arXiv:2602.01095v1 [cs.CV] 1 Feb 2026

PandaPose：通过将二维姿态先验传播至三维锚点空间，实现从单张图像提升三维人体姿态

Jinghong Zheng[1] Changlong Jiang[1] Yang Xiao[1,†] Jiaqi Li [1]Haohong Kuang[1] Hang Xu[1] Ran Wang[3,4] Zhiguo Cao[1] Min Du [5] Joey TianyiZhou [6,7]

[1]人工智能与自动化学院，华中科技大学 [3]新闻与信息传播学院，华中科技大学[4]未来技术学院，华中科技大学 [5]B

y字节跳动有限公司。

[6]新加坡科技研究局前沿人工智能研究中心，新加坡[7]新加坡科技研究局高性能计算研究所，新加坡

[†]通讯作者

{deepzheng,changlongj,Yang_Xiao,lijiaqi_mail}@hust.edu.cn
{haohong_kuang,hang_xu,rex_wang,zgcao}@hust.edu.cn
bingwen.ai@bytedance.com joey_zhou@a-star.edu.sg

摘要

从单幅RGB图像进行三维人体姿态提升是三维视觉中的一项困难任务。现有方法通常基于二维特征，直接建立从二维姿态到三维姿态的关节到关节映射。这种建模方式存在两个根本性局限：输入的预测二维姿态中的误差不可避免地会传播到三维预测中，而且难以处理自遮挡情况。本文提出PandaPose，这是一种通过将二维姿态先验传播至三维锚点空间，并以此作为统一中间表示来实现三维人体姿态提升的方法。具体而言，我们的三维锚点空间包含：（1）规范坐标系中的逐关节三维锚点，为缓解二维姿态估计误差提供准确且鲁棒的先验。（2）深度感知的逐关节特征提升，以分层方式整合深度信息，消除自遮挡歧义。（3）锚点-特征交互解码器，将三维锚点与提升后的特征相结合，生成统一的锚点查询向量，其中封装了逐关节三维锚点集、视觉线索和几何深度信息。随后，利用锚点查询向量促进锚点到关节集成预测。在三个成熟基准测试（即Human3.6M、MPI-INF-3DHP和3DPW）上的实验表明，我们提出的方法具有优越性。与最先进方法相比，在Human3.6M的困难条件下，误差大幅降低了14.7%；定性对比也进一步展现了我们方法的有效性和鲁棒性。

1 引言

从单幅RGB图像进行单目三维人体姿态估计，在动作识别等领域有着广泛应用 [44, 60, 5]，虚拟现实 [11] 以及人机交互 [16, 34]。与基于序列的方法相比 [27, 57, 20, 50]，基于图像的方法 [48, 54, 55] 通常更适用于实时应用，且不受输入序列长度的影响。近年来，得益于二维人体姿态估计的发展 [4, 33]，三维姿态提升研究

第39届神经信息处理系统大会（NeurIPS 2025）。

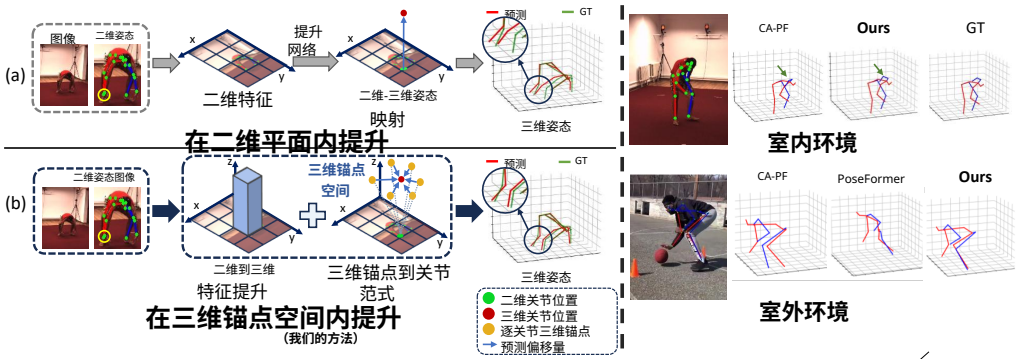


图1：不同2D到3D人体姿态提升方式的比较。以往方法（a）通常关注2D平面内特征，并直接预测三维姿态。我们的方法（b）通过将平面内特征提升至三维空间，并与逐关节三维锚点进行交互，从而缓解深度歧义。随后，在三维锚点空间内，通过锚点到关节集成预测估计三维姿态。我们的方法对室内和室外遮挡场景均具有鲁棒性。

