；反之，当所评定的关节融合权重 $w_{j,c}$ 小于阈值 ε 时，说明该关节估计质量欠佳，需要融合参考视图中的高质量特征，补充缺失的信息，进行优化。

因此，多视图融合优化模块最终会通过阈值判断对相应视图下的关节融合方式进行选择，从而根据不同的估计质量获得优化热图 $\boldsymbol{M}_{j,c}$。具体过程为

$$\hat{\boldsymbol{M}}_{j,c}=\begin{cases}\boldsymbol{M}_{j,c} & w_{j,c}>\varepsilon\\ \boldsymbol{M}_{j,c}^{\text{fuse}} & w_{j,c}\leqslant\varepsilon\end{cases} \tag{2}$$

式中，$\boldsymbol{M}_{j,c}$ 表示视图 $\boldsymbol{c}$ 下的关节 j 的原始热图。$\boldsymbol{M}_{j,c}^{\text{fuse}}$ 表示视图 $\boldsymbol{c}$ 下的关节 j 的多视图融合热图，可通过极线几何求取主视图下原始热图中的每一点在参考视图中对应的极线，并在其上寻找响应值最大点作为匹配点，根据权重进行融合获得，过程为

$$\hat{R}^v(x)=w^vR^v(x)+\sum_{c=1}^{C-1}w^c\max_{x'\in I^c(x)}R^c(x') \tag{3}$$

式中，w^v 为源视图的融合权重，而 w^c 则为各个参考视图的融合权重。$R^v(x)$ 为源视图 $\boldsymbol{v}$ 中 x 点的响应值，$\boldsymbol{I}^c(x)$ 为 x 点在参考视图 $\boldsymbol{c}$ 上的极线，而 x' 为极线上的一个点，$R^c(x')$ 为参考视图中极线上的最大响应值，$\hat{R}^v(x)$ 为融合后源视图 $\boldsymbol{v}$ 中 x 点的响应值。最终，由低质量关节热图中所有点的融合响应值构成融合热图矩阵 $\boldsymbol{M}_{j,c}^{\text{fuse}}$。

通过阈值判断，估计性能优秀的关节热图可以以避免被参考视图中的噪声所干扰。对于单视图中质量欠佳的关节估计，融合之后会被其较低的融合权重所稀释，从而在融合热图中留下较小的响应值，而具有高权重的关节热图会做出更大的贡献。并且，高质量的关节热图之间往往具有较高的一致性，通常会重叠于关节附近，融合输出最大响应值，进而通过 softmax 操作排除较大的误差估计，获得各视图下更为精确的关节点坐标，得到二维人体姿态 $\boldsymbol{X}$。

2.3 二维姿态细化模块

为了实现更为精确的多视图人体姿态估计，在构建多视图间关联性的同时，还需要探索单视图中人体关节的内在相关性。人体骨架具有固定的拓扑结构，每个关节点的位置估计可以从相邻或远处的关节点信息中受益，通过拓扑约束和运动学约束，修

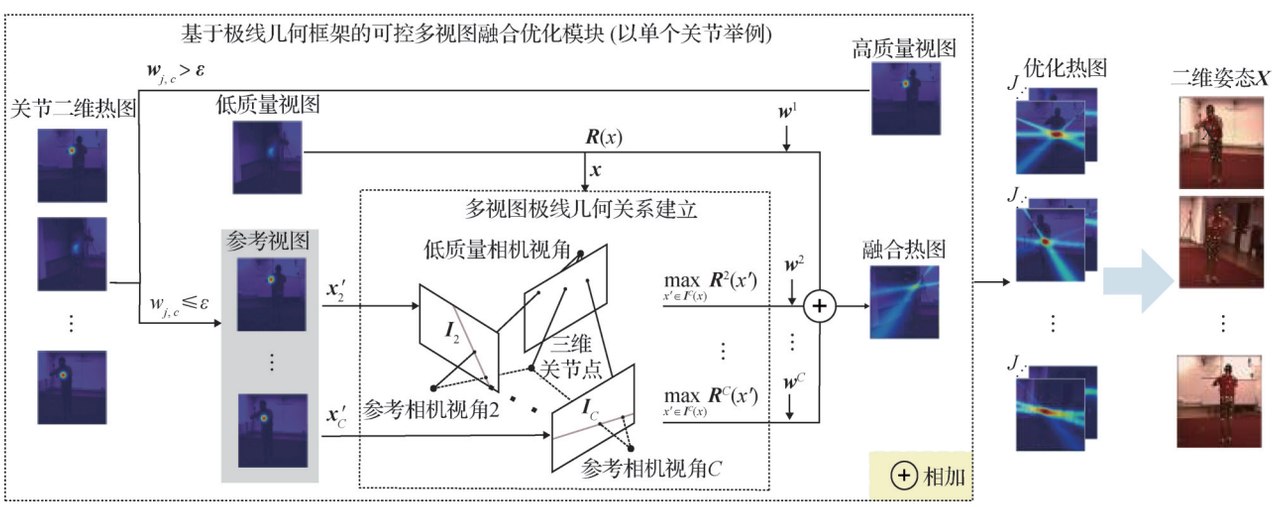


图2 基于极线几何框架的多视图可控融合优化模块

Fig. 2 Multi-view controlled fusion optimization module based on epipolar geometry framework

