基于结构引导的人体姿态估计框架

涂 浚 武港山 王利民
（南京大学计算机软件新技术全国重点实验室 南京 210023）

摘 要 近年来，2D人体姿态估计作为计算机视觉中的基础任务，广泛应用于行为识别、人机交互等领域。尽管基于深度学习的姿态估计方法取得了显著进展，但在多人拥挤、遮挡复杂及低分辨率等实际场景下，现有方法往往面临结构信息利用不足、优化路径粗糙等问题，导致模型的姿态结构建模能力有限、泛化鲁棒性不强。为此，本文提出了一种基于结构引导的多步优化框架，将姿态估计过程建模为从初始粗略预测逐步优化至结构合理目标的多阶段演化路径。该框架通过显式构造渐变图序列，引导网络在每一步预测中持续融合结构先验信息，并在训练与推理阶段保持结构引导路径的一致性，从而有效提升模型的结构建模能力与预测稳定性。系统性消融实验表明，所提出的渐变图序列和路径一致性设计对于提升关键点定位精度和结构约束能力具有显著效果；参数敏感性分析进一步验证了插值步数与调度策略等可调参数对模型性能的影响。在COCO和CrowdPose等主流数据集上的实验结果表明，本文方法在HRNet-W48骨干网络下、CrowdPose验证集上取得77.6 mAP，超过当前先进方法TransPose-H（76.3 mAP）；在COCO验证集256×192分辨率和检测框设定下同样实现了75.6 mAP，与主流Transformer方法持平或更优，验证了本方法在多种复杂场景下的有效性与通用性。

关键词 2D人体姿态估计；结构引导；结构建模；渐变图序列；多步优化
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00760

Structure-Guided Framework for Human Pose Estimation

TU Jun WU Gang-Shan WANG Li-Min
（State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023）

Abstract Recent advances in deep learning have significantly improved 2D human pose estimation, a core task in computer vision with broad applications in action recognition, human-computer interaction, rehabilitation, virtual reality, and intelligent surveillance. However, existing approaches still struggle in crowded scenes, heavy occlusion, and low-resolution conditions due to insufficient utilization of structural priors and coarse optimization trajectories. These limitations often hinder the model's ability to maintain globally coherent pose structures, resulting in reduced robustness and suboptimal generalization.

To address these challenges, this paper proposes a structure-guided multi-step optimization framework that models pose estimation as a progressive evolution from an initial coarse prediction toward a structurally plausible target. The key idea is to construct an explicit gradient map sequence, which encodes a step-wise structural evolution path and guides the network at each stage to incorporate human body priors. By enforcing consistency between training and inference trajectories, the framework enhances both global structural modeling and prediction stability.

The method consists of three main components (1) a fixed single-person pose estimation

收稿日期：2025-06-06；在线发布日期：2025-12-09. 本课题得到科技创新2030——“新一代人工智能”重大项目（2022ZD0160900）、江苏省自然科学基金攀登项目（BK20250009）资助。 **涂 浚**，博士研究生，主要研究领域为人体姿态估计、姿态跟踪、步态识别。E-mail：tujun@smail.nju.edu.cn。 **武港山**，博士，博士生导师，中国计算机学会（CCF）会员，主要研究领域为媒体内容分析、多媒体信息检索等。 **王利民**（通信作者），博士，教授，博士生导师，中国计算机学会（CCF）会员，主要研究领域为视频理解和动作识别等。E-mail：lmwang@nju.edu.cn。

backbone that extracts image features and provides the initial heatmap prediction; (2) a gradient map generator that constructs a multi-stage structural evolution sequence via cosine-based interpolation between a predefined canonical pose and the backbone's predicted pose; and (3) a structure-guided UNet that performs multi-step refinement, where each step receives the image features, the corresponding gradient map, and the time-step index as conditional inputs. Unlike diffusion-based denoising models, the proposed approach does not rely on Gaussian noise modeling. Instead, it focuses on structural evolution driven by explicit geometric priors.

Extensive ablation studies demonstrate that both the gradient map sequence and the path-consistent multi-step optimization mechanism contribute significantly to accuracy improvements. Sensitivity analyses further show the effects of interpolation step count and scheduling strategies, validating that appropriate structural evolution granularity leads to more stable optimization. In addition, the framework consistently boosts performance across different backbones, including ResNet-based SimpleBaseline and HRNet variants, confirming its generality and compatibility.

Experiments on the COCO and CrowdPose benchmarks verify the effectiveness of the proposed framework. With an HRNet-W48 backbone, our approach achieves 77.6 mAP on the CrowdPose validation set, surpassing state-of-the-art models such as TransPose-H (76.3 mAP). On the COCO validation set at 256×192 resolution under ground-truth bounding box settings, our method reaches 75.6 mAP, performing on par with or exceeding several transformer-based approaches. These results highlight the framework's strong structural modeling capability and robustness in complex scenarios.

Overall, this study introduces a new paradigm for structure-conditioned optimization in pose estimation, offering an effective way to combine explicit geometric priors with deep feature learning. With its high extensibility, the proposed structure-guided multi-step refinement mechanism has strong potential for more complex real-world environments, such as dense crowds, motion blur, or extreme lighting, and can be naturally extended with temporal cues or multimodal inputs.

Keywords 2D human pose estimation; structure guidance; structural modeling; gradient map sequence; multi-step refinement

1 引 言

人体姿态估计(Human Pose Estimation)是计算机视觉中的核心任务之一，旨在从图像或视频数据中检测并定位人体关键点(如关节)，以二维或三维坐标形式表示人体姿势结构，从而在像素空间中重建人体结构[1,2]。该任务通常被建模为一个结构化预测问题，其输出不仅包括单点坐标，更反映关键点之间的拓扑和语义关系，如左右对称、骨骼连接、关节角度约束等。因此，姿态估计不仅要求高精度的局部定位能力，还需对人体全局结构具有良好的建模能力，尤其在遮挡、多人交互、复杂动作等场景下更具挑战性[3]。人体姿态估计在动作识别[4]、人机交互[5]、医疗康复[6]、虚拟现实[7]以及智能监控[8]等多个领域具有广泛的应用。其中，2D人体姿态估计指的是从单幅二维图像中恢复人体关键点的二维坐标，以简洁直观的方式表征人体姿态[9]。

在深度学习技术普及之前，人体姿态估计主要依赖于手工设计特征与结构化建模，这类方法通过几何约束与模板匹配进行姿态推断，其中 Pictorial Structures(PS)模型是最具代表性的研究之一。这类方法通过显式建模人体各关键点的空间几何约束，提升姿态估计的精确性和稳定性[10,11]。然而，受限于手工特征设计的局限性以及对复杂环境的泛化能力不足，传统方法难以适应多样化的现实场景。

随着深度学习技术的快速发展，基于卷积神经网络(CNN)的方法逐渐成为人体姿态估计的主流方向。这些方法利用CNN自动学习视觉特征，并通常采用高斯热图(Gaussian Heatmap)作为监督信

号，显著提升了关键点检测的准确性[12-15]。随着数据集规模扩大和计算资源的提升，姿态估计逐渐从单人场景扩展至多人复杂场景。为应对多人定位与身份关联问题，研究者提出了自顶向下与自底向上两类方法：自顶向下方法先检测人体实例，再分别估计每个人体的关键点姿态[14,16-20]；自底向上方法则以关键点为检测目标，先定位所有关键点，再通过组装方式完成关节归属，从而划分为各个人体实例[21-29]。此外，部分工作提出了单阶段方法，试图在一个端到端网络中同时完成人体检测与姿态估计，虽然其表面上为单阶段的，但本质上仍属“先定人再定关键点”的范式变体[30,20,31]。

尽管基于CNN的方法取得了显著进展，现有方法在人体整体结构建模方面仍存在不足，主要依赖于特征图的局部感受野，缺乏显式的人体结构先验建模。这使得在严重遮挡、复杂姿态或多人交互等复杂情境下，姿态估计的鲁棒性与合理性仍面临挑战[32,33]。鉴于姿态估计天然具有图结构特征，近年来研究者重新关注结构性建模方法，尝试引入图卷积网络（GCN）[9]、变换器结构（Transformer）[34,35]或关键点显式连接机制[36]，以增强模型对人体结构的建模能力[37]，从而提升模型对复杂场景的适应能力。

近年来，Diffusion Models凭借其稳定的训练过程与逐步建模的优势，在图像生成、超分辨率与语义编辑等任务中获得显著进展[38-46]，并进一步扩展至视频生成[47]、人体动作建模[48]、音频合成[49,50]、三维形状重建与生成[51,52]等多模态场景，展现出强大的通用性与结构建模潜力。尽管其原生范式主要面向生成任务，但其逐步迭代优化预测结果的特性，为结构化回归类任务提供了一种新的建模视角。在人体姿态估计领域，也曾有工作探索通过迭代方式逐步修正姿态预测结果，例如 Iterative Error Feedback（IEF）将姿态估计建模为从初始状态逐步逼近最终目标的过程，从而提升了输出结构的一致性与可解释性。我们注意到，扩散模型从初始输入出发、通过多步预测逐步生成结构化结果的方式，与IEF等方法在建模思路上具有一定相似性。

基于这一观察，本文提出了一种基于结构引导的人体姿态估计框架，核心思想是通过构建“结构引导图序列（General structure guidance map sequence）”引导姿态预测结果从初始状态逐步演化至结构合理的目标。在该框架中，我们将人体结构的先验知识与深度模型的特征学习能力相结合，显式构造了一系列中间结构状态，用于分阶段约束和优化姿态估计过程。每一步的中间状态不仅反映了姿态结构的渐进变化，也为多步推理过程提供了结构一致性的持续引导。相比于单步预测或仅依赖数据驱动的方式，基于结构引导的多步优化机制提升了模型的可控性和输出结果的结构合理性，尤其在复杂姿态或严重遮挡等场景下，有助于提升预测的一致性与物理可行性。

值得注意的是，尽管本文方法借鉴了逐步演化的思想，但与传统扩散建模以高斯噪声为起点、逐步去噪的范式不同，我们以结构合理的初始姿态和结构先验为起点，引导模型在多步优化过程中持续贴合人体结构约束，更好地适应了姿态估计任务的结构化本质。

本文的主要贡献包括：

（1）提出了一种基于结构引导的人体姿态估计框架，通过构建结构引导图序列，将关键点预测建模为从初始状态向结构合理目标逐步演化的过程，实现对结构一致性与全局协调的有效建模。

（2）显式设计多步结构优化路径，在每一步引入姿态结构约束，使预测过程兼具物理可行性与高鲁棒性，尤其在复杂场景（如多人遮挡、极端姿态）下表现突出。

（3）将结构先验与深度特征表征有机融合，在不依赖噪声建模的前提下，实现结构化建模与感知能力的协同优化，显著提升了关键点定位的准确性和模型的泛化能力。

本研究为复杂结构化回归任务中的结构引导优化提供了一种新范式，为后续相关领域的方法设计提供了理论和实践参考。

2 相关工作

2.1 基于深度学习的人体姿态估计方法

人体姿态估计在深度学习驱动下取得了显著进展，目前主流方法主要分为基于卷积神经网络（CNN-based）和基于变换器结构（Transformer-based）两大类别。

在基于卷积神经网络方法中，最具代表性的路线是通过卷积网络提取图像特征，并以高斯热图为监督信号，实现对关键点的精确定位。这一类方法以 DeepPose、Convolutional Pose Machine（CPM）、Stacked Hourglass Network、SimpleBaseline、HRNet等为代表[1,10-14,16,53]，在多个公开数据集上持续保持较高精度，并广泛应用于实际场景。与此同时，近年

来又出现了一类将关键点横纵坐标分别建模为一维概率分布的方案，如SimCC[17]、RTMPose[19]、RTMO[20]等，即采用有序分类或分布回归方式替代传统热图回归策略，以进一步提升定位分辨率和推理效率。这两类基于卷积神经网络的方法目前在学术界和工业界均被广泛采用，并持续推动关键点定位的精度和实用性提升。

基于变换器结构方法[54]近年来也受到了广泛关注。以ViTPose[18]为代表，采用变换器编码器（Transformer Encoder）作为特征主干，结合卷积网络作为解码器输出关键点热图，充分利用了MAE（Masked AutoEncoder）[55]预训练的变换器在全局关系建模方面的优势。此外，TokenPose[34]、TransPose[35]、Poseur[56]、ED-Pose[57]、PETR[58]等方法通常以CNN作为基础特征提取模块，再结合变换器模块（包括Encoder-only和Encoder-Decoder结构）对关键点间的全局依赖关系进行建模，实现端到端的人体姿态估计。这些方法有效提升了结构表达能力与全局信息建模能力，成为推动性能持续提升的关键技术路线之一。

