基于单目视频的全域协同人体姿态估计算法

刘 鹏[1,2] 周平平[1,2]
LIU Peng ZHOU Pingping

摘 要 受深度信息缺失影响，人体二维姿态提升到三维姿态具有极大的挑战性。如何有效协同输入信息的时空特征成为解决该挑战性问题的关键。为此，文章设计实现了一种全新的算法，称为全域协同 Mamba 网络（GCMam-Net），旨在充分利用 Mamba 网络选择性状态空间的特性和注意力机制提取全局上下文的优势，从全局与局部两个视角对姿态特征进行融合。通过设计，实现了 Tem-AG 模块、Spa-Mam 模块协同提取时空间信息，还将 TS-Fus 模块对姿态相关特征进行整理融合。经过严格的消融实验，所提出的 GCMam-Net 算法在两个基准数据集：Human 3.6M 和 MPI-INF-3DHP 上取得了优异的结果，证明了方法的有效性和泛化性。

关键词 人体姿态估计；协同提取；Mamba 网络；注意力机制

doi：10.3969/j.issn.1672-9528.2026.02.018

0 引言

3D 人体姿态估计，是研究人类自身知识体系、构建人体标识的一个智能研究方向，在运动分析 [1]、行为识别 [2]、人机交互 [3] 等多个领域有着广泛的应用。二维到三维提升方法是该领域最流行的技术，具体为将输入的图片或视频经过二维人体姿态估计网络得到人体 2D 关节点，再使用 3D 姿态网络将 2D 关节点提升到 3D 关节坐标。本文的工作仍采用二维到三维提升方法。

在关节点维度提升的过程中，人体姿态面临着固有的不适定性问题。从人体二维骨架回归到人体三维骨架需要经过空间映射，而同一个二维骨架拥有多个合理的不同三维关节坐标。这使得姿态估计具有极大的挑战性。此外，受个人身高、体重等变化，输入信息拥有极大的不确定性，这也影响了三维关节点的估计准确度。

为应对这些挑战，深度学习方法取得了重大的进展。通过不断优化特征提取机制和特征融合策略，一些最新的深度学习方法已能取得较好的性能。基于 GCN 的方法 [4-6] 通过对人体关节点拓扑结构施加约束，以使物理上相邻的两个或多个关节点具有邻域信息，从而更好地估计姿态关节点的位置。基于注意力机制的方法 [7-9] 利用其特性对姿态关节进行全局的特征提取，帮助算法获得姿态全局的特征信息。

以人体物理拓扑结构为基础的方法和对全局姿态特征进行提取的方法都取得了良好的模型效能，但仍然存在以下问题：

（1）全局与局部姿态特征协同不足。全局信息被作为主要特征应用时，局部信息的精确性不足；局部信息被作为主要特征应用时，缺乏全局调整的能力。

（2）空域与时域特征协同不足。空域信息作为主要特征应用时，时域信息包含的丰富内容被过分忽略，在以长时序为输入的姿态信息序列中，这显然难以取得最好的估计结果。以时域信息作为主要特征应用时，姿态的空间信息又无法达到精细的效能。

（3）当前算法的误差积累仍然较高，还需要新的算法来提高姿态估计任务的准确度。

基于此，本文引入计算机视觉领域的 Mamba 模型 [10]，利用其选择性机制和硬件感知算法，实现对局部空间特征信息的精准提取，并以 Mamba 核心架构为基础，设计了 Spa-Mam 模块。此外，结合人体拓扑姿态信息，与注意力机制协同，设计了全局时间提取模块 Tem-AG，从姿态长序列信息中捕获全局与局部的信息依赖。最后利用时空特征融合模块 TS-Fus 将全域信息进行深度捕获，并以此概念命名了本文提出的模型：全域协同 Mamba 网络（GCMam-Net）。在 Human3.6M[11] 和 MPI-INF-3DHP[12] 两个基准数据集上对本文算法进行了全面的消融实验，实验结果表明，本文取得了较为先进的性能，同时具有较好的泛化能力。