利用输入的预测二维姿态引起了研究者的关注。然而，该领域仍面临诸多挑战，包括估计二维姿态的误差传播、高频自遮挡，等等。[53, 15]。
当前，最先进的基于图像的方法 [54, 55, 59] 尝试通过引入图像特征来补充空间上下文，以执行姿态提升，如图1（a）所示。尽管这些方法展现出良好的性能，但仍存在固有局限。首先，它们试图建立从二维姿态到三维姿态的一一映射，从而增加了对二维姿态预测质量的依赖。因此，输入姿态中的微小噪声可能导致预测三维姿态出现显著偏差，使鲁棒性受到限制。其次，由于单目性质，图像中的人体自遮挡十分常见。现有方法 [55, 50, 28] 主要依赖图像描述线索来表征三维姿态，未能显式建模深度维度，因此难以应对深度歧义和自遮挡问题。为克服这些局限，我们提出通过将二维姿态先验传播至三维锚点空间来促进三维姿态提升。因此，我们提出逐关节三维锚点设置和深度感知特征提升，以提供鲁棒的锚点初始化来抑制噪声，并融合层次化深度以消除自遮挡中的歧义，从而形成三维锚点空间，如图1（b）所示。

对于逐关节三维锚点，我们旨在将输入的二维姿态先验传播为兼具强鲁棒性和高精度的三维关节先验。这缓解了以往姿态提升方法中的误差传播问题，因为这些方法过度依赖输入姿态的准确性。具体而言，我们利用三维锚点集作为与输入姿态相关的粗略初始化，并以集成方式预测每个关节的三维锚点到关节偏移量，而非直接估计三维关节位置。先前工作中的初步探索 [42, 14] 局限于简单的全局固定锚点设置，这通常会导致锚点到关节回归偏移量过大，在对输入姿态利用不足的情况下，精度和鲁棒性均会下降。相比之下，我们提出逐关节局部锚点，以充分利用全局二维姿态上下文的指导。我们根据每个关节的二维位置，以可学习的方式自适应地在其附近设置一组锚点，力求在对输入误差的鲁棒性与初始三维先验的准确性之间取得平衡。

鉴于输入图像和姿态具有单目性质，自遮挡仍然是当前姿态提升方法面临的一项重大挑战。仅依赖平面内特征会导致遮挡区域的语义信息严重受损。在其他领域中，引入空间结构信息是恢复自遮挡区域预测结果的一种潜在解决方案 [17, 6, 51]。然而，简单引入深度图对于毫米级人体姿态估计而言仍然不足，尤其无法解决自遮挡关节中的深度歧义。因此，我们提出了一种逐关节特征提升模块，在中间关节级深度监督的引导下，将图像特征对齐至三维锚点特征域。对于监督信号，在缺乏稠密真实深度图的情况下，我们提取真实关节的深度值，而不是使用完整深度作为监督。这种方法不仅通过深度分层固有地提供了重建自遮挡关节所需的关键先决条件，还促进了以更实用的方式拟合实际的关节深度分布。在特征层面，我们利用二维姿态先验对视觉特征进行特征采样，以抵抗背景噪声，同时降低

计算内存开销。采样特征在投影到三维锚点空间之前与深度信息进行交互，从而实现面对遮挡时的稳健关节定位。
由逐关节三维锚点设置得到的初始锚点查询向量，在三维锚点空间中与深度特征和提升图像特征进行多轮基于注意力的交互。由此得到的输出查询向量融合了视觉、深度和几何锚点数据中的跨模态信息，随后被转换为针对各关节的锚点偏移和权重。这种从锚点到关节的预测机制能够生成稳健的三维姿态预测，既能抵御输入二维姿态误差，也能有效应对自遮挡。