尽管上述两类方法在关键点定位精度、推理效率等方面不断取得进展，但并未有对于人体结构的显式建模。无论采用何种特征主干，大多数方法依然主要依赖数据驱动，难以充分表达人体拓扑约束和空间结构先验。因此，在复杂遮挡、关节歧义和极端姿态变化等场景下，结构一致性和鲁棒性仍有待进一步提升。

2.2　姿态估计中的结构建模与姿态优化（Pose Refine）策略

在人体姿态估计任务中，结构一致性的建模能力对于提升模型性能和鲁棒性具有决定性作用。为了克服卷积网络局部感受野和缺乏显式结构约束的局限，近年来研究者提出了多种结构建模与姿态精修策略，力图将人体骨架的空间拓扑与先验知识引入关键点预测流程[33]。

多步优化与反馈机制是提升结构一致性的重要路径。典型方法如Iterative Error Feedback（IEF）-将姿态估计建模为一个多步修正过程，每一步根据当前预测与目标之间的残差进行反馈引导，通过逐步优化有效提升了输出的结构一致性。这类方法显著增强了结构表达能力，但在实际应用中仍面临训练复杂度高、收敛效率有限等挑战。

随着对结构信息建模需求的提升，基于图神经网络（Graph Neural Network，GNN）的结构感知方法被广泛关注。此类方法以人体骨架关键点为节点，通过图卷积捕捉关节点间的空间与语义依赖，实现结构先验与特征表达的有效结合。Graph-PCNN[32]、GCN-Pose[59]、SGCN[37]、Dual Graph Networks（DGN）[60]等方法通过多分支图结构与自适应邻接矩阵设计，不仅实现了结构先验与特征表达的有机融合，也在复杂场景下有效提升了结构一致性和鲁棒性。

与此同时，变换器结构（Transformer）及注意力机制的引入为结构建模带来了新的突破。变换器结构能够全局建模长距离依赖关系，在姿态估计中支持更丰富的结构信息融合。部分研究将变换器结构作为结构建模模块，与卷积网络（CNN）或图网络（GNN）联合用于结构一致性提升，进一步推动了端到端建模和全局信息表达能力的提升[46,50]。

尽管上述结构建模与多步优化方法在提升结构一致性和鲁棒性方面取得了显著进展，但仍难以在效率、全局协同和结构先验利用等方面实现最佳平衡。近期扩散模型在逐步优化与结构建模领域[61,62]提供了新的启发，强调通过多阶段演化过程提升输出的结构一致性。本文提出的结构引导多步优化框架，正是基于上述多步反馈、结构感知和逐步优化思想的有机融合，进一步推动了结构化建模策略在深度姿态估计中的发展。

3　关键方法

3.1　方法概览

本文提出了一种基于结构引导的多步优化框架，用于提升人体姿态估计的结构一致性与鲁棒性，其整体框架如图1所示。模型由三个主要模块组成：特征提取的姿态估计网络、渐变图序列生成模块和多步优化模块。本方法基于自顶向下范式展开，假设已获得人体检测框信息，输入图像为裁剪后的单人区域图像，经过缩放等数据增强处理后送入后

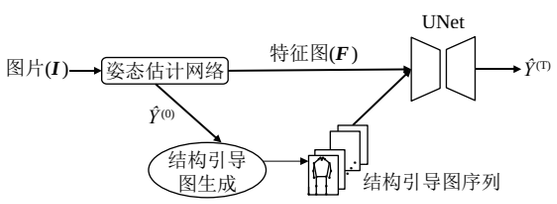

图1　整体结构示意图（输入图像$\boldsymbol{I}$经过姿态估计网络提取特征$\boldsymbol{F}$与初始热图$\hat{\boldsymbol{Y}}^{(0)}$，并生成渐变图序列，与特征$\boldsymbol{F}$一同作为UNet的条件输入，输出最终热图$\hat{\boldsymbol{Y}}^{(T)}$。）

续模块用于姿态预测。整个流程以图像为输入，逐步生成精细且结构一致的姿态估计结果。

模型的输入为RGB图像$\boldsymbol{I}\in\mathbb{R}^{3\times H\times W}$，其中$H$，$W$分别为图像高度与宽度。图像首先输入姿态估计网络（如ResNet等），提取得到特征图$\boldsymbol{F}\in\mathbb{R}^{C\times h\times w}$，并同时预测初始关键点热图$\hat{\boldsymbol{Y}}^{(0)}\in\mathbb{R}^{K\times h\times w}$，其中$C$为特征通道数，$K$为关键点数量，$h$，$w$为下采样后的热图空间尺寸。随后，在渐变图序列生成模块中，基于初始估计$\hat{\boldsymbol{Y}}^{(0)}$和结构先验，显式构造出从初始状态到目标结构的渐变图序列$\left\{\boldsymbol{H}^{(t)}\right\}_{t=0}^{T-1}$。这些图以时间维度组织，形成一个四维张量，其中$T$为优化步数。多步优化模块以渐变图序列为引导，在每一步$t(0\leq t<T)$上，网络不仅利用前一阶段的预测结果$\hat{\boldsymbol{Y}}(t)$和特征图$\boldsymbol{F}$，还结合结构引导信号$\boldsymbol{H}^{(t)}$进行优化。通过递归优化，每一步输出都受到结构先验的持续约束，从而使得最终预测结果具备更高的结构一致性和物理可行性。值得一提的是，本方法的多步优化模块受扩散模型、IEF等逐步优化思想的启发，但不同于传统扩散建模依赖高斯噪声逐步去噪，本文以结构先验和初始估计为基础，显式建模关键点演化路径，有效提升了复杂场景下的姿态估计表现。

综上，本文提出的结构引导多步优化框架通过特征提取、渐变图序列建模和多步优化三个阶段，实现了深度特征与结构先验的有机融合。为更清晰地呈现模型设计，以下第2至第4节将分别介绍姿态估计网络、渐变图生成模块与Unet模块的具体实现，第5节则进一步描述模型在推理阶段的多步优化流程。

3.2 姿态估计网络

姿态估计网络用于从输入图像中提取语义特征并生成初始姿态估计结果，是整个模型流程的起始模块。该模块对应于自顶向下范式中的单人姿态估计（Single Person Pose Estimation，SPPE）部分，在检测到人体框后对每个个体独立进行姿态预测。该部分结构成熟、实现稳定，广泛应用于主流自顶向下方法中，如SimpleBaseline、HRNet等，本文借助其中的特征提取与热图预测能力，作为后续结构建模与多步优化的输入基础。

与本方法的核心模块不同，姿态估计网络作为外部引入的组件，其参数在训练过程中保持冻结状态，不参与本文模型的训练，仅在前向过程中提供特征表示与初始预测结果。其整体行为可抽象为如下映射关系：

$$f_{pose}:\boldsymbol{I}\longrightarrow\left(\hat{\boldsymbol{Y}}^{(0)},\ \boldsymbol{F}\right)$$

其中，输入为RGB图像$\boldsymbol{I}\in\mathbb{R}^{3\times H\times W}$，输出为初始关键点热图$\hat{\boldsymbol{Y}}^{(0)}\in\mathbb{R}^{K\times h\times w}$和图像特征表示$\boldsymbol{F}\in\mathbb{R}^{C\times h\times w}$。

图像特征$\boldsymbol{F}$并非由额外分支网络生成，而是从姿态估计网络主干中截取的中间表示。例如，在SimpleBaseline架构中，热图预测模块由三个上采样卷积层，均可作为姿态估计网络的输出特征图$\boldsymbol{F}$；而在HRNet中，最终预测前的高分辨率融合阶段产生了多尺度的综合特征，这些各尺度的特征均可视为由HRNet这个网络提取的、与姿态估计相关的特征图$\boldsymbol{F}$。本文在不改变网络结构的前提下，直接利用上述特征作为后续结构建模与多步优化的条件输入。

热图预测部分输出的$\hat{\boldsymbol{Y}}^{(0)}$为姿态估计网络回归预测的置信热图，其中每个通道对应一个关键点，其像素值表示该关键点出现在对应位置的置信度分布。该热图既是渐变图序列的生成基础，也是后续结构引导多步优化模块的输入，同时还可作为性能基线（baseline），用于衡量后续模块的改进效果。

3.3 渐变图生成模块

为了模拟姿态估计过程中的结构性演化路径，本文引入了渐变图生成模块，用于构建一组由初始姿态逐步过渡至目标热图的中间表示，即结构引导图序列。由本模块构造的结构引导图序列，统一称为“渐变图序列”（Gradient Map Sequence）。该模块位于姿态估计网络与多步优化模块之间，其输出在训练和推理过程中均作为结构性约束信号，引导模型逐步优化当前预测结果。通过人为构造这种“结构连续性路径”，我们能够显式刻画多步优化过程中关键点空间分布的变化趋势，从而提升模型对姿态结构的建模能力。

3.3.1 姿态坐标插值

渐变图序列的构造首先依赖于目标姿态的关键点坐标。为此，本文将姿态估计网络输出的置信热图解码为二维坐标表示。解码过程中，采用最大响应位置作为关键点预测位置，并对每个关键点对应的最大响应值进行置信度判定。为提升插值质量并避免噪声干扰，仅将置信度高于0.2的关键点保留其预测坐标。其余关键点则被置为零坐标（即坐标设为(0,0)），以确保所有样本在关键点数量上的一致性。最终得到的目标姿态坐标记为$\boldsymbol{P}_{\text{pred}}\in\mathbb{R}^{K\times 2}$，其中$K$为关键点总数，维度在不同样本间保持不变。

在获得目标姿态后，本文引入一组预定义的初始姿态坐标$\boldsymbol{P}_{\text{start}}\in\mathbb{R}^{K\times 2}$，用于基于结构的多步优化路径的起始状态。该初始姿态保持不变，通常为四肢自然下垂、结构清晰的标准站立姿势。

为构造从初始姿态到目标姿态的演化路径，本文借助扩散模型中的余弦调度（cosine schedule）构造插值权重序列。设插值数量为$S+1$，即在初始姿态与目标姿态之间插入$S-1$个中间坐标，共生成$S+1$个姿态。使用余弦算法生成一组单调递增的插值权重$\{\alpha_s\}_{s=0}^{S}\in[0,1]$，并据此定义每一阶段的插值坐标为

$$\boldsymbol{P}^s=(1-\alpha_s)\cdot\boldsymbol{P}_{\text{start}}+\alpha_s\cdot\boldsymbol{P}_{\text{pred}}。$$

其中，$s=0,1,\cdots,S$。这里使用姿态估计网络的输出姿态，而不用真值（Ground Truth）姿态参与插值，主要是考虑到训练与推理过程的一致性。若在训练过程中使用真值，而推理过程中无法使用真值，则会导致训练与推理过程的不一致，从而使得训练出的模型的泛用性降低，在推理过程中无法获得更好的成绩。

该插值序列作为中间结构表示的基础，将在后续阶段转化为高斯热图（详见3.3.2），并根据固定的映射关系对应到结构引导路径中的时间步（详见3.3.3）。本文采用的非线性插值策略受扩散建模机制的启发，在保证结构过渡连续性与可控性的同时，为结构引导多步优化提供了可调控的路径设计。

3.3.2　高斯热图的生成

在上一节姿态插值的基础上，我们将每阶段的关键点坐标表示转换为图像空间中的密集热图，以供模型感知结构约束。具体而言，本文对每个插值姿态$\boldsymbol{P}^{(s)}\in\mathbb{R}^{K\times 2}$生成其对应的高斯响应图$\boldsymbol{H}^{(s)}\in\mathbb{R}^{K\times h\times w}$，以更贴近主流姿态估计网络的输出格式，并便于与后续结构建模模块进行对齐。

每个关键点k在热图上的响应被建模为以其坐标(x_k,y_k)为中心、标准差为σ的二维高斯分布，其表达式为

$$\boldsymbol{G}_k^{(s)}(i,j)=\exp\left(-\frac{(i-x_k)^2+(j-y_k)^2}{2\sigma^2}\right)。$$

其中，(i,j)表示热图中像素位置，满足$i\in\{0,\cdots,h-1\}$，$j\in\{0,\cdots,w-1\}$。标准差σ控制了高斯响应的扩散范围，其取值并非固定常量，而是依据热图分辨率动态设定。本文采用的输入图像大小为256×192，输出热图分辨率为64×48，在此设定下经验性地选取$\sigma=2$。当输入图尺寸或网络下采样比率变化时，σ值亦应随之进行相应缩放，以保持响应在图像空间中的结构一致性。