正错误的估计结果。因此，为了能够更好地挖掘单视图中的人体关节相关性，本文构建了一个图卷积和注意力机制联合的二维姿态细化模块，如图3所示。

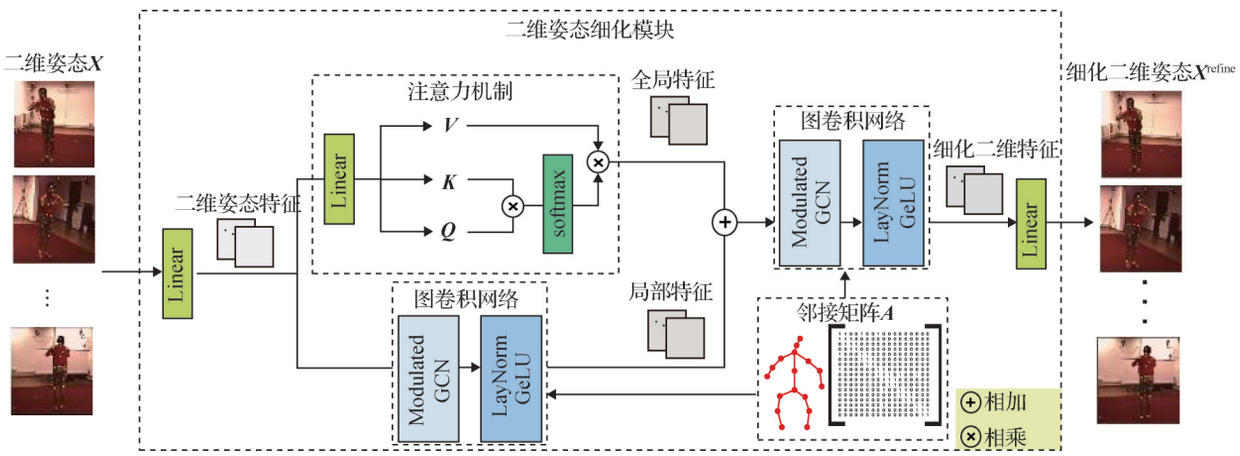


图3　二维姿态细化模块

Fig. 3　2D pose refinement module

人体的拓扑结构由关节点的位置和相连表征，这种表征信息可以通过点和边的图形结构展现，而图卷积能够很好地学习这种图特征，进而通过聚合相邻关节点之间的特征学习人体的局部信息。人体运动姿态的变化正是在这种局部关联性的基础上所呈现出的全局表征，而注意力机制可以帮助网络在处理全部输入特征时关注于与输出信息相关的部分，从而可以采用注意力机制构建整个人体关节点之间的联系，有效学习人体的全局信息。因此，基于图卷积和注意力联合的二维姿态细化网络能够通过从局部到全局的相关性学习，细化各个关节的估计结果，获取更为精准的二维姿态。

首先，该模块将描述二维姿态的关节点坐标向量通过线性层和归一化操作转换为二维姿态特征。具体为

$$\boldsymbol{F}^{\text{pose}} = LN(Liner(X)) \tag{4}$$

式中，$\boldsymbol{F}^{\text{pose}}$为二维姿态特征，$LN(\cdot)$为层归一化，$Liner(\cdot)$为线性层。

随后，融入人体骨架邻接矩阵，构建图卷积网络。所有关节点的二维特征构成节点集，描述人体结构信息的关节之间的邻接矩阵为边集，二者共同表示为图卷积所输入的无向图。因此，图卷积网络能够从二维姿态特征中提取所蕴含的人体局部信息，具体为

$$\boldsymbol{F}^{\text{GCN}} = GeLU(MGCN(F^{\text{pose}}, \boldsymbol{A})) \tag{5}$$

式中，$\boldsymbol{A}$为人体骨架的标准化邻接矩阵，$MGCN(\cdot)$为调制图卷积（Zou和Tang，2021），$GeLU(\cdot)$为GeLU（Gaussian error linear unit）激活函数，$\boldsymbol{F}^{\text{GCN}}$为以骨架拓扑信息为约束的人体二维姿态局部特征。

注意力机制将输入的二维特征映射为查询（Q）、键（K）、值（V），利用QKV来构建特征之间相关性，从而以相似度获取注意力权重，并通过权重与特征元素的加权平均和求得人体二维姿态的全局特征。因此，注意力机制可以从二维姿态特征的相关性构建中获取人体的全局信息，具体为

$$\boldsymbol{F}^{\text{Att}} = Att(\boldsymbol{F}^{\text{pose}} + E) \tag{6}$$

式中，E为人体关节坐标的位置嵌入，$Att(\cdot)$为多头注意力机制（Vaswani等，2017），$\boldsymbol{F}^{\text{Att}}$为注意力机制学习的二维姿态全局特征。

最后，将局部特征和全局特征馈入图卷积网络，并通过层归一化和线性处理，提升估计精度，得到细化后的二维姿态。具体为

$$\boldsymbol{F}^{\text{final}} = LN(MGCN((\boldsymbol{F}^{\text{GCN}} + \boldsymbol{F}^{\text{Att}}), \boldsymbol{A})) \tag{7}$$

$$\boldsymbol{X}^{\text{refine}} = Liner(\boldsymbol{F}^{\text{final}}) \tag{8}$$

式中，$\boldsymbol{F}^{\text{final}}$为最终的二维特征，$\boldsymbol{X}^{\text{refine}}$为细化后的二维姿态。

2.4　结构化三角剖分

采用三角剖分技术，可以利用多个视图中的二维关节点坐标和相应相机投影矩阵计算出最终的三维关节点坐标，实现多视图二维姿态到三维姿态的转化。其推理过程为

$$\min_{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_J}\sum_{q=1}^{J}\sum_{c=1}^{C}p_{q,c}\left\|\boldsymbol{P}_c\boldsymbol{y}_q-\boldsymbol{x}_{q,c}\right\| \tag{9}$$

式中，$\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_J$为待定的三维关节点。$p_{q,c}$为视图$\boldsymbol{c}$下的$q$关节的置信度。$\boldsymbol{P}_c$为相机视图$\boldsymbol{c}$的投影矩阵。$\boldsymbol{x}_{q,c}$为视图$\boldsymbol{c}$下第$q$个关节的二维坐标。