总的来说，工作的主要贡献如下：

（1）引入 Mamba 架构，设计 Spa-Mam 模块提取姿态的状态空间，引入新颖的局部空域状态概念。

（2）设计实现 Tem-AG 模块，捕获全局与局部姿态特征信息，共同提取姿态长序列中隐含的人体姿态拓扑信息。

（3）提出全域协同建模，构建全局与局部、空间和时间双向的协同融合，以此引入全域协同 Mamba 网络（GCMam-

1. 重庆移通学院 重庆 401420
2. 公共大数据安全技术重庆市重点实验室 重庆 401420

Net），并通过大量的实验结果证明了该算法的有效性与泛化性。

1 模型设计

使用现成的 CPN 网络 [13]，从长序列姿态信息中提取人体 2D 关节点，再以 2D 关节点作为输入传输到全域协同 Mamba 网络（GCMam-Net）中。提出的架构共细分为 3 个模块：Spa-Mam 模块、Tem-GC 模块以及 TS-Fus 模块，通过逐级残差连接，将 3 个模块共同构成了一个基本块。再通过 M=20 个基本块串形连接构成了架构的主体。模型架构如图 1 所示。

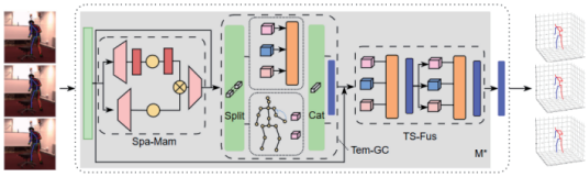

图 1 模型架构图

对于初始长序列输入 $P=\{p_{t,j}\in\mathbb{R}^2 \mid t=1,2,\cdots,T;\ j=1,2,\cdots,N\}$，先将其通过 Spa-Mam 模块获得姿态状态空间。而后结合初始信息将混合特征输入 Tem-AG 模块，从全局和局部视角对混合特征进行提取。最终使用 TS-Fus 模块对当前提取出来的多特征依赖进行捕获，并将其与初始特征一同输入到下一层。该过程表示为：

$$\begin{aligned} P' &= \text{Spa}-\text{Mam}(P) \\ P'' &= \text{Tem}-\text{AG}(P+P') \\ \overline{P} &= \text{TS}-\text{Fus}(P+P'') \end{aligned} \tag{1}$$

在通过 M=20 个基本块后，使用一个线性回归头来回归最终的人体三维姿态关节点坐标 $\boldsymbol{Z}\in\mathbb{R}^{T\times J\times 3}$。通过协同全域时空特征、全局与局部特征之间的平衡，使算法积累更低的误差，从而达到更好的估计效果。

2 实验设计

2.1 数据集和评价指标

Human3.6M 数据集 [11] 是三维人体姿态估计领域广泛流行的一个基准数据集，包含 11 位演员的 15 个不同动作，共收录了 360 万帧率的捕捉数据。使用 Human3.6M 数据集作为基准数据集，以第 1、5、6、7、8 位演员的捕捉数据作为训练数据，以 9 和 11 位演员的捕捉数据作为测试数据，来评估架构的有效性。使用平均每关节位置误差 MPJPE(Protect #1) 和对齐后的平均每关节位置误差 P-MPJPE(Protect #2) 来评估本文方法。此外，MPI-INF-3DHP 数据集被本文设计为泛化性测试数据集，主要包含更多样的户外场景和动作类别。MPI-INF-3DHP 数据集 [12] 使用正确关键点百分比（PCK）的 3D 扩展和曲线下面积（AUC）作为主要的评价指标。

2.2 实现细节

本文在单张 NVIDIA RTX 4090 GPU 上完成了整个方法的训练、测试和消融全过程。单次训练 + 评估平均花费时间 12.5 min，模型训练次数为 45 次。网络整体采用 AdamW 优化器，设置 GeLU 函数为默认激活函数，单批次输入大小为 12。遵循以往的方法，本文使用级联金字塔网络（CPN）[13] 对数据集进行基准测试。