从(i,j)的取值范围可以看到，对于坐标为$(0,0)$的关键点，本文在生成高斯热图时直接跳过其响应计算，即该关键点对应的整个响应图为零张量，避免无效预测对结构建模产生干扰。该处理方式不仅增强了模型对异常关键点的鲁棒性，也有助于保留姿态结构中的稳定部分信息。为提升处理效率，我们在实际实现中采用向量化的批处理方式，将所有插值步对应的坐标统一转换为张量形式，并一次性生成高斯热图序列，显著减少了逐图循环带来的计算开销。最后，我们获得插值步数为$S+1$的高斯热图序列，记为$\{\boldsymbol{G}^{(0)},\cdots,\boldsymbol{G}^{(S)}\}$，将作为下一阶段“渐变图”生成模块的输入，并贯穿整个结构引导多步优化过程。

3.3.3　渐变图序列的张量构造

在上一节生成的高斯热图的基础上，我们将所有插值阶段对应的热图序列统一组织为模型可处理的张量结构。具体而言，本文将每一阶段热图$\boldsymbol{G}^{(s)}\in\mathbb{R}^{K\times H\times W}$按顺序堆叠，构建一个四维张量$\boldsymbol{H}\in\mathbb{R}^{(S+1)\times K\times H\times W}$，其中$S+1$表示插值序列的总长度，包含初始与目标两个端点。

为了将该插值张量进一步转换为多步优化过程中的渐变图序列，本文引入了该序列的时间建模机制。设多步优化过程包含T个时间步，本文设定$T=S+2$，以确保在插值阶段与多步优化阶段之间建立起合理映射。在每个时间步$t\in\{0,\cdots,T-1\}$，我们生成对应的渐变图$\boldsymbol{H}^{(t)}$，用于对该阶段的预测提供结构性约束。

为此，本文设计了一种加权融合策略：每个渐变图$\boldsymbol{H}^{(t)}$并不直接对应某一个插值步s，而是由所有插值热图$\{\boldsymbol{G}^{(s)}\}$按权重进行最大值融合，得到

$$\boldsymbol{H}^{(t)}=\max_s\left(w_{t,s}\cdot\boldsymbol{G}^{(s)}\right)。$$

其中，权重函数$w_{t,s}$的定义方式如下：

$$w_{t,s}=\begin{cases}1, & \text{当}(t=0,s=0)\text{或}\\ & (t=T-1,s=S-1)\\ \dfrac{S-2-|s-(t-1)|}{S-1}, & 1\le t\le T-2,\text{且}\\ & |s-(t-1)|\le S-2\\ 0, & \text{其他。}\end{cases}$$

该权重定义确保边界时间步完全由对应插值热图控制，而中间时间步则根据s与$t-1$的距离线性衰减，并仅对满足跨度条件的插值阶段赋予非零权重。上述“融合”操作中的最大值计算指对所有插值阶段的加权热图进行逐元素比较，保留每个像素位置上响应值最大的那一个，从而保留关键点在结构演化过程中的显著响应区域。

为了增强该机制的直观性，图2给出了当插值步数$S=4$时，各时间步所对应的插值权重分布情况。此时结构引导多步优化过程的时间步数为$T=6$，插值热图为$\{\boldsymbol{G}^{(0)},\cdots,\boldsymbol{G}^{(3)}\}$。例如：

1. 在$t=2$时：

$$\boldsymbol{H}^{(2)}=\max\left(\frac{2}{3}\boldsymbol{G}^{(0)},\ \frac{1}{3}\boldsymbol{G}^{(1)}\right)。$$

2. 在$t=4$时：

$$\boldsymbol{H}^{(4)}=\max\left(\frac{1}{3}\boldsymbol{G}^{(2)},\ \frac{2}{3}\boldsymbol{G}^{(3)}\right)。$$

这种分段线性加权策略不仅在数值上实现了从粗略结构到精细结构的平滑过渡，也为多步优化过程提供了结构连贯性的先验控制。在图2中，我们可视化展示了每一时间步对应的插值热图权重分布，以进一步说明该机制的合理性。通过上述机制，我们构建了一个时间对齐的渐变图序列$\{\boldsymbol{H}^{(t)}\}_{t=0}^{T=1}$，后续将在模型的每一步优化中提供结构约束信息，具体使用方式将在第5节中详细介绍。

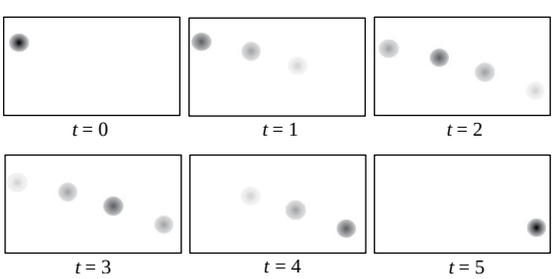

图2 某一关键点在时间步$T=6$下的渐变图序列（从左上至右下依次为$t=0\sim5$。每张图展示了该关键点在对应时间步的高斯热图分布，由多个插值阶段的热图按权重进行最大值融合得到。该序列由3.3节定义的权重函数生成。此图直观展现了关键点响应如何从初始状态向目标热图逐步过渡，体现了本文所设计的“结构演化路径”。）

3.4 结构引导多步优化模块

为进一步提升姿态估计中关键点预测的结构一致性与鲁棒性，本文在特征提取与渐变图生成的基础上，引入结构引导多步优化模块。该模块借鉴了扩散模型和IEF等逐步优化策略的思想，但不再对高斯噪声或扩散概率进行建模，而是聚焦于结构先验的多步约束与逐步优化机制。通过多步优化，实现从粗略预测到结构合理的高精度姿态热图的动态演化。具体而言，网络在每个时间步t接收姿态估计网络输出的特征图$\boldsymbol{F}$与当前时刻的渐变图$\boldsymbol{H}^{(t)}$，生成更新后的预测热图$\hat{\boldsymbol{Y}}(t)$，并逐步迭代逼近目标热图。

本文提出的结构建模机制具有显式引导性：通过人为构造的渐变图序列，在每一步显式控制生成行为，并结合深层特征表征，实现对预测路径的稳健调控。在网络设计上，本模块以标准Unet架构为基础，充分利用多尺度感知能力与跳跃连接策略，从而在每个时间步上实现空间一致、结构合理的热图重建。为更好地揭示该模块的建模逻辑与目标函数，以下将系统介绍多步优化目标、损失设计及网络实现的具体策略。

3.4.1 结构引导多步优化原理与训练目标

本文将姿态热图的生成过程建模为一个多步迭代过程，其核心思想是：通过逐步优化，使预测热图逐步趋近于结构合理的目标热图分布。本文采用由“渐变图生成模块”显式构造的渐变图序列$\{\boldsymbol{H}^{(t)}\}_{t=0}^{T-1}$作为引导路径信息，以控制每一时间步上的预测行为，从而提升生成过程的结构一致性与可控性。

在训练过程中，我们随机选取时间步t，并使用对应的渐变图$\boldsymbol{H}^{(t)}$与特征图$\boldsymbol{F}$作为条件输入，监督网络输出$\boldsymbol{P}^{(t)}$逼近目标热图$\boldsymbol{H}^{*}$。损失函数采用均方误差（MSE）形式：

$$\mathcal{L}=\left\|\boldsymbol{P}^{(t)}-\boldsymbol{H}^{*}\right\|_2^2。$$

该目标函数鼓励网络在任意结构阶段下均能生成结构合理的姿态热图，从而提升结构引导过程的整体稳定性与泛化能力。

3.4.2 Unet网络结构与条件融合方式

为实现姿态热图在结构空间中的逐步优化，本文采用Unet架构[63]作为结构引导多步优化模块的核心。该网络继承了典型的编码-解码结构，并在每个尺度上引入跳跃连接以融合多尺度信息。相较于传统分割任务中的Unet，本工作实现的Unet变体更侧重于时序信息、结构引导以及多尺度特征的联合建模，充分满足结构引导多步优化对特征融合和动态条件建模的需求。

Unet的输入由三部分组成：主干姿态估计网络输出的图像特征图列表$\{F_n\}_{n=0}^{N-1}$、渐变图序列中第t步的渐变图$H^{(t)}$，以及当前的结构引导阶段索引（时间步）t。因此，整个输入结构可形式化地表示为

$$\hat{Y}(t)=\mathrm{Unet}_\theta\left(\{F_n\}_{n=0}^{N-1},\ H^{(t)},\ t\right)。$$

在具体实现中，本文采用了一种多尺度条件融合方式，将姿态估计网络输出的特征图列表注入至Unet编码路径（Downscale）的不同尺度层中。图像特征图列表包括多个分辨率层，分别来自姿态估计网络不同阶段的输出（如Simple Baselines的最后三层或HRNet的不同分辨率分支）。随后，Unet编码路径上的每一层输出的中间特征与主干网络输出的相应分辨率特征图进行融合，确保时序信息与结构引导得以贯通至各尺度特征表达中。此类融合在下采样过程中不断进行，并为后续上采样阶段提供跳跃连接支持，实现解码恢复。

在渐变图序列的使用上，本文仅将分辨率为$h\times w$的目标热图$H^{(t)}$与相同分辨率下的图像特征图进行通道维度的拼接。该拼接操作发生在Unet编码路径的开始，作为网络的输入之一参与多尺度下的后续建模。为了建模时间演化行为，网络内部通过位置编码机制将时间步t映射为向量e_t，并在多个残差模块中注入该时间嵌入，从而实现时间条件控制。由于所有输入张量在空间上均已对齐，网络整体结构无需额外的尺寸调整或对齐操作，推理流程保持一致性与效率。

在分辨率下降路径中，充分融合了多尺度条件特征和结构引导信息，有助于提升编码阶段的空间与结构感知能力。对于Unet的其余部分，包括编码与解码之间的中间连接、解码（Upscale）模块以及各级跳跃连接（skip-connection），均沿用结构引导模型中主流的标准Unet架构设计。这一结构不仅保证了优化过程的稳定性与高效性，也为结构引导下的多步优化机制提供了坚实的网络基础。图3示意了上述条件注入与特征融合的结构实现，图中详细展示了多尺度条件注入、渐变图拼接及主干Unet的信息流向。

3.5　模型训练与推理

为了实现关键点预测的结构一致性与逐步收敛能力，本文在训练与推理阶段均引入了结构引导多步优化机制。该机制以结构引导的“渐变图序列”为条件输入，驱动模型从初始状态出发，逐步趋近于目标姿态。考虑到训练与推理所依赖的先验信息差异显著，我们分别设计了差异化的时间步选择策略与输入构建流程，以适应模型在有监督与无监督场景下的演化建模需求。

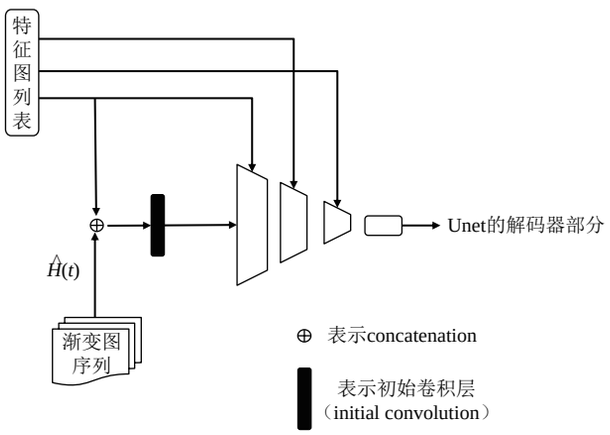

图3　本文所用Unet编码路径（Downscale部分）的条件融合结构示意图（当前时间步的预测热图与渐变图在通道维拼接后，经初始卷积作为输入。多尺度图像特征在各尺度分别注入至对应编码层，与中间特征图拼接融合。为突出条件注入的细节，图中省略了解码路径（Upscale部分），该部分结构与标准Unet保持一致，详见正文描述。）

本节将分别介绍模型在训练阶段的监督设计与样本构造方式，以及推理阶段的多步预测流程。

3.5.1　训练阶段

姿态估计网络$\mathcal{B}(\cdot)$接收输入图像I，并生成初始预测姿态坐标$P_{\text{pred}}\in\mathbb{R}^{K\times 2}$，同时提取图像特征$F=\mathcal{B}(I)$。随后，算法用预先定义的初始姿态坐标$P_{\text{start}}\in\mathbb{R}^{K\times 2}$与$P_{\text{pred}}$构建渐变图序列。构建过程分为三个步骤：首先，以$P_{\text{start}}$和$P_{\text{pred}}$为端点，通过余弦插值（cosine schedule）生成$S+1$帧插值姿态序列$\left\{P^{(0)},\cdots,P^{(S)}\right\}$；其次，基于该序列生成一组高斯热图$\left\{G^{(0)},\cdots,G^{(S)}\right\}$；第三步，由高斯热图，通过渐变图生成算法（见3.3）最终形成T个时间帧的渐变图序列$\left\{H^{(0)},\cdots,H^{(T-1)}\right\}$。在每一次训练样本的构建中，算法从序列中均匀随机采样一个时间步$t\in\{0,1,\cdots,T-1\}$，然后从当前的渐变图序列中提取对应的渐变图$H^{(t)}$。图像特征F、渐变图$H^{(t)}$以及时间步t输入多步优化网络，预测当前关键点热图分布$G_{\text{out},t}\in\mathbb{R}^{K\times H\times W}$。