从式(9)可以看出，这种传统的三角剖分过程仅考量了单个关节点的推理，忽视了人体的固有结构和人体骨长的先验知识。为此，如图4所示，本文通过引入结构化三角剖分(Chen等，2022)来替代传统的代数三角剖分，提取人体的骨骼长度作为约束条件，在推理过程中加入人体结构的先验知识。约束条件为

$$\left\|\boldsymbol{y}_i-\boldsymbol{y}_q\right\|=L_B \tag{10}$$

式中，L_B为相邻关节i和q之间的骨骼B的长度，即骨长先验约束。在此限制下，结构化三角剖分的目标为使用已知的所有骨骼长度来最小化加权平方重投影误差。此时，通过转换矩阵$\boldsymbol{G}_{iq}\in\mathbf{R}^{3\times3}$，两个相邻关节点$i$和$q$的三维坐标信息可以转换为包含人体结构信息的骨向量$\boldsymbol{b}_{iq}$。优化后的推理过程为

$$\boldsymbol{b}=\boldsymbol{G}\boldsymbol{y} \tag{11}$$

$$\min_{\boldsymbol{b}} f(\boldsymbol{b})=\frac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{b}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{b}+d \tag{12}$$

$$\text{s.t.}\quad \left\|\boldsymbol{b}_s\right\|^2=\boldsymbol{L}_s^2 \tag{13}$$

式中，$\boldsymbol{G}\in\mathbf{R}^{3J\times3J}$由$\boldsymbol{G}_{iq}$构成，指示了关节点之间的指向，从而将所有的关节点的三维坐标转化为包含人体信息的骨向量集合$\boldsymbol{b}$，$\boldsymbol{A}\in\mathbf{R}^{3J\times3J}$是半正定矩阵，$\boldsymbol{\beta}\in\mathbf{R}^{3J}$和$d\in\mathbf{R}$是常数，$s=1,\cdots,J-1$为骨骼编号，$\boldsymbol{L}$为所有骨骼长度的集合。

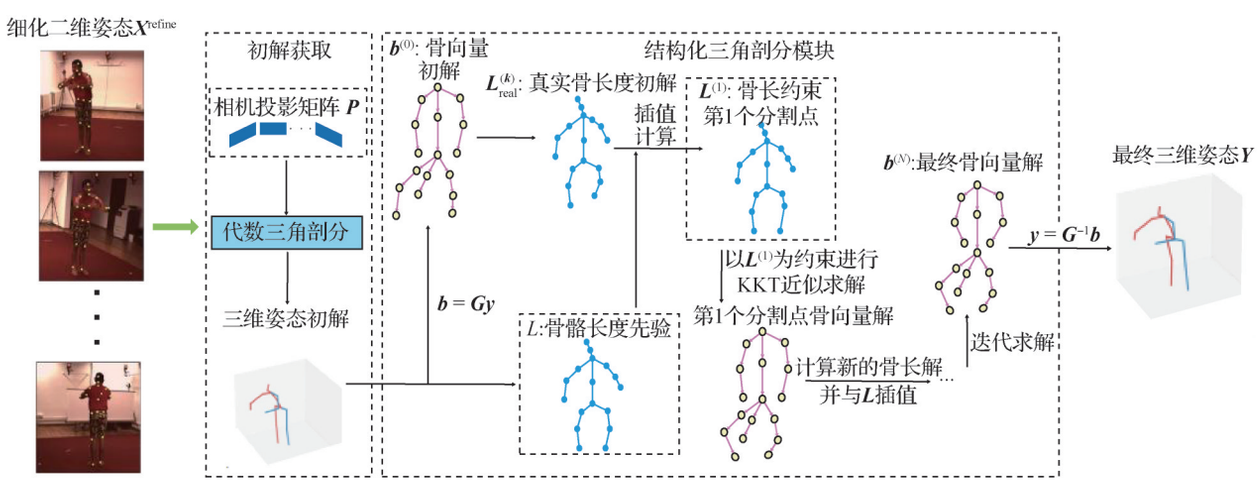


图4　结构化三角剖分模块

Fig. 4　Structural triangulation module

Karush-Kuhn-Tucker(KKT)条件给出的分析搜索范围，在适当的近似中实现了线性化。但是，当初解的骨骼长度集合$\boldsymbol{L}_{\text{real}}^{(0)}$与引入的骨长先验约束$L$相差过大时，原本假设的近似将不成立。因此，需要使用步长约束SCA(step constraint algorithm)算法优化求解过程，通过在$\boldsymbol{L}^{(0)}$与$\boldsymbol{L}$之间进行插值，由结构化三角剖分从解$\boldsymbol{b}^{(0)}$开始递推，直至求得最终解$\boldsymbol{b}^{(N)}$。步长约束算法为

$$\boldsymbol{L}^{(k)}=\frac{a_k}{a_{k-1}}\boldsymbol{L}_{\text{real}}^{(k-1)}+\frac{a_{k-1}-a_k}{a_{k-1}}\boldsymbol{L} \tag{14}$$

式中，$k\in N$为当前步，N为总步长。a_k取自递减级数$\{a_i\}_{i=0}^{N}$，其中，$a_0=1$，$a_N=0$。$\boldsymbol{L}_{\text{real}}^{(k-1)}$为由$\boldsymbol{b}^{(k-1)}$直接计算得到的骨骼长度集合，而$\boldsymbol{L}^{(k)}$则为第$k$步推导出的分割点。新分割点可以作为一个新的约束条件进行结构化三角剖分求解，由$\boldsymbol{b}^{(k-1)}$求得新分割点对应的骨向量解$\boldsymbol{b}^{(k)}$，进一步获得相应的骨骼长度集合$\boldsymbol{L}_{\text{real}}^{(k)}$以计算下一个分割点$\boldsymbol{L}^{(k+1)}$。最终，经过$N$步递推获得优化解$\boldsymbol{b}^{(N)}$。并根据$\boldsymbol{y}=\boldsymbol{G}^{-1}\boldsymbol{b}$逆推得到三维关节点坐标，生成三维人体姿态。