我们在三个广为认可的数据集上验证了所提出的 PandaPose 的优越性（即，Hu- man3.6M [13], MPI-INF-3DHP [26], 3DPW [39]）。实验表明，我们的方法总体上优于所有最先进的基于图像的方法。尤其是在 Human3.6M 的挑战性场景（例如遮挡）下，我们的方法在 MPJPE 上提升了 11.3%，在 PA-MPJPE 上提升了 14.7%。总的来说，本文的主要贡献包括：

- 我们提出 PandaPose，通过将二维姿态先验传播至三维锚点空间并将其作为中间表示，解决基于图像的三维人体姿态提升问题；同时通过锚点到关节集成预测，实现有效且抗遮挡的三维姿态估计；
- 我们设计了逐关节三维锚点设置，以提供准确、稳健的三维关节映射，从而减轻输入二维姿态误差的影响；
- 我们提出了一种新颖的二维到三维特征提升方法，通过稀疏深度监督估计逐关节深度分布，以抵御自遮挡和深度歧义问题。

2 相关工作

二维到三维人体姿态提升。随着二维人体姿态估计技术的进步，将二维姿态提升至三维已成为一个重要的研究领域，相关方法通常分为基于序列的方法和基于图像的方法。基于序列的方法使用较长的时间序列，并利用图卷积网络或 Transformer 对二维姿态中的时空相关性进行建模。尽管使用更长的序列能显著提升性能，但这类方法也面临更高的计算复杂度和内存需求。对于单帧输入，一些工作尝试对人体骨架内部的空间关系进行建模，以捕捉二维姿态与三维姿态之间的空间相关性。另一些方法则提出利用二维姿态估计器提取的视觉特征，弥补二维姿态位置描述能力不足的问题。然而，由于平面内特征存在局限，这些方法通常缺乏对深度的显式建模，因此难以处理深度歧义和自遮挡问题。在我们的工作中，我们显式引入逐关节深度分布，以增强特征的深度感知能力，并采用自适应锚点到关节回归方式降低预测难度。

基于锚点的姿态估计。在手部姿态估计中，一些研究 [42, 12, 8] 设计了锚点到关节范式，将二维锚点视为局部回归器来估计手部姿态。A2J- Transformer [14] 进一步提出将三维锚点设置与 Transformer [23] 相结合，以增强抵抗遮挡的能力。然而，这些方法中的锚点设置都是固定的，无法根据图像内容进行调整，因此在某些场景中的性能并不理想。我们首次探索了在三维人体姿态提升中使用锚点，并利用二维姿态先验将锚点设计从静态设置增强为可学习的自适应设置。

三维增强视觉特征。在姿态估计中，一些工作将视觉特征扩展为三维体素 [37, 45, 32, 25]，并应用三维卷积神经网络回归三维人体姿态。尽管效果显著，但受内存使用量和计算成本的限制，三维人体姿态估计的准确性可能受到体素分辨率的制约。为高效构建三维特征表示，一些三维物体识别领域的工作 [29, 17, 51, 6, 46] 进行了积极探索。这些方法通常利用给定或预测的深度图，将平面内特征扩展到三维空间。然而，单张前景深度图不足以实现准确的三维姿态估计，尤其是在存在显著深度歧义的自遮挡情况下。因此，我们提出预测逐关节深度分布，并将其与平面内特征融合，从而将这些特征提升为深度感知三维特征。