为了保持目标的一致性，监督信号采用由真值（ground truth）P^{gt}生成的目标热图G^{gt}，损失函数定义如下：

$$\mathcal{L}=\left\|G_{\text{out},t}-G^{\text{gt}}\right\|_2^2。$$

其中，$\|\cdot\|_2^2$表示逐像素均方误差。最终，训练目标是在全体样本的随机时间步上最小化该损失。该机制避免了传统多步训练中的步间梯度干扰问题，使模型在每一演化阶段都能独立收敛，从而提升结构建模能力。整体流程如算法1所示。

算法1. 基于随机步监督的结构引导多步优化训练流程

输入：图像I，Ground Truth姿态P^{gt}，时间频数T

输出：更新后的模型参数

1. 提取图像特征：$F=\mathcal{B}(I)$；
2. 计算初始预测坐标：$P_{pred}=coordinate\_from\_backbone(F)$；
3. 生成插值序列：$\{P^{(t)}\}_{t=0}^{T-1}=interplate\_poses(P_{start},P_{pred},T)$；
4. 生成热图序列：$\{G^{(t)}\}=generate\_gaussian\_maps(\{P^{(t)}\})$；
5. 构造渐变图序列：$\{H^{(t)}\}=compose\_graduated\_sequence(\{G^{(t)}\})$；
6. 随机采样时间步：$t\in\{0,1,\cdots,T-1\}$；
7. 拼接结构引导图与图像特征：$C=concat(F,H^{(t)})$；
8. 预测当前步热图：$G_{out}^{(t)}=UNet(C)$；
9. 构造监督热图：$G^{gt}=generate\_gaussian\_maps(P^{gt})$；
10. 计算损失函数：$\mathcal{L}=MSE(G_{out}^{(t)},G^{gt})$；
11. 执行梯度反向传播并更新模型参数。

3.5.2 推理阶段

推理阶段采用固定步长的迭代机制，以实现结构引导下多步优化过程的逐步演化。本文设计了一种基于初始估计自引导的多步优化过程，在结构一致性约束下，迭代生成最终预测结果。推理阶段与训练过程高度对应，但不再依赖监督热图。具体过程如下：

与训练阶段类似，输入图像I首先经由姿态估计网络$\mathcal{B}(\cdot)$提取图像特征F，并获得初始姿态估计P_{pred}。同样地，算法通过P_{start}和P_{pred}，经由插值序列$\{P^{(0)},\cdots,P^{(S-1)}\}$从而构造出渐变图序列$\{H^{(0)},\cdots,H^{(T-1)}\}$。

与训练阶段不同，这里通过迭代，使得多步优化网络在每一个步长$t\in\{0,1,\cdots,T-1\}$上均接收图像特征F、步长t以及对应渐变图$H^{(t)}$作为输入，并输出当前热图预测$G_{out,t}$。最终，经过T轮迭代后，输出的最后一步G_{out}，再经过常规的热图到坐标的转换方法，获得最终的估计$P_{out}=P_T$作为输出的预测结果。整个推理过程如算法2所示。

算法2. 基于结构引导的多步优化推理流程

输入：图像I，时间步数T

输出：最终姿态估计结果P_{pred}

1. 从图像I提取特征$F=\mathcal{B}(I)$；
2. 通过姿态估计网络获得初始预测坐标：$P_{pred}=coordinate\_from\_backbone(F)$；
3. 初始化当前预测姿态$P^{(0)}\leftarrow P_{start}$；
4. FOR $t=0$TO $T-1$DO：
5. 插值P_{start}与$P^{(t)}$得到序列$\{P^{(t,s)}\}$；
6. 生成热图序列：$\{G^{(t,s)}\}=generate\_gaussian\_maps(\{P^{(t,s)}\})$；
7. 构造结构引导图：$\{H^{(t,s)}\}=compose\_graduated\_sequence(\{G^{(t,s)}\})$；
8. 拼接结构图与图像特征：$C=concat(F,H^{(t)})$；
9. 输入结构引导多步优化网络预测热图：$G_{out}^{(t)}=UNet(C)$；
10. 解码出新坐标 $P^{(t+1)}=heatmap2coords(G_{out})$；
11. END FOR
12. 返回最终预测结果：$P_{pred}\leftarrow P^{(T)}$

4 实验评估

4.1 实验设置

本文在COCO-Keypoints数据集和CrowdPose数据集上开展实验，用以全面评估所提出方法在不同场景下的有效性与适应能力。COCO-Keypoints数据集是COCO（Common Objects in Context）数据集的一个子集，专门用于人体关键点估计任务。该数据集涵盖了丰富的姿态变换情形，提供了大规模标注数据，支持单人和多人场景下的姿态估计建模。COCO-Keypoints训练集包含118,287张图像，共计149,813个注释实例（persons），而验证集（val2017）包含5,000张图像，对应64,817个标注实例。CrowdPose数据集则面向高密度人群环境下的人体姿态估计任务。该数据集共包含约17,000张图像，划分为训练集（10,000张）、验证集（2,000张）和测试集（5,000张），共计标注约166,000个关键点。相较于COCO-Keypoints而言，CrowdPose中的人物遮挡率更高，个体间交互更加复杂，能够有效评估模型在复杂场景下的适应能力与鲁棒性。

在实验设置中，我们将 COCO-Keypoints 数据集作为主要实验平台，所有消融实验均在该数据集上完成，以保证在典型且分布稳定的训练环境中评估各模块的独立贡献。而在整体性能比较中，我们分别在 COCO-Keypoints 与 CrowdPose 两个数据集上进行测试，以全面验证模型在通用姿态估计场景与复杂人群场景中的适应能力与鲁棒性，从而展示所提方法的通用性与结构稳定性。为了全面衡量模型的预测性能，实验采用 OKS（Object Keypoint Similarity）作为人体姿态估计任务的主评价指标。OKS 通过关键点间的欧式距离、人体尺度以及关键点可见性权重来衡量关键点预测的准确性。所有实验中报告的评估结果均基于 OKS 计算得出。

在训练设置方面，所有实验均采用 Adam 优化器，学习率设定为 1×10^{-4}，并在整个训练过程中保持不变。模型训练共进行 130 轮（epochs），训练过程中使用多种数据增强策略，包括随机翻转、尺度缩放和随机裁剪，以提升模型对多样化输入的鲁棒性。所有输入图像在送入网络前，均被标准化至固定尺寸，并根据关键点热图生成对应的监督信号。损失函数采用均方误差（MSE），用于衡量预测热图与目标热图之间的像素差异。所有实验在配备 6 张 NVIDIA Titan XP GPU 的服务器上完成，训练过程中统一采用同一数据并行训练以提高训练效率。批量大小（Batch Size）设定为 48，以适配 GPU 内存，同时保证训练过程的稳定性与收敛性。

4.2　基于 Unet 的结构引导配置消融实验

为验证模型性能提升是否主要来源于结构引导机制的引入与多步优化流程，我们设计了如表 1 和图 4 所示的对照实验。本组实验将基础模型（Baseline）与引入 Unet 模块后的不同结构引导方案（A1～A4）进行比较，并进一步与完整结构引导多步优化模型（Ours）进行对照分析。其中，基础模型采用的是 Simple Baselines-ResNet-50，各方案的具体配置见表 1。消融实验的设计旨在分离网络容量扩展与结构引导机制本身对性能的贡献，并考察结构引导路径构造方式（如插值方式、路径一致性）对模型效果的影响。实验结果如图 4 所示，可以直观展示不同结构引导策略下模型性能的变化趋势。

表 1　不同结构引导配置下的模型

实验方案	模型配置
Baseline	仅使用姿态估计网络（无结构引导模块）
A1	Unet，无结构引导图（仅特征输入）
A2	Unet，预测热图作为结构引导图
A3	Unet，插值热图序列作为结构引导图
A4	Unet，基于 GT 的渐变图序列作为结构引导图
Ours	Unet，基于预测的渐变图序列作为结构引导图（路径一致）

注：不同结构引导配置下的模型方案对比。Baseline 为不使用结构引导模块，仅使用姿态估计网络；A1～A4 为不同结构引导图与路径配置的消融实验；Ours 采用预测渐变图序列作为结构引导图，实现训练-推理路径的一致性。

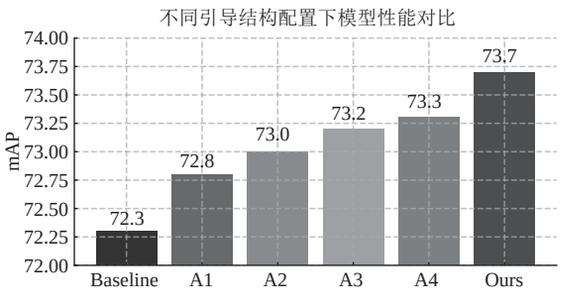

图 4　不同结构设计下模型在 COCO 验证集上的 mAP 性能对比（不同结构设计配置见表 1，对比结果验证了结构引导机制及路径一致性设计在提升模型性能方面的有效性。）

4.2.1　仅引入 Unet 模块（A1）

在 A1 中，我们在姿态估计网络（backbone）输出之后引入了一个 Unet 模块，其唯一输入为姿态估计网络提取的图像特征图，输出为关键点热图。该模块不接收任何形式的结构引导图、时间步嵌入或路径控制信号，也不具备逐步生成机制，因此不构成结构引导框架的完整建模流程。其设计目的在于模拟“增加网络容量但不引入结构机制”的情形，从而评估结构设计是否为性能提升的关键。

从实验结果来看，A1 相较于 Baseline 在 COCO 验证集上提升了 0.5 mAP（由 72.3 mAP 提升至 72.8 mAP），说明即便不引入结构引导图，网络结构复杂性增加所带来的中间表示增强能力在一定程度上仍可带来性能改善。然而，这种提升是有限的，且不足以解释最终模型在结构建模能力上的显著优势。当我们将 A1 与完整方法（Ours）进行比较时，性能差距进一步扩大至 0.8 mAP（73.7 mAP）。考虑到两者在主干结构和参数规模上的差异极小，这一显著差距表明，本文提出的结构引导机制与多步优化过程才是性能提升的决定性因素。结构引导图为每一阶段提供了清晰的姿态演化方向，多步预测机制则使模型能够逐步细化关键点间的空间关系，从而有效增强模型在复杂姿态估计场景下的结构建模能力与预测稳定性。

4.2.2 将预测热图作为输入的Unet模块(A2)

在前一节中,我们验证了单独引入Unet模块(A1)虽可带来一定程度的性能提升,但若缺乏结构引导信息,仍难以发挥其建模潜力。为进一步分析结构引导信号的贡献,我们在A2中引入由姿态估计网络预测得到的静态热图作为结构引导输入,并与其输出的特征图一起送入Unet,从而构成一个结构更为简化的引导版本。

我们使用这一方案模拟一类典型方法:直接利用已预测的热图信息作为额外通道,用于增强Unet对关键点空间分布的感知能力。值得一提的是,该方案未构造阶段性演化的路径,也未进行插值操作,仅提供了当前预测结果对应的热图作为静态输入,因此不具备结构引导路径中的时间动态性与结构约束机制。实验结果显示,A2相较A1在COCO验证集上带来了0.2 mAP的提升(由72.8 mAP提升至73.0 mAP),验证了即使是静态结构引导,也有助于提高特征融合与关键点位置建模的能力。然而,相较于后续引入插值构造与路径建模的结构引导方式(如A3与A4),该方案的性能提升依然有限,说明仅引入预测热图仍不足以有效引导网络逐步优化姿态结构,也无法建立一致且清晰的结构先验机制。

由于A2所使用的结构引导图尚未体现本文提出的核心设计理念,即通过连续性渐变图序列来建模结构演化路径,因此其提升幅度仍相对有限。在下一节中,我们将继续比较A2与A3,以进一步揭示渐变机制在结构优化中的作用。

4.2.3 带有插值热图作为输入的Unet模块(A3)

在上一节中,我们进一步在Unet模块输入中加入了姿态估计网络预测生成的热图(A2),用于增强模型对当前关键点空间分布的感知能力。虽然该方案在不引入任何结构演化机制的情况下,已取得了比单纯引入特征图(A1)更优的性能,但其结构引导仍是静态且粗粒度的,尚未体现本文方法所强调的多阶段结构演化建模思想。