综上所述，引入人体骨架结构和关节点间的骨长先验知识，细化后的二维人体姿态经过结构化三角剖分能够实现更为准确的三维人体姿态估计。

3　实　验

3.1　数据集和评估标准

Human3.6M数据集(Ionescu等，2014)提供了

一个具有4个摄像头的拍摄场景，共收录360万幅图像，由11名不同的受试者在进行15种不同行为活动时获取所得。本文采用S1，S5，S6，S7，S8这5名对象的相关图像作为训练集，S9和S11两名对象的相关图像作为测试集。

Total Capture数据集（Trumble等，2017）中的图像由8个不同方位的摄像头所提供。该数据集共有5名受试者参与拍摄，其中包含4名男性以及1名女性。5名受试者共进行了Freestyle、ROM、Walking、Acting和Running等5种活动，且每组动作重复3次，数字标号i表示第i次的重复动作。按照惯例，本文的训练集由对象S1，S2，S3的活动ROM1 2 3，Freestyle1 2，Walking1 3，Acting1 2和Running1图像集合组成，测试集则由对象S1，S2，S3，S4，S5的活动Freestyle3，Acting3，Walking2构成。

Occlusion-Person数据集为合成数据集（Zhang等，2021），设计了9个不同的环境场景，由身着13种不同服装的人体模型在CMU Motion Capture数据集姿态驱动下生成不同的行为姿态，同时这些人体模型会被沙发、桌子等东西遮挡住部分关节。整个场景由相隔45°放置的8个摄像头所拍摄。该数据集提供了15个关节点坐标作为三维姿态真值，并且所有的关节都提供了遮挡标签。

本文采用平均关节误差MPJPE作为实验的评估指标，由估计的三维关节坐标与真值关节坐标之间的欧氏距离表示。MPJPE结果能够评价三维人体姿态的估计质量，从而能够更为直观地展现、比较各种方法的性能。

3.2 实验设置

本文采用预训练的Simple Baseline（Xiao等，2018）作为二维检测器以提取二维热图。其中，关节质量判断阈值$\varepsilon = 0.99$，二维姿态细化模块的模块层数$N = 3$。在网络主体训练中，本文采用Adam优化器进行训练，迭代次数30，初始学习率为0.000 1，在第10个和15个迭代时乘以0.1的比例进行学习率衰减。

3.3 实验结果

3.3.1 Human3.6M上的实验结果

为了展现本文方法在三维人体姿态估计中的优越性能，在公共数据集Human3.6M与几种先进方法（state-of-the-art，SOTA）进行实验比较，结果如表1所示。

从表1可以看出，本文方法在基线Adafuse指标MPJPE为19.5 mm的基础上，实现了2.4 mm的误差降低，对比数据集中出现的所有姿态，相对于基线系统有1.4～4.1 mm程度的性能提升。与使用了参数化人体模型和体积三角剖分的LMT（learnable human mesh triangulation）方法相比，本文方法将误差降低了0.5 mm，且计算成本相对更低，更具实际意义。与没有使用参数化人体模型的其他各类方法相比，本文方法都能有至少10%的提升。并且由于本文网络引入了丰富的人体先验知识，很好地构建了人体关节点间的内在联系，使得本文方法在结合

表1 CFJCNet与先进方法在Human3.6M数据集上的性能比较

Table 1 Comparisons of the CFJCNet with the SOTA methods on the Human3.6M dataset

/mm

方法	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	平均
Adafuse（Zhang等，2021）	17.8	19.5	17.6	20.7	19.3	16.8	18.9	20.2	25.7	20.1	19.2	20.5	17.2	20.5	17.3	19.5
CrossView（Qiu等，2019）	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	26.9	31.0	25.6	25.0	28.1	24.4	26.2
Multi-view PPT（Ma等，2022）	21.8	26.5	21.0	22.4	23.7	23.1	23.2	27.9	30.7	24.6	26.7	23.3	21.2	25.3	22.6	24.4
HT（Wan等，2023）	19.5	20.9	19.5	18.3	21.1	20.0	17.9	21.3	23.9	30.1	21.6	19.9	18.9	22.8	19.5	21.1
ST（Chen等，2022）	17.4	19.7	15.6	**17.5**	19.6	18.8	**17.0**	20.2	**21.9**	26.7	19.3	18.9	17.9	20.6	18.7	19.4
Epipolar transformer（He等，2020）	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	19.0
LMT（Chun等，2023）	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	17.6
CFJCNet（本文）	**16.4**	**17.0**	**13.5**	**18.7**	**16.9**	**14.9**	17.5	**17.9**	22.7	**16.8**	**17.2**	**18.6**	**14.9**	**17.6**	**15.5**	**17.1**

注：加粗、下划线字体分别表示各列最优、次优结果，“–”表示原作者未提供实验结果。

多视图融合优势的情况下能进一步优化人体姿态估计结果。尤其是针对 Disc、Eat、Phone、Purch、SitD、Walk 等行为活动，由于 CFJCNet 能够修正手腕、膝盖等部位的错误估计，并能进一步减少遮挡带来的影响，因此，网络性能存在明显优势。但是在 Sit、Pose 和 Greet 等姿态估计上，网络取得了次优或与次优相当的结果，原因在于这些姿态中的部分关节在 4 个视图中的热图估计质量均不理想，多视图融合的优势无法发挥，反而会导致噪声的相互干扰，从而引起估计质量的下降。从数据集中的 15 种姿态对比分析，CFJCNet 估计效果最为稳定。