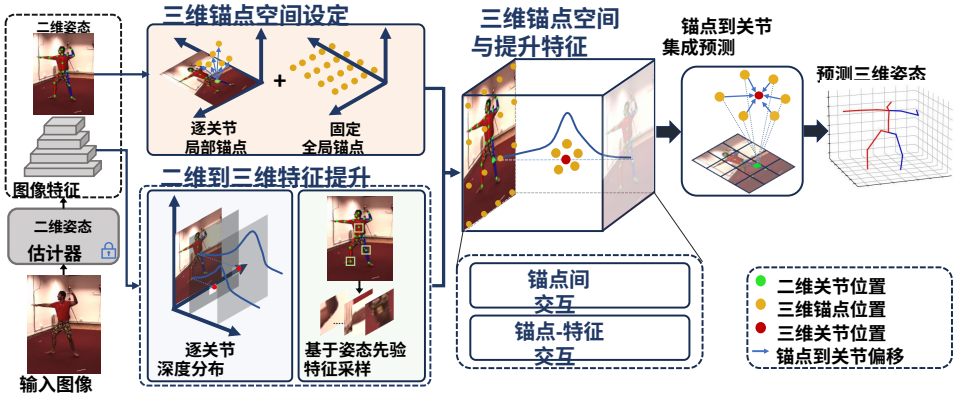


图 2：PandaPose 流程概览。给定输入的单帧二维姿态和中间图像特征，我们在三维空间中自适应采样锚点。通过估计逐关节深度分布，并采用基于二维姿态先验的采样策略，我们将特征从二维域提升至三维域。经过三维锚点特征交互后，我们通过锚点到关节集成预测得到预测的三维姿态。

3 方法

3.1 主流程概述

主要技术流程如图 2 所示。遵循最先进的姿态提升方法 [55],所采用的两阶段流程，给定大小为 H × W ×3的输入 RGB图像 I，现成的二维姿态估计器 [33]会生成相应的二维人体姿态 $P_J^{2D} \in \mathbb{R}^{N_J \times 2}$ 以及中间金字塔特征图，其中 N_J 表示关节数量。给定输入二维姿态，我们提出通过自适应逐关节局部锚点和固定全局锚点设置（第 3.2 节）构建三维锚点集 A。为将输入的二维平面视觉特征提升为深度感知三维特征，我们预测逐关节深度分布图 $Dist_D$ 和深度嵌入 F_D（第 3.3 节）。我们还设计了一种基于二维姿态先验的特征采样策略，以提取关节相关视觉特征 F_I（第 3.4 节），旨在过滤背景噪声的同时最大限度地降低计算成本。三维锚点被编码为可学习锚点查询 Q_{anchor}，并在三维锚点空间中与提升后的三维特征进行交互，从而使锚点查询成为一种统一表示，涵盖三维锚点、视觉线索和深度信息（第 3.5 节）。最后，通过锚点到关节集成预测（第 3.6 节）得到预测的三维姿态 $P_J^{3D} \in \mathbb{R}^{N_J \times 3}$。

3.2 逐关节自适应三维锚点设置

与当前通常建立从二维姿态到三维姿态的直接关节到关节映射的姿态提升方法不同，我们将三维锚点构建为与输入二维姿态相关的粗略初始化，然后以集成方式预测每个关节的三维锚点到关节偏移。以往基于锚点的方法 [42, 14] 通常采用稀疏分布于三维空间中的固定锚点设置，缺乏对特定姿态模式的适应性，从而导致远离人体区域的锚点产生较大偏移（图 3）。鉴于预测较大偏移会带来更大的误差余地，这可能会降低关节定位的整体性能，尤其是在存在遮挡或二维姿态误差的情况下。为解决这一问题，我们提出利用二维姿态先验，将锚点设置从静态演变为动态自适应。我们对不同锚点设置在Human3.6M [13] 上进行了比较（图3）。通过选取前50个锚点到关节权重，我们识别出对关节预测贡献最大的信息性锚点，并计算其占总数的比例。我们的自适应设置将偏移从154.6毫米显著降至69.7毫米。自适应三维锚点生成过程如图 4 所示。输入的二维姿态 P_J^{2D}首先归一化为 $[-1, 1]$。然后自适应生成三维局部锚点，为每个关节 $j \in J$:采样得到一组 偏移 $\delta \in \mathbb{R}^{K \times 3}$，包含 K

$$\delta_J = Linear(P_J^{2D}), \quad \delta_J \in \mathbb{R}^{N_J \times K \times 3}. \tag{1}$$

由于考虑了全局二维姿态上下文，即使二维关节在局部存在不准确的情况，模型仍具有良好的适应性，如图10所示。同时，归一化的二维关节位置（j_x、j_y）在深度 0 处被初始化为三维位置，形成（j_x、j_y、0）。随后，通过以下方式生成三维局部锚点集A_{local}：