为更进一步评估结构引导机制的建模潜力,我们在A3中引入了一组由初始姿态与最终预测姿态之间插值得到的中间关键点热图,作为结构引导图与特征图一起传入Unet,用以引导每一阶段的结构优化。该热图序列通过插值方式构造,刻画了从起点姿态到目标姿态的结构变化过程,尽管尚未构造完整的"渐变图"机制,但已具备一定的阶段性结构演化特征。该设置旨在探究,即使不引入精细的结构演化设计,仅通过引入一组插值热图提供一定的结构先验,是否也能显著提升多步优化过程的效果。

实验结果表明,A3相较于A1的性能进一步提升了0.4 mAP(由72.8 mAP提升至73.2 mAP),较A2的性能也提升了0.2 mAP,表明即使是较为粗略的结构引导,也能够增强结构优化路径中的结构一致性,从而改善关键点热图的空间布局与响应位置。与A1、A2相比,A3在中间预测阶段获得了更明确的姿态演化方向,进一步印证了结构信息对于结构引导多步优化机制建模能力的重要性。

尽管A3所使用的结构引导图虽已具备阶段性建模特征,但仍未体现本文提出的核心设计理念,即通过连续性渐变图序列来建模结构演化路径,因此其提升幅度仍相对有限。在下一节中,我们将继续比较A3与A4,以进一步揭示渐变机制在结构引导框架中的作用。

4.2.4 基于初始姿态与真值的渐变图序列作为输入的Unet模块(A4)

在前一节中我们验证了,即使仅使用插值热图序列作为结构引导,也能有效提升模型性能。为进一步评估我们所提出的"渐变图"对提升模型性能的核心作用,本节进一步构造一组更完整的结构演化路径,即"渐变图序列",输入Unet的则是渐变图序列与特征图。为了更全面地考查渐变图序列,我们采用了两种不同的版本。在A4版本中,训练阶段用来构造渐变图序列$\{\boldsymbol{H}_{\text{gt}}^{(t)}\}$的信息,由初始姿态$\boldsymbol{P}_{\text{start}}$与真值(Ground Truth)姿态$\boldsymbol{P}_{\text{gt}}$之间的插值提供;推理阶段,由于真值不能参与,因此渐变图序列$\{\boldsymbol{H}_{\text{pred}}^{(t)}\}$基于初始姿态$\boldsymbol{P}_{\text{start}}$与姿态估计网络输出的预测姿态$\boldsymbol{P}_{\text{gt}}$构造。另一个版本,即本文正式提出的完整版本Ours中,训练与推理阶段均使用预测结果$\boldsymbol{P}_{\text{pred}}$,始终基于初始姿态$\boldsymbol{P}_{\text{start}}$与$\boldsymbol{P}_{\text{pred}}$构造渐变图序列$\{H_{\text{pred}}^{(t)}\}$。

上述两种路径构造方式的关键差异在于,在训练阶段与推理阶段均会产生"训练-验证差距"(train-valid gap),只不过二者产生差距的方式有所不同。在A4版本中,差距的产生主要源于$\boldsymbol{P}_{\text{gt}}$与$\boldsymbol{P}_{\text{pred}}$的不同。而在我们的完整模型中(Ours),差距的产生主要源于训练集的$\boldsymbol{P}_{\text{pred}}$的精度远高于验证集的$\boldsymbol{P}_{\text{pred}}$。这主要是由于模型是在训练集上训练所导致的。

实验结果表明,A4的性能相较于A3仅提升了0.1 mAP(由73.2 mAP提升至73.3 mAP),而Ours则达到73.7 mAP,高于A4。由此可见,尽管A4在

训练阶段使用了渐变图序列，其结构引导路径在训练与推理阶段的不一致性仍限制了最终性能的发挥。相比之下，完整模型虽然使用的是预测结果作为路径终点，但由于构造方式的一致性（均是由模型预测给出的），整体训练-验证差距反而更小，从而带来更稳定的泛化性能。

这一对比结果表明，在结构引导框架中，路径构造方式的一致性较路径本身的理想程度更为关键。结构引导路径若前后保持一致性，不仅有助于稳定训练过程，也为推理阶段的结构演化路径提供了可复用的建模基础。

4.2.5　小　结

通过对多个对照实验的分析可见，本文提出的结构引导框架在多个维度上展现出显著的性能优势。首先，模型性能的提升并非仅来源于网络结构的扩张（A1），而是依赖于结构引导图本身所提供的姿态演化信息。其次，A2的实验进一步验证了，即使仅引入静态的预测热图，也能够在一定程度上改善特征融合效果，说明结构引导信号的加入本身即具备增强建模的能力。再次，实验表明结构引导图的构造方式对模型性能具有关键影响：仅使用静态热图作为结构提示（A2）虽可带来一定性能增益，进一步引入插值构造的中间热图序列（A3）后，模型获得了更明确的结构阶段性演化引导，而引入连续演化的渐变图序列（A4）则在建模路径动态一致性上进一步提升了结构表达能力，显著增强了多步优化路径中结构信息的阶段性建模效果。

此外，通过比较A4与完整模型（Ours）在训练阶段所采用的不同路径构造策略可以看出，尽管A4在训练阶段使用Ground Truth构造了更理想的结构路径，但由于推理阶段无法获得真值，其引导路径构造方式发生变化，导致训练-验证阶段间存在显著差异，从而限制了泛化性能的进一步提升。相比之下，完整模型（Ours）在训练与推理阶段均使用预测姿态构造渐变图序列，保持路径构造方式一致，有效缓解了结构信息引导路径的不一致问题，在泛化性能上实现了最好的表现。

综上，实验表明结构引导框架的性能提升主要来自两方面：一是通过渐变图序列引入结构演化信息，增强结构引导多步过程的建模能力；二是在训练与推理阶段保持路径构造方式的一致性，提升模型的稳定性与可复用性。

4.3　训练策略与架构设计的进一步分析

在4.2中，我们围绕结构引导路径的设计展开了系统的消融实验，验证了不同结构引导图构造方式对结构引导多步优化过程建模能力的影响。然而，除了结构设计本身，训练策略与架构设计同样会对模型性能与稳定性产生重要作用。本节将围绕以下三个方面展开分析：一是渐变图序列的插值步数设置，探讨结构演化阶段的建模效果；二是多步优化步长调度策略（schedule）的选择，分析其对训练稳定性与最终性能的影响；三是不同姿态估计网络在框架中的适应性，以验证所提方法的通用性与增益效果。通过这些补充性实验，我们进一步完善对结构引导框架的性能特性与可调参数的分析。

4.3.1　渐变图序列的插值步数敏感性分析

渐变图序列作为结构引导框架中的核心建模模块，其插值步数S决定了结构演化路径的阶段划分，进而影响结构引导多步优化过程中的结构信息表达能力。插值步数过少可能导致结构变化不足，影响引导效果；而步数过多则可能引入冗余信息，增加训练复杂度，甚至影响模型收敛性。为系统评估插值步数对模型性能的影响，本文开展了步数敏感性实验。我们的实验是从$S=3$开始，未纳入$S=1$和$S=2$两种配置。主要原因在于，$S=1$实际上等价于仅使用最终预测结果生成的热图作为结构引导（对应于4.2节的A2方案），而$S=2$仅包含初始姿态$\boldsymbol{P}_{\text{start}}$与预测姿态$\boldsymbol{P}_{\text{pred}}$之间的直接插值，但阶段过少，无法形成有效的渐变图序列，难以体现结构演化路径的多阶段建模能力。

实验在COCO验证集上进行，插值步数的设置为$S=3, 4, 5, 6, 7, 8$。如图5所示，模型在$S=3$和$S=5$时达到性能峰值（73.7 mAP），当步数进一

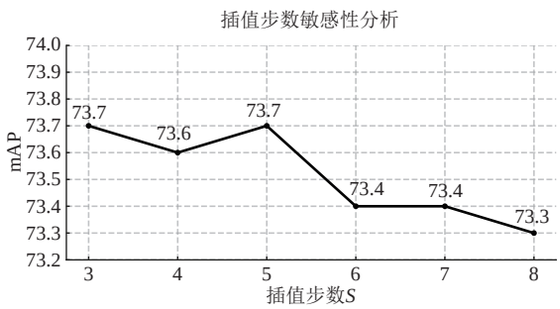

图5　在COCO验证集上不同插值步数S对结构引导框架性能的影响（实验结果表明，步数在$S=3$与$S=5$时达到性能峰值（73.7 mAP），当步数超过$S=5$后，性能开始下降。这说明适当的结构演化阶段有助于提升建模效果，但过长的演化路径会引入冗余，增加训练复杂度，反而限制了模型性能。最终，综合考虑性能与计算开销，本文采用$S=3$作为默认设置。）

步增加至$S=6$及以上时，性能出现下降，最终在$S=8$时降至73.3 mAP。这一趋势进一步说明，过长的路径会引入冗余，扰乱训练过程，反而限制了模型性能。

考虑到综合性能表现与训练复杂度，本文最终选用S=3作为渐变图序列的默认插值步数。该配置不仅具备最优的性能表现，同时也以最小的插值阶段数量，降低了结构引导图生成与模型训练的开销，兼顾了建模效果与效率。

4.3.2 多步优化调度策略对比分析

在结构引导框架中，步长调度策略（schedule）用于控制多步优化过程各阶段的步长变化，从而平衡结构演化路径的细粒度建模与整体优化稳定性。合理的调度策略能够在训练初期维持较大步长，鼓励模型在结构空间中探索更多姿态变换可能性；而在训练后期逐步缩小步长，以细化模型在关键结构区域的预测能力。因此，调度策略的选择在一定程度上影响模型的训练效率与最终性能。

本文在实验中对比了两种常见的调度策略：一是广泛应用于相关模型训练的余弦调度（cosine schedule），其步长在训练初期保持较大，后期逐渐收缩；二是线性调度（linear schedule），即步长按照线性递减方式进行调控。两者在多步优化阶段的动态调整曲线有所差异，可能会对模型训练过程产生不同影响。在COCO验证集上的实验结果如图6所示，余弦调度最终取得了73.7 mAP，而线性调度略低，为73.4 mAP。虽然两者在性能上的差距较小，但仍表明余弦调度在训练后期保留更大步长，能够在结构引导路径演化的后期阶段维持模型的结构调整能力，略优于线性调度。综合考虑训练稳定性与性能表现，本文最终采用余弦调度作为默认配置。

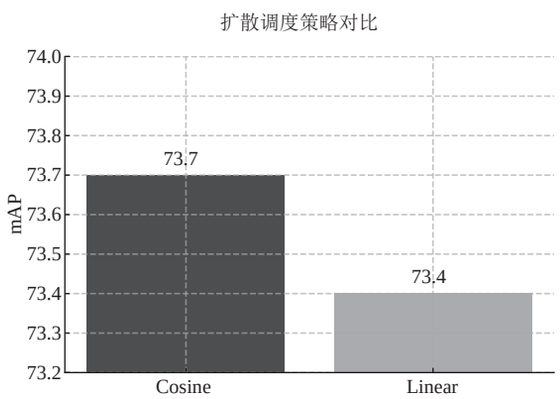

图6 在COCO验证集上采用不同步长调度策略（schedule）时，结构引导框架的性能对比结果（余弦调度在训练初期维持较大步长，后期逐步收缩，最终取得73.7 mAP，略优于采用线性递减步长的线性调度（73.4 mAP）。结果表明，余弦调度在结构演化后期阶段保留更多结构调整空间，训练更稳定，性能略优，因而被选为本文的默认配置。）

4.3.3 不同姿态估计网络下的通用性分析

为了进一步验证本文提出的结构引导多步优化框架在不同姿态估计网络下的通用性与增益效果，我们在多种主流姿态估计网络上进行了对比实验，涵盖了Simple Baselines（ResNet-50/101/152）及HRNet W32/W48。实验结果如表2所示，我们的方法在各类姿态估计网络下均带来一致性增益。其中，在基于ResNet-50的Simple Baselines上，我们的方法提升了1.3 mAP，达到73.7 mAP；在更深层次的ResNet-101和ResNet-152上，分别提升了0.8 mAP和0.4 mAP。尽管HRNet系列具备更强的特征提取能力，我们的方法在HRNet W32和HRNet W48上依然取得了良好的增益，分别提升了0.7 mAP和0.5 mAP，最终达到76.3 mAP和77.6 mAP。这些结果进一步验证了所提方法在不同规模与不同类型姿态估计网络下的适应性与增益效果。

表2 不同姿态估计网络下的模型性能对比

方法	基线（mAP）	Ours（mAP）	增益（mAP）
SimpleBaseline（ResNet-50）	72.4	73.7	+1.3
SimpleBaseline（ResNet-101）	73.5	74.3	+0.8
SimpleBaseline（ResNet-152）	74.2	74.6	+0.4
HRNet W32	75.6	76.3	+0.7
HRNet W48	77.1	77.6	+0.5