3.3.2　Total Capture 上的结果

为了证明本文方法的普适性，基于关节点相关性的三维人体姿态估计模型在公共数据集 Totalcapture 上进行了训练与测试，性能比较结果如表 2 所示。从表中可以发现，CFJCNet 在受试对象的各种行为活动测试中均取得了具有竞争力的结果，获得了最优或次优估计。

在实验结果中，虽然 CrossView 算法在对象 S4、S5 的 FS3 活动中性能最好，但对于其余对象的各种行为估计效果均较差。同时，AdaDeepFuse 算法和 GeoFuse 算法都需要使用额外传感器 IMU（inertial measurement unit）获得辅助信息，而本文方法在缺少 IMU 信息的条件下依然能够较 AdaDeepFuse 算法平均提高 3.8 mm，相对 Baseline 平均提高 2.6%。实验结果表明，CFJCNet 在不同数据集上的估计性能均表现优秀。

表 2　CFJCNet 与先进方法在 Total Capture 数据集上的性能比较

Table 2　Comparisons of the CFJCNet with SOTA methods on the Total Capture dataset

/mm

方法	IMUs	受试者(S1, S2, S3)			受试者(S4, S5)			平均
		W2	A2	FS3	W2	A2	FS3	
Adafuse（Zhang 等，2021）	–	**7.2**	**10.8**	18.5	22.8	26.6	42.9	19.2
CrossView（Qiu 等，2019）	–	19.0	28.0	21.0	32.0	54.0	**33.0**	31.0
LWCDR（Remelli 等，2020）	–	10.6	16.3	30.4	27.0	34.2	65.0	27.5
MTF-Transformer（Shuai 等，2023）	–	13.9	18.1	29.2	29.2	35.6	49.5	27.0
GeoFuse（Zhang 等，2020）	√	14.3	17.5	25.9	23.9	27.8	49.3	24.6
AdaDeepFuse（ Bao等，2023）	√	10.2	13.7	26.3	**21.7**	26.8	49.2	22.5
CFJCNet（本文）	–	8.7	12.2	**18.0**	22.5	**26.2**	35.8	**18.7**

注：加粗、下划线字体分别表示各列最优、次优结果，“√”表示采用，“–”表示未采用。

3.3.3　Occlusion-Person 上的结果

为了更好地展现 CFJCNet 面对被遮挡关节时的估计性能，本节在一个具有遮挡标签的合成数据集中进行了实验，结果如表 3 所示。从表中结果可以明显看出，本文方法可以很好地修复因为遮挡等问题导致的错误估计，从而取得良好的估计结果；而对于受遮挡较少的关节，该方法能通过二维姿态的细化修正，取得估计效果的进一步提升。与基线方法

表 3　CFJCNet 与先进方法在 Occlusion-Person 数据集上的性能比较

Table 3　Comparisons of the CFJCNet with the SOTA methods on the Occlusion-Person dataset

/mm

方法		Root	Belly	Neck	Hip	Knee	Ankle	Shlder	Elbow	Wrist	平均
	遮挡率	14.3%	13.7%	7.6%	23.0%	25.0%	23.5%	16.8%	25.3%	21.7%	
Adafuse（Zhang等，2021）		7.2	10.6	11.6	11.7	13.8	15.7	14.2	**9.9**	14.4	12.6
CFJCNet（本文）		**6.7**	**10.1**	**9.0**	**9.2**	**12.0**	**12.2**	**10.3**	10.5	**12.1**	**10.2**

注：加粗字体表示各列最优结果。

相比，对于绝大部分关节都可以取得最优估计，在基线基础上性能提升了19%。

3.3.4　光照影响实验结果

为了探究CFJCNet网络在不同光照强度下的性能，本节将Human3.6M的测试集进行了光亮度调整，以此模拟实际的光照变化，并将原始数据训练的CFJCNet和基线网络在该测试集上进行性能测试，结果如表4所示。首先，单独将某个方向视图的光亮度进行了50%幅度的削减，与Baseline相比，CFJCNet的姿态估计性能MPJPE改进了2.87 mm；进一步将光亮度减少80%，此时CFJCNet性能较Baseline的性能改进了3.28 mm。因此，在不同光照强度影响下，本文方法相比于基线方法具有更好的适应性。

表4　在Human3.6M数据集上的光照对比实验

Table 4　Light comparison experiments on Human3.6M dataset

方法	MPJPE/mm	
	光亮度削减50%	光亮度削减80%
Baseline	23.38	32.55
CFJCNet	20.51	29.27

3.4　消融实验

为了验证各个模块以及结构化三角剖分的引入在CFJCNet中的有效性，本节对这些不同模块进行了性能评估，如表5所示。

首先，实验1在Baseline的多视图融合过程中嵌入了基于极线几何框架的多视图可控融合优化模块。该模块是一个轻量模块，其对权重学习网络得到的融合权重进行了简单处理、筛选，从而区分了各关节在不同视图下的热图估计质量，进而避免高质量关节因融合导致的噪声引入，相较于Baseline实现了0.18 mm的误差减少。在实验2中，添加了二维姿态细化模块，以此构建人体骨架关节之间的内在联系，作为先验知识修正了部分错误姿态，使得误差从19.54 mm下降到19.37 mm。实验3则引入了结构化三角剖分以改进Baseline中的代数三角剖分，使得网络能够采用人体的骨长信息约束部分不合理的姿态，在Baseline的基础上得到了2.19 mm的性能提升。随后，为了验证多个模块联合使用时的有效性，实验4在结构化三角剖分的基础上添加了可控多视图融合优化模块，系统估计误差进一步降低了0.12 mm。实验5则在结构化三角剖分的基础上添加了二维姿态细化模块，成功使得误差从17.35 mm下降到17.1 mm。最后，联合3个模块的CFJCNet实现了17.09 mm的最优估计效果。