2.3 损失函数

本文使用平均每关节位置误差（MPJPE）来最小化估计结果和真实关节点位置之间的误差。损失函数表示为：

$$\ell = \frac{1}{\omega}\sum_{i=1}^{\omega}|G_i - E_i|^2 \tag{2}$$

式中：ℓ代表损失；G_i 代表第 i 个关节点的真实关节点位置；E_i 代表第 i 个关节点的估计结果。

3 结果分析

本文对全域协同 Mamba 网络（GCMam-Net）进行了全面的消融实验，同时与其他先进的方法进行了多方面的比较，最后展示了该架构的可视化结果。

3.1 与先进方法的比较

Human3.6M 的比较。本文将本文提出的全域协同 Mamba 网络（GCMam-Net）与其他先进的方法进行了比较，如表 1~2 所示。

表 1 与目前先进方法的 P1 损失定量比较

Protocol #1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.
HSTFormer[14]	39.5	42.0	39.9	40.8	44.4	50.9	40.9	41.3
UpliftandUp[15]	38.6	41.0	**37.6**	39.7	44.2	**47.9**	40.9	39.8
ConvFormer[16]	41.0	43.2	39.0	42.4	44.5	52.2	41.7	40.8
GLSTE[17]	38.1	41.7	39.1	39.5	43.7	49.7	40.4	39.7
GCMam-Nett（Ours）	**36.6**	**39.5**	38.1	**34.4**	**42.1**	51.0	**37.7**	**37.1**
Protocol #1	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
HSTFormer[14]	54.7	58.8	43.6	40.7	43.4	30.1	30.4	42.7
UpliftandUp[15]	51.7	60.3	43.1	41.1	41.6	28.4	29.2	41.7
ConvFormer[16]	53.0	60.6	44.8	41.3	43.7	29.6	30.9	43.2
GLSTE[17]	54.2	58.7	42.8	42.1	42.0	28.7	28.5	41.9
GCMam-Net（Ours）	**50.6**	**55.5**	**42.2**	**38.3**	**38.1**	**27.2**	**27.5**	**39.7**

表 2 与目前先进方法的 P2 损失定量比较

Protocol #2	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.
UpliftandUp[15]	31.6	33.7	31.8	33.3	34.7	**38.7**	32.2	31.2
HSTFormer[14]	31.1	33.7	33.0	33.2	**33.6**	38.8	31.9	31.5
ConvFormer[16]	31.4	34.2	**32.0**	35.2	34.0	40.3	32.7	31.3
GLSTE[17]	**30.8**	34.6	32.4	32.9	34.0	39.5	31.5	**31.0**
GCMam-Net（Ours）	31.0	**33.6**	32.6	**29.7**	35.2	40.1	**31.4**	31.6

Protocol #2	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
UpliftandUp[15]	**41.9**	48.9	35.5	32.6	33.7	23.4	24.0	33.8
HSTFormer[14]	43.7	**46.3**	35.7	31.5	33.1	24.2	24.5	**33.7**
ConvFormer[16]	42.6	49.0	36.2	**31.3**	34.8	23.4	24.9	34.2
GLSTE[17]	44.2	48.7	**35.2**	32.4	33.5	**23.1**	**23.5**	33.8
GCMam-Net（Ours）	42.9	48.8	36.8	31.5	**32.8**	**23.1**	23.7	**33.7**

此外，本文将真实二维关节点坐标作为输入数据对模型进行训练，以评估模型在不同输入数据中的泛化性，与其他方法的比较结果如表 3 所示。全域协同 Mamba 网络（GCMam-Net）取得了 22.4 mm 的平均关节误差，在所有的 15 个测试动作中都取得了最优效果。该结果证明了本文提出的架构面对不同输入都能取得较为优异的精度，体现了该架构的泛化性。

表 3 以 2D 真实坐标为输入的方法 P1 定量比较

Protocol #1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.
MHFormer [9]	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2
HSTFormer[14]	24.9	27.4	28.1	25.9	28.2	33.5	28.9	26.8
ConvFormer[16]	28.9	31.8	28.0	28.2	29.5	33.0	32.9	30.1
GLSTE[17]	27.8	29.5	27.1	26.5	27.4	31.0	30.8	27.5
GCMam-Net（Ours）	**20.4**	**22.9**	**20.2**	**21.2**	**21.3**	**25.0**	**23.9**	**22.5**
Protocol #1	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg.**
MHFormer [9]	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
HSTFormer[14]	33.4	38.2	27.2	26.7	27.1	20.4	20.8	27.8
ConvFormer[16]	36.8	37.4	29.8	29.6	28.2	21.7	21.5	29.8
GLSTE[17]	31.9	35.0	28.2	28.5	27.0	20.3	20.3	27.9
GCMam-Net（Ours）	**27.6**	**29.7**	**23.2**	**21.8**	**20.9**	**13.6**	**14.5**	**21.9**