注：本表展示了在多种主流姿态估计网络（Simple Baselines：ResNet-50/101/152，HRNet W32/W48）下，结构引导框架（Ours）与基础模型（Baseline）的mAP性能对比结果。结果表明，所提方法在各类特征提取器上均取得了一致性的性能增益。其中，在Simple Baselines系列网络上，随着网络深度的增加，性能提升幅度逐渐减小，符合模型优化空间收敛的预期；在HRNet系列中，尽管基线性能更高，所提方法依然实现了显著的性能增益（W32提升0.7 mAP，W48提升0.5 mAP），验证了方法的通用性与稳定性。

为系统分析增益趋势，我们设计了两类实验：一是在同一系列的Simple Baselines姿态估计网络上，探究在同一体系内，特征提取能力增强对结构引导增益幅度的影响；二是在HRNet系列（W32/W48）上，补充验证方法在不同姿态估计网络结构下的通用性。从整体趋势来看，本文提出的结构引导框架在同一系列网络内部，随着网络规模的加大，性能增益呈现出逐渐减小的规律（如Simple Baselines中

ResNet-50 至 ResNet-152，HRNet W32 至 W48）。这一现象符合“规模越大，模型自身优化空间越小”的常规预期，表明我们的方法在已有模型性能基础上能够提供合理且稳定的性能增益。

然而，进一步对比Simple Baselines与HRNet两种不同网络结构，我们观察到增益幅度并不与模型自身性能强弱直接相关。尽管HRNet W32和W48的基线性能均优于Simple Baselines（ResNet-152），但在这两类高性能网络中，我们的方法依然实现了更大的性能提升（分别为0.7 mAP和0.5 mAP），超过了Simple Baselines（ResNet-152）对应的增益（0.4 mAP）。这一结果表明，我们所提出的框架在不同网络结构下均具有良好的适应性与增益效果，并未因基线性能的提升而失效，进一步验证了方法的有效性与通用性。

4.4　可视化分析

为进一步验证所提结构引导多步优化框架在推理过程中的结构修正能力与关键点恢复能力，本文从两种可视化角度进行定性分析：一方面，通过热图演化过程直观展示模型在结构优化和关键点激活方面的动态变化；另一方面，通过最终姿态估计结果与真实标注的对比，展示模型在复杂遮挡、多人交互等高难度场景下的结构鲁棒性和关键点恢复能力。

本节首先分析基于热图的结构优化过程（4.4.1，4.4.2），随后在4.4.3节进一步补充了多组复杂场景下的姿态估计实例。为了提高可视化效果，我们在每组热图中使用统一的颜色映射上限，并设为该组图中最大响应值，以保证热图颜色对比的连贯性。而在姿态估计实例中，我们分别使用蓝色、绿色与红色，代表Baseline（优化前）、优化后的结果与真值（Ground Truth）。

4.4.1　误检修正

如图7所示，模型在某些遮挡场景中会出现将他人身体部位误判为目标关键点的情况。图7(a)展示了姿态估计网络初始输出的热图，其中出现了错误的脚部激活区域；图7(b)-(d)展示了在结构引导下逐步消除错误响应的过程，最终在图7(d)中形成清晰、准确的热图聚焦区域；图7I则通过骨架图显示了初始预测（蓝色）与优化后结果（绿色）的结构差异，验证了所提出方法对误检关键点的有效抑制能力。

4.4.2　缺失补全

如图8所示，在某些遮挡严重或服饰复杂的情况下，姿态估计网络的初始结果可能无法识别出某些关键点。图8(a)中人物左腿完全被裙摆遮挡，初始热图中无有效响应（响应值低于阈值）；在结构引导的多步演化下，模型逐步强化对应区域的响应，最终在图8(d)中显著激活目标关键点位置。图8I骨架图清晰展示了从“完全缺失”到“结构恢复”的演化过程，进一步验证了结构引导的激活能力。

需要指出的是，在部分样例中，结构引导过程还展现出对错误关键点“从有到无”的能力，即抑制真值（Ground Truth）中未标注的关键点响应。然而，由于此类关键点不计入评估指标（如mAP），因此其抑制行为虽具合理性，但不影响量化结果，本文不再另行展示。

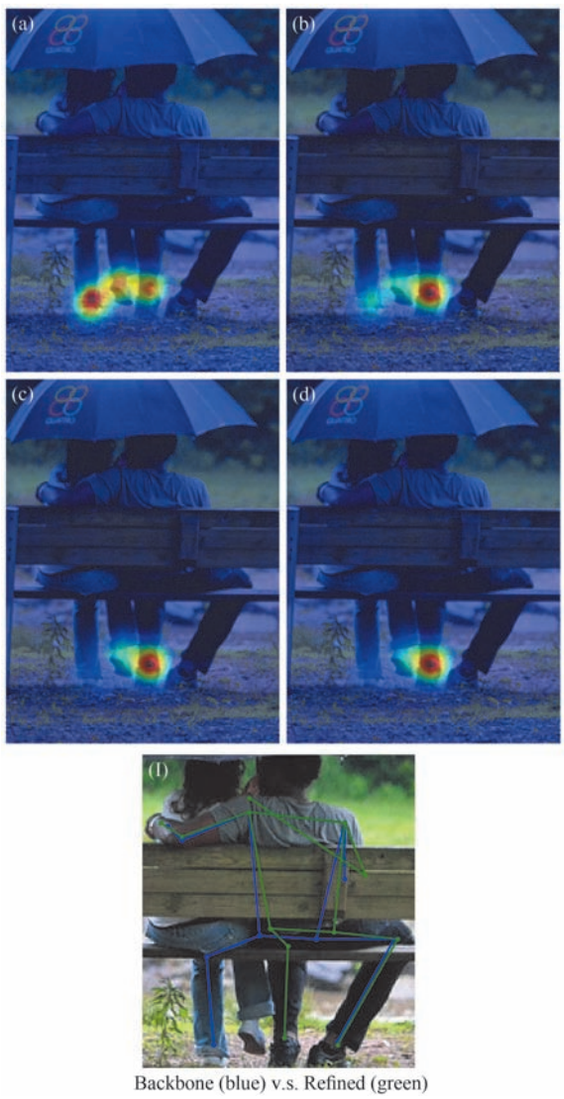

图7　结构引导多步优化过程中的误检修正示例（(a)-(d)展示关键点热图的逐步优化过程。I展示初始估计与优化后姿态的骨架结构对比。蓝色为姿态估计网络的输出，绿色为优化结果。）

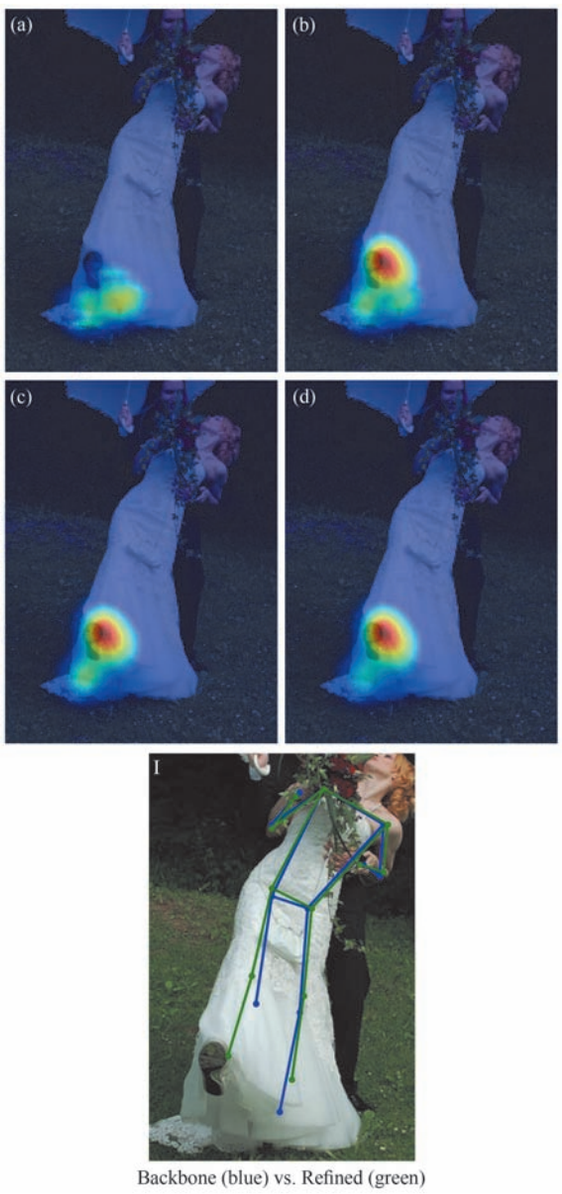

图 8 结构引导多步优化过程中的关键点激活示例((a)-(d)展示关键点从无到有的激活过程。I展示初始估计与优化后姿态的骨架结构对比。)

4.4.3 复杂场景下的其他实例

为进一步展示结构引导多步优化框架在更多复杂条件下(比如低分辨率、强遮挡以及运动模糊等)、实际应用场景中的结构一致性与鲁棒性,我们选取了三组具有代表性的复杂样例,展示Baseline方法与所提方法在最终骨架预测结果上的对比。

如图9所示,在低分辨率和多人遮挡场景下,Baseline方法(蓝色)将目标人物的一只脚关键点误判,导致骨架结构严重错位。我们的方法(绿色)通过结构引导多步优化过程,准确修正了关键点位置,并且优化了骨架结构,使其更接近真实标注(红色),展现出强大的结构恢复能力。图10中,在餐桌遮挡和儿童姿态变化显著的样例中,Baseline方法无法检测出被遮挡的左膝关键点,导致下肢骨架断裂。我们的方法能够显著补全缺失的关键点,骨架结构完整,与真实标注高度接近,验证了模型在关键点缺失情况下的补全与鲁棒性。而图11则展示了在摩托车骑行等装备遮挡场景下,Baseline方法无法正确识别下肢关键点,骨架连线错位严重。所提方法

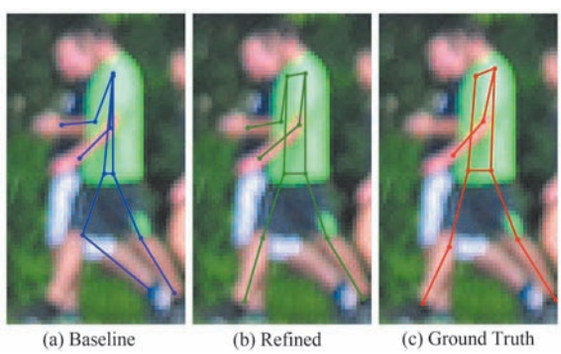

图 9 低分辨率的多人场景下的(a)优化前(Baseline)、(b)由本方法优化后(Refined)、(c)真值(Ground Truth)

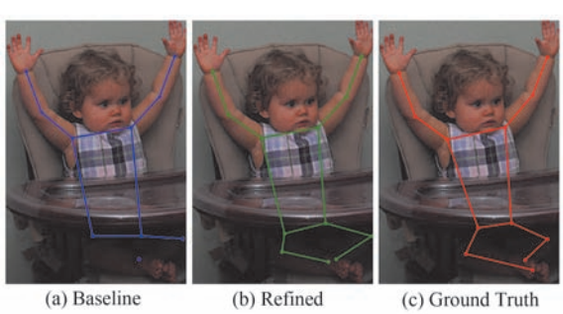

图 10 儿童被桌子遮挡场景下的(a)优化前(Baseline)、(b)由本方法优化后(Refined)、(c) 真值(Ground Truth)