表5　在Human3.6M数据集上的消融实验

Table 5　Ablation study on the Human3.6M dataset

方法	可控多视图融合优化模块	二维姿态细化模块	结构化三角剖分模块	MPJPE/mm
Baseline	–	–	–	19.54
实验1	√	–	–	19.36
实验2	–	√	–	19.37
实验3	–	–	√	17.35
实验4	√	–	√	17.23
实验5	–	√	√	17.10
CFJCNet	√	√	√	17.09

注：“√”表示采用，“–”表示未采用。

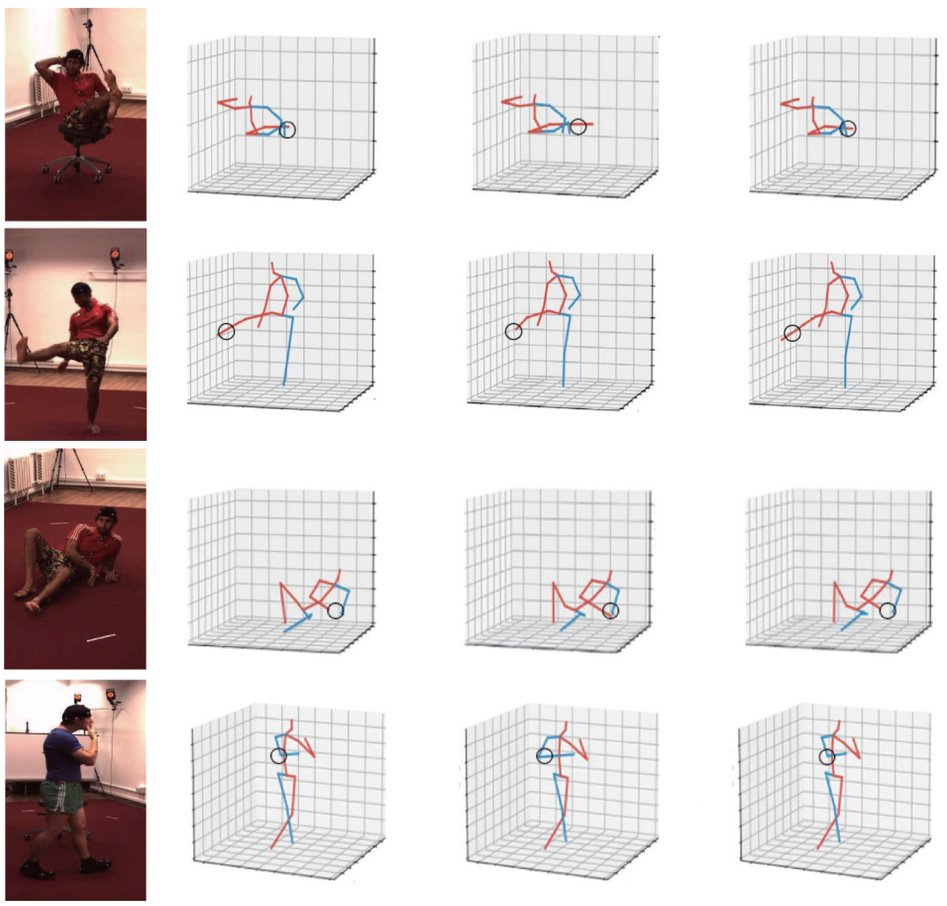
3.5　可视化分析

为了更直观地展现本文方法的估计效果，在Human3.6M数据集和Total Capture数据集上进行了可视化结果分析。图5和图6分别展示了两个数据集中部分姿态估计的可视化效果。

由图5前两个示例可见，基线方法估计的人体姿态相较于真实姿态在右小腿上存在明显的错误，而本文方法基于人体骨骼长度的结构约束，通过细化人体姿态，实现了右脚的正确估计。在第3、4个示例中，CFJCNet修正了Baseline对两只手腕和左手肘估计的位置偏差和角度偏差，实现了更为精准的估计。由此可见，在数据集human3.6M上，本文方法可以纠正错误估计，从而取得与真实3D姿态相近的估计效果。同样，图6展现了CFJCNet在数据集Total Capture上对错误估计姿态的修正能力。在两个数据集上的定性实验结果皆表明了CFJCNet能够利用先验约束修正不合理的错误姿态估计。

4　结　论

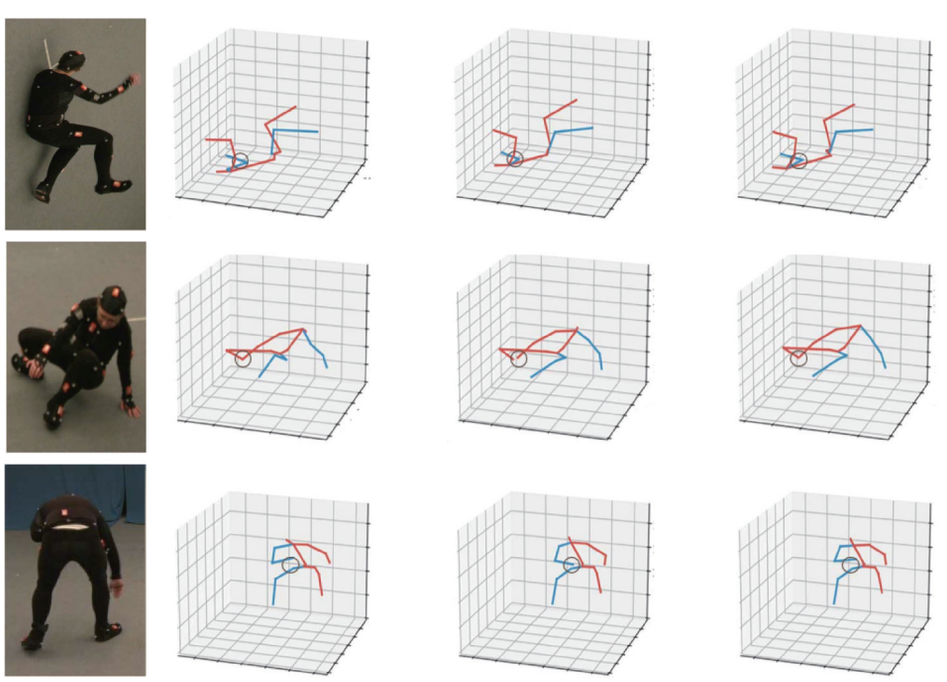
本文提出了一个联合多视图可控融合和关节相关性的三维人体姿态估计算法，在嵌入多视图可控融合优化模块和2D姿态细化模块基础上，引入了结构化三角剖分，获得了优秀的3D人体姿态估计性