MPI-INF-3DHP 的比较。本文以 MPI-INF-3DHP 数据集对模型的跨数据集泛化能力进行测试，测试结果如表 4 所示。本文的算法能够实现 98.6% 的正确率，高于比较的其他模型，证明其具有优于其他算法的泛化性。

表 4 在 MPI-INF-3DHP 数据集上的性能比较

Method	T	PCK↑	AUC↑
UpliftandUp[15]	81	95.4	67.6
STCFormer[8]	81	98.7	83.9
GLSTE[17]	81	98.2	79.0
GCMam-Net（Ours）	81	**98.6**	**88.5**

3.2 消融实验

本文在 Human3.6M 数据集上对本文的全域协同 Mamba 网络（GCMam-Net）进行了严格的消融实验，以研究该模型的极限性能。

（1）帧率设置。以不同帧率对模型进行训练，能够得到不同的估计性能。所提方法在 243 帧时性能最优，而测试其他帧率的估计效果成为必要过程。因此，本文以 81 帧、27 帧为输入，对架构进行测试，得到结果如表 5 所示。以不同帧率为输入时，架构的参数量都为 8.6×10^6，较其他大部分方法拥有更多参数。在模型效果上，以 81 帧为输入时能取得最优的结果；以 27 帧为输入时，能取得第 3 的结果。

表 5 不同帧率方法的比较

Method	Frames T	Parameters/10^6	P1
MHFormer[9]	27	18.9	45.9
STCFormer[8]	27	4.8	44.1
ConvFormer[16]	27	2.7	47.7
GLSTE[17]	27	2.4	43.7
GCMam-Net（Ours）	27	8.6	44.3
MHFormer[9]	81	19.7	44.5
STCFormer[8]	81	4.8	42.0
ConvFormer[16]	81	3.4	45.0
GLSTE[17]	81	2.5	43.1
GCMam-Net（Ours）	81	8.6	41.8

（2）模型配置。不同配置对架构参数量、积累损失会造成巨大影响。为更好地测试模型的性能，本文对不同的模型配置进行了测试。实验结果如表 6 所示。

表 6 不同配置模型的比较

Layer	Channel	Parameters/10^6	P1	P2
16	192	19.2	40.2	33.9
16	128	9.5	**39.7**	**33.7**
16	64	**2.2**	43.9	37.2
20	128	10.7	42.2	36.8
12	128	6.4	40.9	34.1
12	256	25.5	41.1	34.6

在模型层数为 16 层，通道宽度为 128 时，取得整体最优值，同时也是本文架构的最优配置。当改变通道宽度时，参数量随宽度增加而增加，两者呈正相关，但在估计精度上，都未能达到更好的效果。当改变模型层数时，参数量随层数增加而增加，也呈现为正相关，同时改变两个参数取更均衡的配置时，模型效果都不及最佳配置好。但从整体而言，不同的模型配置都能取得较为优异的效果。

3.3 定性结果

本文使用 Human3.6M 数据集第 9 和第 11 位演员的部分动作，对本文的架构进行定性分析。如图 2 所示。本文对初始输入、使用现成 2D 关节检测器 CPN 检测到的 2D 姿态和 3D 估计结果进行了视觉比较。能够看出，本文的架构在不同

动作中都能准确地估计出该环境下的 3D 人体姿势，这表明了全域协同 Mamba 网络（GCMam-Net）的有效性。

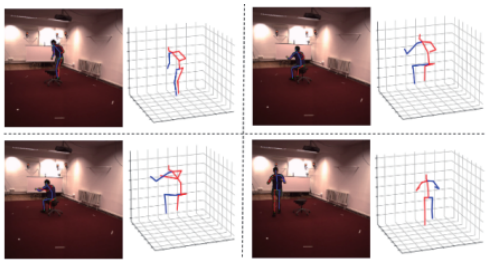

图 2 在 Human 3.6M 数据集上的可视化结果

4 结论

本文提出了一种基于单目视频的全域协同三维人体姿态估计算法，设计实现了全域协同 Mamba 网络（GCMam-Net）。通过对长时间序列隐含的状态空间特征和人体拓扑信息进行捕获，设计实现了分别用于特征提取的 Spa-Mam 模块、Tem-GC 模块和用于特征融合的 TS-Fus 模块。首次引入局部空域状态概念，用于描述姿态提取过程中的特定状态。本文在 Human3.6M 基准数据集上对本文提出的架构进行了测试，测试结果证明，本文的方法具有精确的姿态估计性能和较强的泛化性。只是在模型参数量和估计精度上仍还存有改进的空间，在未来的研究中还需要进一步精简模型的设计和争取更高的效能。