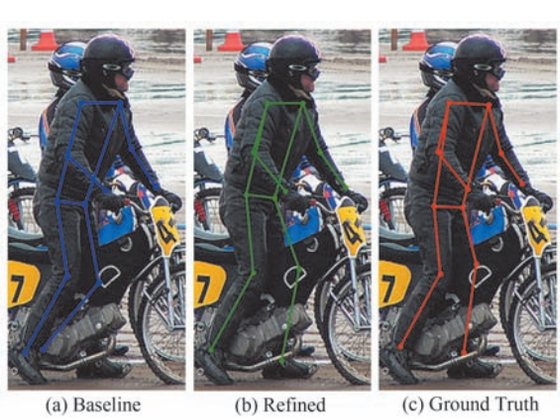

图 11 摩托车遮挡且多人场景下的(a)优化前(Baseline)、(b)由本方法优化后(Refined)、(c)真值(Ground Truth)。

有效恢复了被遮挡的腿部结构和关键点位置,提升了整体结构一致性和鲁棒性。

这些典型实例直观展现了结构引导多步优化框架在实际复杂场景下的结构恢复能力、关键点补全效果与鲁棒性提升。更多复杂场景下的可视化分析结果见附录A中的图12～图16。

4.5　与现有主流方法的对比实验

4.5.1　CrowdPose数据集上的对比

在CrowdPose验证集上，我们对所提结构引导多步优化框架进行了系统评估，结果如表3所示。对比方法涵盖了基于CNN的传统架构（如HRNet、HRNeXt）、图神经网络方法（如OPEC-Net、DGN），以及近期提出的Transformer方法（如TransPose-H/A6、ED-Pose、RTMO-L）。在统一输入分辨率与真值框（ground truth bounding box）条件下，本文方法在ResNet-152和HRNet-W48两种主干网络下分别达到74.8和77.6 mAP。其中，基于HRNet-W48的结果不仅显著超过所有现有方法（包括主流Transformer与GNN架构），相比第二高的TransPose-H/A6（76.3 mAP）提升1.3 mAP，相比ED-Pose（73.1 mAP）、RTMO-L（73.2 mAP）等最新Transformer方法，提升也非常显著。与第三高的GNN方法DGN（72.4 mAP）相比，优势超过5 mAP，进一步体现了所提结构引导多步优化机制在复杂场景中的结构建模能力与精度优势。

表3　CrowdPose验证集上不同方法的性能对比

方法	Backbone	架构类型	AP (mAP)
HigherHRNet[23]	HRNet-W32	CNN-based	65.9
CrowdPose Baseline[19]	HRNet-W32	CNN-based	66.0
DEKR[55]	HRNet-W32	CNN-based	68.0
HRNeXt[52]	HRNeXt	CNN-based	70.4
OPEC-Net[52]	ResNet-101	GNN-based	70.6
I^2R-Net[53]	HRNet-W48	CNN-based	72.3
DGN[50]	HRNet-W48	GNN-based	72.4
ED-Pose[12]	Swin-L [64]	Transformer-based	73.1
RTMO-L[59]	CSPDarknet	CNN-based	73.2
TransPose-H/A6[35]	HRNet-W48	Transformer-based	76.3
Ours (ResNet)	ResNet-152	CNN-based	74.8
Ours (HRNet)	**HRNet－W48**	**CNN－based**	**77.6**

注：本表展示了CrowdPose验证集上不同方法的性能对比，所有方法在统一输入分辨率下进行评估。本文方法基于真值框（ground truth bounding box）测试，采用ResNet-152和HRNet-W48两种主干网络，均取得了优异性能。

这些结果表明，所提结构引导多步优化模型在多人遮挡、姿态交错等复杂场景中具备强鲁棒性和高效建模能力，为密集场景下的人体姿态估计提供了新的性能提升思路。值得一提的是，尽管部分对比方法可能采用更高输入分辨率（如OPEC-Net为320×256），本文仍在256×192这一所有比较方法中最低分辨率设定下取得领先成绩，进一步体现了模型在结构建模层面的有效性与泛化能力。

4.5.2　COCO-Keypoints数据集上的对比

在COCO验证集上，我们进一步比较了所提方法与主流姿态估计方法在256×192分辨率与检测器框条件下的性能，结果如表4所示。对比方法涵盖了CNN-based（如HRNet、SimpleBaseline、RTMPose）以及Transformer-based（如HRFormer-B、ViTPose-B、TransPose-H/A6）等多种代表性结构。所提方法在ResNet-152和HRNet-W48两种骨干网络下分别取得了74.2 mAP和75.6 mAP，其中在HRNet-W48下的结果与当前最佳的Transformer方法HRFormer-B持平，略低于TransPose-H/A6和ViTPose-B（均为75.8 mAP）。同时，所提方法也优于主流CNN方法RTMPose-m/l和HRNet-W48。

表4　COCO验证集（256×192）下不同方法的性能对比

方法	Backbone	架构类型	AP (mAP)
SimpleBaseline[21]	ResNet-152	CNN-based	73.5
RTMPose-m[58]	CSPNeXt-m	CNN-based	73.6
HRNet[24]	HRNet-W32	CNN-based	74.4
RTMPose-l[58]	CSPNeXt-l	CNN-based	74.8
HRNet[24]	HRNet-W48	CNN-based	75.1
TransPose-H/A4[35]	HRNet-W48	Transformer-based	75.3
HRFormer-B[53]	HRFormer-B	Transformer-based	75.6
TransPose-H/A6[35]	HRNet-W48	Transformer-based	75.8
ViTPose-B[57]	ViT-modified	Transformer-based	**75.8**
Ours (ResNet)	ResNet-152	CNN-based	74.2
Ours (HRNet)	HRNet-W48	CNN-based	75.6

值得强调的是，所提方法完全基于CNN架构，在不引入Transformer模块的前提下即可实现与最新Transformer方法接近或持平的性能，验证了结构引导多步优化机制在COCO数据集下的稳定性与结构建模能力。进一步地，尽管本文方法在CrowdPose数据集上优于TransPose-H（A6配置），但在COCO数据集上的表现略逊于该方法。结合两个数据集的差异，尤其是COCO对遮挡关键点的标注较弱，我们推测，评估结果的这一差异可能与结构建模类方法在COCO上受限于标注准确性有关。

类似现象也在DGN[50]与OPEC-Net[52]等工作中有所指出。

5 结 论

本文提出了一种结构引导多步优化框架，用于提升2D人体姿态估计的结构建模能力。该框架通过构造从初始姿态到目标姿态的插值热图序列，引导多步优化过程逐步逼近结构合理的估计结果，并在训练与推理阶段保持引导路径的一致性，从而显著增强了模型的稳定性与泛化能力。

在COCO与CrowdPose数据集上的实验验证了所提方法的有效性。其中，在CrowdPose数据集上，基于HRNet-W48的模型在256×192分辨率下取得了77.6 mAP，超过目前已知的所有公开方法。在COCO验证集上，所提方法在256×192分辨率和预测框设定下亦达到75.6 mAP，与多种代表性Transformer架构方法持平或更优，展现出良好的结构建模能力与稳健性。

尽管本文主要基于被广泛使用的Simple Baselines以及HRNet这两种基于CNN的姿态估计网络进行验证，实验结果表明该结构引导多步优化框架可与更复杂的姿态估计网络结构（包括基于Transformer架构的）兼容，具备进一步提升这些网络表现的潜力。更为重要的是，本文提出的“结合任务特性设计结构引导与多步优化机制”的思路，可为结构优化与约束类任务提供具有普适性的建模参考。未来可在结构约束、优化路径或语义引导等方面，构建更具针对性的多步优化方式，进一步拓展该方法在复杂结构建模任务中的适用范围与表达能力。

此外，基于结构引导的多步优化机制本身具备良好的可扩展性。在更复杂的现实场景中，例如密集人群、快速运动引起的模糊、极端光照变化，以及多模态融合等更具挑战性的设置下，该机制在结构恢复、关键点补全与全局一致性维护方面仍具有进一步发挥的潜力。未来可结合视频时间信息、跨尺度特征或跨模态感知（如深度图像或事件相机数据），探索该框架在更广泛实际环境中的应用能力。

参 考 文 献

[1] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA, 2014: 1653-1660

[2] Martinez J, Hossain R, Romero J, Little JJ. A simple yet effective baseline for 3D human pose estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2640-2649

[3] Zhang S, Wang L, Wang G, et al. OPEC-Net: Towards occlusion-aware pose estimation and crowd understanding//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea, 2019: 7241-7250

[4] Girdhar R, Ramanan D. CATER: A diagnostic dataset for compositional actions and temporal reasoning//Proceedings of the International Conference on Learning Representations (ICLR). Virtual, 2020

[5] Shen S, Tan J. iMapper: A deep learning framework for interactive mapping in human-robot collaboration//Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China, 2021: 11315-11322

[6] Ravindra V, Castellano K, Krishnan S. A deep learning approach for assessing therapy progress in stroke rehabilitation. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2021, 29: 1234-1243

[7] Xu W, Huang Q. Real-time 3D hand pose estimation for virtual reality interaction//Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Atlanta, USA, 2020: 343-352

[8] Li Y, Zhang S. Anomaly detection in surveillance videos using deep learning. IEEE Transactions on Image Processing. 2021, 30: 5556-5569

[9] Zhao L, Peng X. Deep learning based human pose estimation: A survey//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018

[10] Felzenszwalb P, Huttenlocher D. Pictorial structures for object recognition. International Journal of Computer Vision. 2005, 61(1): 55-79

[11] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA, 2011: 1385-1392

[12] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, Netherlands, 2016: 483-499

[13] Wei SE, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 4724-4732

[14] Xiao B, Wu H, Wei Y. Simple baselines for human pose estimation and tracking//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 466-481

[15] Li W, Wan F, Zhang X, et al. Pose estimation via improved integral regression and occlusion-aware strategies//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021: 20249-20258

[16] Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 5693-5703

[17] Li J, Wang W, Wei X, et al. SimCC: A simple coordinate classification perspective for human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 13469-13478

[18] Xu B, Wang Y, Zhang T, et al. ViTPose: Simple vision transformer baselines for human pose estimation//Advances in Neural Information Processing Systems (NeurIPS). New Orleans, USA, 2022, 35: 28309-28324

[19] Gao X, Lin K, Qian C, et al. RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv: 2303.00890, 2023

[20] Wang X, Zeng X, Wang Z, et al. RTMO: Real-time multi-person pose estimation and tracking with decoupled keypoint regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2024

[21] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 7291-7299

[22] Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2020: 5385-5394

[23] Papandreou G, Zhu T, Chen LC, et al. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based model//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 269 - 286

[24] Geng C, Sun K, Xiao B, et al. Bottom-up human pose estimation via disentangled keypoint regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021: 14676-14686

[25] Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping//Advances in Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017

[26] Jin S, Liu W, Xie E, Wang W, Qian C, Ouyang W, Luo P. Differentiable hierarchical graph grouping for multi-person pose estimation //Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 718-734

[27] Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 11977-11986

[28] Luo Z, Wang Z, Huang Y, Wang L, Tan T, Zhou E. Rethinking heatmap regression for bottom-up human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021: 13264-13273

[29] Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B. DeepCut: Joint subset partition and labeling for multi-person pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 4929-4937

[30] Nie X, Feng J, Jin X, Yan S. Single-stage multi-person pose machines//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Republic of Korea, 2019: 6951-6960

[31] Maji D, Nagori S, Mathew M, Poddar D. YOLO-Pose: Enhancing YOLO for multi-person pose estimation using object keypoint similarity loss//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, USA, 2022

[32] Wang Y, Zhang Z, Li X, et al. Graph-PCNN: Two-stage human pose estimation with graph pose refinement//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK, 2020: 492-508

[33] Andriluka M, Roth S, Schiele B. Pictorial structures revisited: People detection and articulated pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA, 2009: 1014-1021

[34] Li Z, Liu J. TokenPose: Learning keypoint tokens for human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 20249-20258

[35] Yang S, Quan Z, Nie M, Yang W. TransPose: Keypoint localization via transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 11835-11844

[36] Guo H, Wang J. Symmetrical-physical GCNs for hand pose estimation. Journal of Computer Research and Development. 2023, 60(7)

[37] Shi L, Zhang Y, Cheng J, et al. Skeleton-based action recognition with directed graph neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 7912-7921

[38] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on Machine Learning (ICML). Lille, France, 2015: 2256-2265

[39] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models//Proceedings of the 34th Advances in Neural Information Processing Systems (NeurIPS). Virtual, 2020

[40] Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. arXiv:2102.09672, 2021

[41] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752, 2022

[42] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487, 2022

[43] Wang P, Zhang Y, Fu Y, Liu W. Towards realistic face photo-sketch synthesis via composition-aided diffusion models. arXiv: 2203.13817, 2022

[44] Amit R, Goyal A. SegDiff: Image segmentation with diffusion probabilistic models. arXiv:2112.00390, 2021

[45] Meng C, Song Y, Song J, Wu J, Zhu JY, Ermon S. SDEdit: Guided image synthesis and editing with stochastic differential equations. arXiv:2108.01073, 2021

[46] Saharia C, Ho J, Chan W, Salimans T, Fleet DJ, Norouzi M. Palette: Image-to-image diffusion models. arXiv: 2111.05826, 2021

[47] Ho J, Jain A, Abbeel P. Video diffusion models. arXiv: 2204.03458, 2022

[48] Tevet G, Amit H, Savin H, Dekel T. Motion diffusion model: Generating human motions from text. arXiv:2208.15001, 2022

[49] Kong Z, Ping W, Huang J, Zhao K, Catanzaro B. DiffWave: A versatile diffusion model for audio synthesis. arXiv:2009.09761, 2020