(a) 图像　(b) 真实姿态　(c) Adafuse　(d) CFJCNet

图 5　Human3. 6M 上的定性结果

Fig. 5　Qualitative results on the Human3. 6M dataset((a)images;(b)ground truth;(c)Adafuse;(d)CFJCNet)



(a) 图像　(b) 真实姿态　(c) Adafuse　(d) CFJCNet

图 6　Total Capture 上的定性结果

Fig. 6　Qualitative results on the Total Capture dataset((a)images;(b)ground truth;(c)Adafuse;(d)CFJCNet)

能。在多视图融合模块中，通过权重信息进行质量评估，从而使得主视图中的低质量关节估计能够从参考视图中获得有益的信息补充，同时，避免了高质量关节热图因不良热图的融合而导致的二维姿态估计错误。而2D关节细化模块则能利用图卷积和注意力机制的联合优势，帮助网络更好地学习人体关节点的内在联系以细化2D姿态估计。最后，引入结构化三角剖分，利用骨长向量约束，实现了3D姿态估计的合理有效。在公共数据集Human3.6M和Total Capture与合成数据集Occlusion-Person上的实验结果表明，相较于其他多种先进方法，CFJCNet在评估指标上存在明显优势，证明了CFJCNet的优越性和泛化性。

虽然整体实验结果较为优秀，但如果部分关节在多个视图中均存在严重遮挡时，多视图融合的优势无法充分发挥。因此，在后续的研究工作中将考虑研究时间信息在获取多帧相关特征中的作用，进而克服因为遮挡导致的特征信息缺失的问题，实现更有效的三维人体姿态估计。

参考文献（References）

Bao Y M, Zhao X and Qian D H. 2023. FusePose: IMU-vision sensor fusion in kinematic space for parametric human pose estimation. IEEE Transactions on Multimedia, 25: 7736-7746 [DOI: 10.1109/TMM.2022.3227472]

Chen H Y, He J Y, Xiang W M, Cheng Z Q, Liu W, Liu H B, Luo B, Geng Y F and Xie X S. 2023a. HDFormer: high-order directed transformer for 3D human pose estimation [EB/OL]. [2024-01-15]. https://arxiv.org/pdf/2302.01825.pdf

Chen Y X, Gu R S, Huang O H and Jia G Y. 2023b. VTP: volumetric transformer for multi-view multi-person 3D pose estimation. Applied Intelligence, 53 (22): 26568-26579 [DOI: 10.1007/s10489-023-04805-z]

Chen Z, Zhao X and Wan X Y. 2022. Structural triangulation: a closed-form solution to constrained 3D human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 695-711 [DOI: 10.1007/978-3-031-20065-6_40]

Chun S, Park S and Chang J Y. 2023. Learnable human mesh triangulation for 3D human pose and shape estimation//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2849-2858 [DOI: 10.1109/WACV56688.2023.00287]

Du X X, Vasudevan R and Johnson-Roberson M. 2019. Bio-LSTM: a biomechanically inspired recurrent neural network for 3-D pedestrian pose and gait prediction. IEEE Robotics and Automation Letters, 4(2): 1501-1508 [DOI: 10.1109/LRA.2019.2895266]

Hassan M T and Ben Hamza A. 2023. Regular splitting graph network for 3D human pose estimation. IEEE Transactions on Image Processing, 32: 4212-4222 [DOI: 10.1109/TIP.2023.3275914]

He Y H, Yan R, Fragkiadaki K and Yu S I. 2020. Epipolar transformers//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7776-7785 [DOI: 10.1109/CVPR42600.2020.00780]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Ionescu C, Papava D, Olaru V and Sminchisescu C. 2014. Human 3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1325-1339 [DOI: 10.1109/TPAMI.2013.248]

Iskakov K, Burkov E, Lempitsky V and Malkov Y. 2019. Learnable triangulation of human pose//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7717-7726 [DOI: 10.1109/ICCV.2019.00781]

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

Kim H W, Lee G H, Oh M S and Lee S W. 2023. Cross-view self-fusion for self-supervised 3D human pose estimation in the wild//Proceedings of the 16th Asian Conference on Computer Vision. Macau, China: Springer: 193-210 [DOI: 10.1007/978-3-031-26319-4_12]

Li W H, Liu H, Tang H, Wang P C and van Gool L. 2022. MHFormer: multi-hypothesis transformer for 3D human pose estimation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13137-13146 [DOI: 10.1109/CVPR52688.2022.01280]

Lu J X, Wan H, Li P Y, Zhao X B, Ma N and Gao Y. 2023. Exploring high-order spatio-temporal correlations from skeleton for person re-identification. IEEE Transactions on Image Processing, 32: 949-963 [DOI: 10.1109/TIP.2023.3236144]