参考文献：

[1] 习会峰，骆家龙，牛佳妮，等 . 融合机器视觉与肌骨逆动力学的运动分析方法研究 [J]. 实验力学 ,2025,40(5):539-549.

[2] 张贺，翟正利 . 基于软邻接时空图卷积神经网络的动作识别算法 [J]. 信息技术与信息化 ,2023(2):183-186.

[3]YAN S, LIU M Y, WANG Y, et al. MLP: motion label prior for temporal sentence localization in untrimmed 3D Human motions[J]. IEEE transactions on circuits and systems for video technology, 2024, 34(11): 11535-11550.

[4] 丁德波，史耀群 . 基于时空图卷积网络与多层次特征融合的快递员 3D 人体姿态估计 [J]. 传感技术学报，2025, 38(8): 1457-1462.

[5] 赵扬飞，刘任波，张伟峰，等 . 面向三维人体姿态估计的特征传递平衡图卷积网络 [J]. 中国科学 : 信息科学，2025, 55(5): 1033-1050.

[6] 冉茂科，成新民 . 基于姿态估计与时空图卷积网络的扶梯危险行为识别分析 [J]. 数字技术与应用，2025, 43(2): 166-168.

[7] 邹宇，周先春，潘志庚，等 . 基于双阶时空交叉卷积 Transformer 的三维人体姿态估计方法 [J]. 电子测量与仪器学报，2025, 39(9): 159-171.

[8]TANG Z H, QIU Z F, HAO Y B, et al. 3D Human pose estimation with spatio-temporal criss-cross attention[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE, 2023: 4790-4799.

[9]LI W H, LIU H, TANG H, et al. MHFormer: multi-hypothesis transformer for 3D Human pose estimation[EB/OL].(2022-06-28)[2025-12-22].https://doi.org/10.48550/arXiv.2111.12707.

[10]GU A, DAO T. Mamba: linear-time sequence modeling with selective state spaces[EB/OL].(2024-05-31)[2025-12-23]. https://doi.org/10.48550/arXiv.2312.00752.

[11]IONESCU C, PAPAVA D, OLARU V, et al. Human3. 6M: large scale datasets and predictive methods for 3D Human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(7): 1325-1339.

[12]ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition.Piscataway:IEEE,2014: 3686-3693.

[13]CHEN Y L, WANG Z C, PENG Y X, et al. Cascaded pyramid network for multi-person pose estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2018: 7103-7112.

[14]QIAN X Y, TANG Y B, ZHANG N, et al. HSTFormer: hierarchical spatial-temporal transformers for 3D Human pose estimation[EB/OL].(2023-01-18)[2025-12-22].https://doi.org/10.48550/arXiv.2301.07322.

[15]EINFALT M, LUDWIG K, LIENHART R. Uplift and upsample: efficient 3D Human pose estimation with uplifting transformers[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway:IEEE,2023: 2903-2913.

[16]DIAZ-ARIAS A, SHIN D. ConvFormer: parameter reduction in transformer models for 3D Human pose estimation by leveraging dynamic multi-headed convolutional attention[J]. The visual computer, 2024, 40: 2555-2569.

[17]WANG Y, KANG H B, WU D D, et al. Global and local spatio-temporal encoder for 3D Human pose estimation[J]. IEEE transactions on multimedia, 2023,26: 4039 - 4049.

【作者简介】

刘鹏（1997—），男，重庆彭水人，硕士，助教、初级研究员，研究方向：计算机视觉。

周平平（1998—），女，重庆永川人，硕士，助教、初级研究员，研究方向：计算机视觉。

（收稿日期：2025-12-18 修回日期：2026-02-05）