[50] Chen N, Zhang Y, Zen H, Weiss R, Norouzi M, Chan W. WaveGrad: Estimating gradients for waveform generation. arXiv:2009.00713, 2020

[51] Poole B, Jain A, Barron JT, Mildenhall B, Liu C, Tan D, Abbeel P. DreamFusion: Text-to-3D using 2D diffusion. arXiv: 2209.14988, 2022

[52] Zhao W, Gadelha M, Xu H, Hartley R, Wang R. 3D shape generation and completion through point-voxel diffusion. arXiv: 2207.09446, 2022

[53] Xu L, Cheng A, Li J, et al. HRNeXt: High-resolution network with next-generation neural architecture design//Proceedings of the European Conference on Computer Vision (ECCV). Tel Aviv, Israel, 2022: 204-221

[54] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need// Proceedings of the 31st Advances in Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 5998-6008

[55] He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 16000-16009

[56] Mao W, Ge Y, Shen C, Tian Z, Wang X, Wang Z, van den Hengel A. Poseur: Direct human pose regression with transformers //Proceedings of the European Conference on Computer Vision (ECCV). Tel Aviv, Israel, 2022: 72-88

[57] Yang J, Zeng A, Liu S, Li F, Zhang R, Zhang L. Explicit box detection unifies end-to-end multi-person pose estimation// Proceedings of the International Conference on Learning Representations (ICLR). Kigali, Rwanda, 2023

[58] Shi D, Wei X, Li L, Ren Y, Tan W. End-to-end multi-person pose estimation with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA, 2022: 11069 - 11078

[59] Zhao L, Peng X, Tian Y, et al. Semantic Graph Convolutional Networks for 2D Human Pose Estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA, 2021

[60] Tu J, Wu G, Wang L. Dual graph networks for pose estimation in crowded scenes. International Journal of Computer Vision. 2024, 132: 633-653

[61] Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502, 2020

[62] Zheng C, Yang Y, Yu Y, Dai Q. Structured 3D human body reconstruction from single images with diffusion models. arXiv: 2201.07738, 2022

[63] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Munich, Germany, 2015: 234-241

[64] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical vision transformer using shifted windows//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada, 2021: 10012-10022

附录A.

这里展示了补充的复杂场景样例，一共五组（图12～图16），从不同方面展示了在多人遮挡、姿态交错等复杂场景中具备强鲁棒性和高效建模能力。

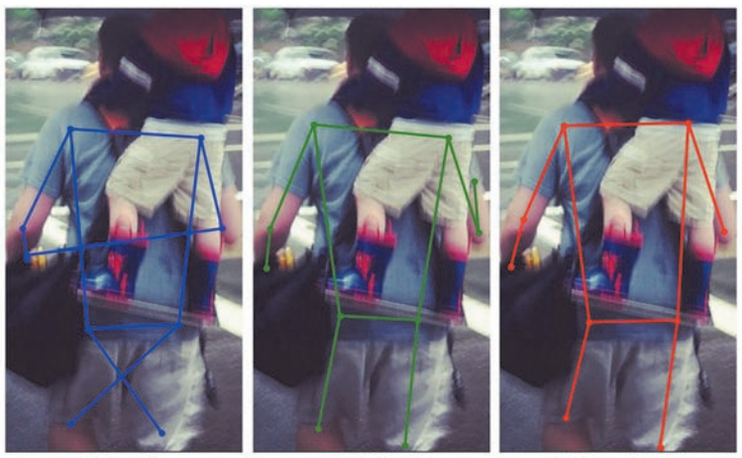

图12　运动模糊及被人遮挡的多人场景下的(a)优化前(Baseline)、(b)优化后(Refined)、(c)真值(Ground Truth)

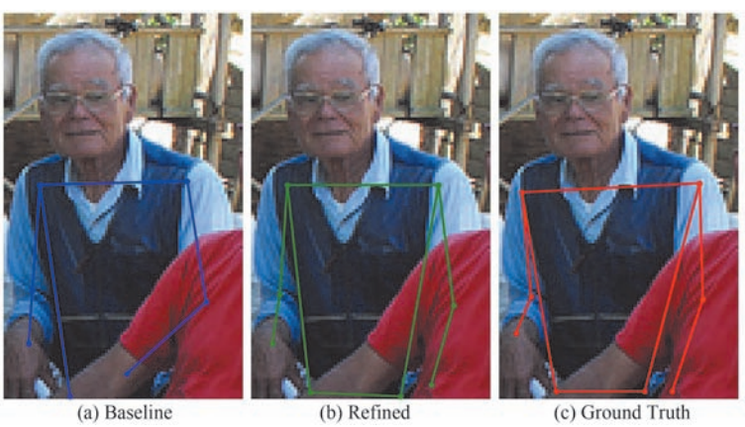

图13　人物左臂(位于图像右侧)被遮挡的场景下的(a)优化前(Baseline)、(b)优化后(Refined)、(c)真值(Ground Truth)

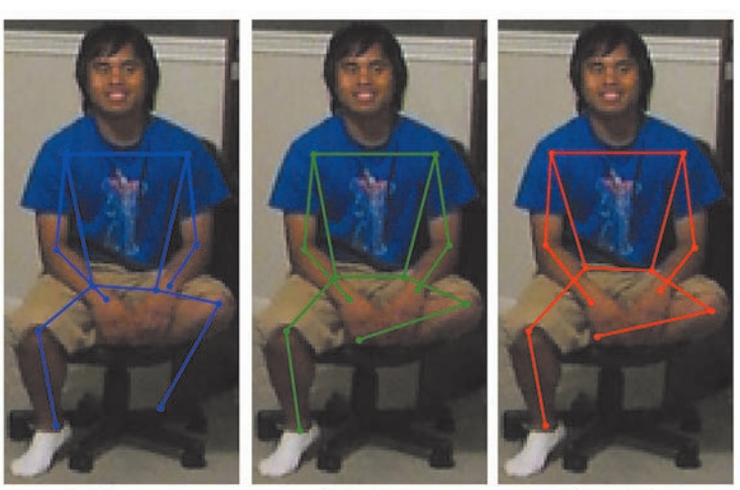

图14　坐姿、双手遮挡小腿的场景下的(a)优化前(Baseline)、(b) 优化后(Refined)、(c)真值(Ground Truth)

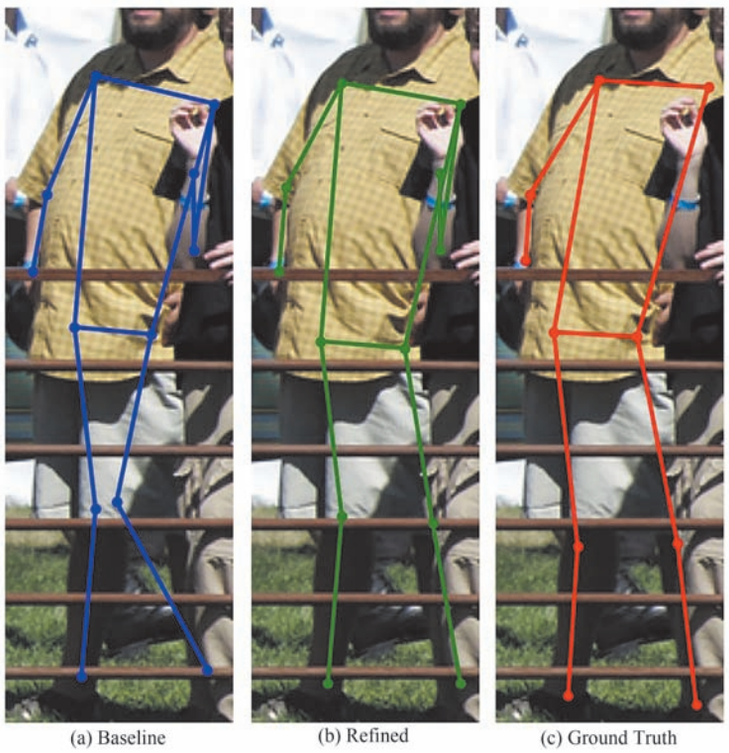

图 15 多人、物体遮挡及强光线阴影场景下的(a)优化前(Baseline)、(b) 优化后(Refined)、(c)真值(Ground Truth)

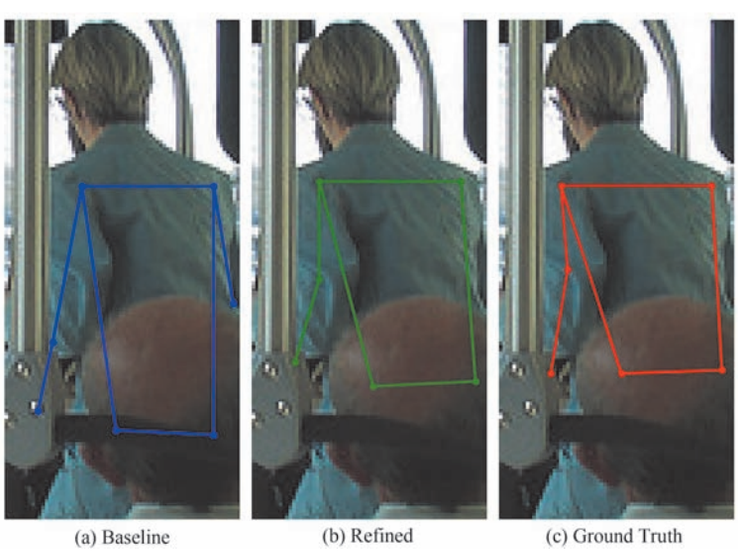

图 16 主体人物被他人头部和设备严重遮挡,髋部(胯部)等关键点完全无法直接观测,只能依靠结构约束进行推断:(a)优化前,髋部预测与真值差距较大(Baseline);(b)优化后,基于结构引导优化,接近真值(Refined);(c)真值(Ground Truth)

TU Jun, Ph. D. candidate. His major research interests include human pose estimation, human pose tracking, and gait recognition.

WU Gang-Shan, Ph. D., professor, Ph. D. supervisor. His major research interests include media content analysis, multimedia information retrieval, etc.

WANG Li-Min, Ph. D., professor, Ph. D. supervisor. His major research interests include video understanding, action recognition, etc.

Background

Human pose estimation is a fundamental task in computer vision, aiming to localize human body keypoints from images or videos. As a core component in human-centric visual understanding, pose estimation supports a wide range of downstream applications such as action recognition, motion analysis, human-computer interaction, and animation synthesis. Despite significant progress in 2D human pose estimation under controlled settings, real-world performance is still limited by challenges such as occlusion, complex body configurations, and structural ambiguity, particularly in crowded or low-resolution environments. These difficulties highlight the continuing gap between academic progress and the demands posed by practical deployment scenarios.

While recent approaches leverage deep convolutional and Transformer-based architectures to improve keypoint localization, most existing methods treat pose estimation as a regression or detection problem, often neglecting the intrinsically structured nature of human pose. Some works attempt to encode structural priors using graph neural networks or intermediate representations such as heatmaps, yet the integration of explicit structural information with mainstream architectures remains insufficient and lacks flexibility. As a consequence, the predicted poses may satisfy local accuracy but still violate global anatomical plausibility, especially when the visual evidence is incomplete or noisy. This limitation becomes more pronounced in challenging multi-person settings, where interactions and mutual occlusions further complicate keypoint inference.

To address these limitations, this paper introduces a structure-guided multi-stage optimization framework for human pose estimation. Instead of treating pose prediction as a one-step regression, our method models the estimation process as a sequence of structure-guided refinements, progressively transforming an initial coarse prediction into a structurally plausible and accurate pose. The core of our approach is the use of intermediate guidance—explicitly constructed sequences of structure-aware intermediate representations (gradient heatmaps or 'guidance maps')—that direct each optimization stage toward maintaining structural consistency. Crucially, the framework enforces consistency in guidance paths between training and inference, reducing discrepancies and improving generalization across diverse scenarios.

Extensive experiments on COCO and CrowdPose datasets demonstrate the effectiveness of the proposed framework. On CrowdPose, our approach achieves 77.6 mAP with an HRNet-W48 backbone, surpassing existing state-of-the-art methods. On COCO, it attains 75.6 mAP, showing strong generalization ability across different image conditions. The framework is compatible with various backbone architectures and consistently improves baseline methods by enhancing structural modeling and keypoint localization accuracy.

In summary, this research advances the field of human pose estimation by explicitly modeling and enforcing structural consistency throughout the pose prediction process. By integrating structure-guided multi-stage optimization, the proposed framework bridges the gap between data-driven learning and structural priors, offering a robust and generalizable solution for challenging pose estimation scenarios.

This work is supported by the National Key. RD Program of China (No. 2022ZD0160900), and Jiangsu Provincial Natural Science Foundation Climbing Project (No. BK20250009).