Ma H Y, Chen L J, Kong D Y, Wang Z, Liu X W, Tang H, Yan X Y, Xie Y S, Lin S Y and Xie X H. 2021. TransFusion: cross-view fusion with transformer for 3D human pose estimation [EB/OL]. [2024-01-15]. https://arxiv.org/pdf/2110.09554.pdf

Ma H Y, Wang Z, Chen Y F, Kong D Y, Chen L J, Liu X W, Yan X Y, Tang H and Xie X H. 2022. PPT: token-pruned pose transformer for monocular and multi-view human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 424-442 [DOI: 10.1007/978-3-031-

20065-6_25］

Ma M，Li Y B，Wu X Q，Gao J F and Pan H P. 2020. Human pose tracking based on multi-feature fusion in videos. Journal of Image and Graphics，25(7)：1459-1472（马淼，李贻斌，武宪青，高金凤，潘海鹏. 2020. 视频中多特征融合人体姿态跟踪. 中国图象图形学报，25(7)：1459-1472）［DOI：10.11834/jig.190494］

Minar M R and Ahn H. 2021. CloTH-VTON：clothing three-dimensional reconstruction for hybrid image-based virtual try-on//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto，Japan：Springer：154-172［DOI：10.1007/978-3-030-69544-6_10］

Pascual-Hernández D，de Frutos N O，Mora-Jiménez I and Cañas-Plaza J M. 2022. Efficient 3D human pose estimation from RGBD sensors. Displays，74：#102225［DOI：10.1016/j.displa.2022.102225］

Qiu H B，Wang C Y，Wang J D，Wang N Y and Zeng W J. 2019. Cross view fusion for 3D human pose estimation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea（South）：IEEE：4341-4350［DOI：10.1109/ICCV.2019.00444］

Remelli E，Han S C，Honari S，Fua P and Wang R. 2020. Lightweight multi-view 3D pose estimation through camera-disentangled representation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：6039-6048［DOI：10.1109/CVPR42600.2020.00608］

Shuai H，Wu L L and Liu Q S. 2023. Adaptive multi-view and temporal fusing transformer for 3D human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence，45(4)：4122-4135［DOI：10.1109/TPAMI.2022.3188716］

Tang Z H，Li J，Hao Y B and Hong R C. 2023. MLP-JCG：multi-layer perceptron with joint-coordinate gating for efficient 3D human pose estimation. IEEE Transactions on Multimedia，25：8712-8724［DOI：10.1109/TMM.2023.3240455］

Trumble M，Gilbert A，Malleson C，Hilton A and Collomosse J. 2017. Total capture：3D human pose estimation fusing video and inertial sensors//Proceedings of the 28th British Machine Vision Conference. London：UK，BMVA：1-13［DOI：10.5244/c.31.14］

Tu H Y，Wang C Y and Zeng W J. 2020. VoxelPose：towards multi-camera 3D human pose estimation in wild environment//Proceedings of the 16th European Conference on Computer Vision—ECCV 2020. Glasgow，UK：Springer：197-212［DOI：10.1007/978-3-030-58452-8_12］

Vaswani A，Shazeer N，Parmar N，Uszkoreit J，Jones L，Gomez A N，Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach，USA：Curran Associates Inc.：6000-6010

Wan X Y，Chen Z and Zhao X. 2023. View consistency aware holistic triangulation for 3D human pose estimation［EB/OL］.［2024-01-15］. https://arxiv.org/pdf/2302.11301.pdf

Wu W，Zhou D S，Zhang Q，Dong J and Wei X P. 2022. High-order local connection network for 3D human pose estimation based on GCN. Applied Intelligence，52(13)：15690-15702［DOI：10.1007/s10489-022-03312-x］

Xiao B，Wu H P and Wei Y C. 2018. Simple baselines for human pose estimation and tracking//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：472-487［DOI：10.1007/978-3-030-01231-1_29］

Yan J J，Zheng W M，Xin M H and Qiu W. 2013. Bimodal emotion recognition based on body gesture and facial expression. Journal of Image and Graphics，18(9)：1101-1106（闫静杰，郑文明，辛明海，邱伟. 2013. 表情和姿态的双模态情感识别. 中国图象图形学报，18(9)：1101-1106）［DOI：10.11834/jig.20130906］

Zhang Z，Wang C Y，Qin W H and Zeng W J. 2020. Fusing wearable IMUs with multi-view images for human pose estimation：a geometric approach//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：2197-2206［DOI：10.1109/CVPR42600.2020.00227］

Zhang Z，Wang C Y，Qiu W C，Qin W H and Zeng W J. 2021. Adafuse：adaptive multiview fusion for accurate human pose estimation in the wild. International Journal of Computer Vision，129(3)：703-718［DOI：10.1007/s11263-020-01398-9］

Zou Z M and Tang W. 2021. Modulated graph convolutional network for 3D human pose estimation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada，IEEE：11457-11467［DOI：10.1109/ICCV48922.2021.01128］

作者简介

董婧，女，副教授，主要研究方向为多模态感知与学习。
E-mail：dongjing@dlu.edu.cn
方小勇，通信作者，男，副教授，主要研究方向为人工智能、数据采集与分析。E-mail：xyfang@hnit.edu.cn
张鸿儒，男，硕士研究生，主要研究方向为计算机视觉。
E-mail：17843125429@163.com
周东生，男，教授，主要研究方向为计算机图形学、智能计算和人机交互。E-mail：zhouds@dlu.edu.cn
杨鑫，男，教授，主要研究方向为计算机图形学与视觉、类脑计算和人机交互。E-mail：xinyang@dlut.edu.cn
张强，男，教授，主要研究方向为生物启发计算及应用。
E-mail：zhangq@dlu.edu.cn
魏小鹏，男，教授，主要研究方向为大数据处理及分析、智能CAD和计算机图形学。E-mail：xpwei@dlut.edu.cn