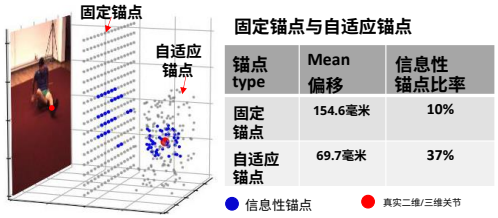


图3：不同锚点设置的对比。对于远离人体区域的关节，自适应锚点更有效。

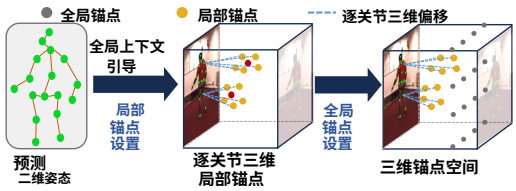


图4：自适应三维锚点集的示意图，包括逐关节局部锚点和全局固定锚点。

将采样偏移量加到对应的关节三维位置（j_x，j_y，0）：

$$A_{local} = \left\{ a \mid P_a = (j_x, j_y, 0) + \delta_{j,k}, j \in J, k \in K \right\}. \tag{2}$$

为增强模型的泛化能力和训练稳定性，我们引入一组全局固定锚点，以稳定的全局上下文补充局部自适应锚点。具体而言，我们预设三维全局锚点 $A_{global} \in \mathbb{R}^{256\times3}$，它们均匀分布在三维空间中根关节所在的平面上，平面内步长为 $S_x = H/16$ 和 $S_y = W/16$。最终，三维锚点集 A 由全局锚点和局部锚点组合而成：

$$A = A_{global} \cup A_{local}. \tag{3}$$

随后，利用三维锚点集 A，通过锚点查询Q_{anchor}的形式实现锚点-特征交互。其三维位置 P_A^{3D}是锚点到关节预测的初始位置。

3.3 逐关节深度分布估计

在三维人体姿态提升中，关键挑战之一是解决自遮挡问题。当前方法通常受限于对受约束二维特征的依赖，因此难以处理深度歧义。然而，仅将单幅深度图作为图像特征的补充进行估计，仍不足以满足精确姿态估计的要求。首先，姿态估计更关注关节之间的相对深度关系，而单幅深度图无法为被遮挡或在二维平面上彼此靠近的关节提供有效的深度预测。其次，在缺少真实深度图的情况下，如何使模型实现更准确的深度估计仍然是一个挑战。如图5所示，由于投影畸变，三维空间中相距较远的关节在二维平面上可能显得很近，并且常常在下采样特征图中重叠。这会导致训练过程出现歧义，即一个点对应多个真值深度；也会导致推理不准确，即多个关节共享同一

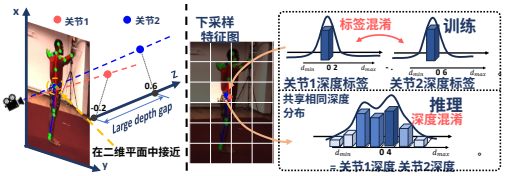


图5：现有方法通常预测单幅深度图。由于投影畸变，三维空间中相距较远的关节在二维平面上可能看起来彼此接近，并且常常重叠在下采样特征图中，从而在训练和推理过程中造成显著混淆。

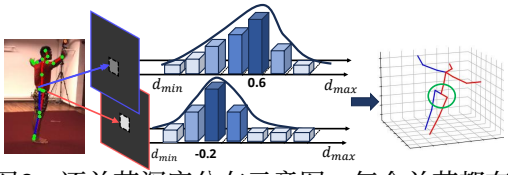


图6：逐关节深度分布示意图。每个关节都有对应的预测深度图，使二维平面上被遮挡或彼此靠近的关节能够具有不同的深度分布。

同样不正确的深度。因此，固有的深度歧义无法得到真正妥善的解决。

因此，我们创新性地设计了一种逐关节方法，为每个关节预测单独的深度分布，如图6所示。逐关节深度分布估计过程如图7(a)所示。我们采用轻量级深度网络来估计关节级深度（其架构见附录A.2）。我们利用预训练二维姿态估计器的多尺度特征，在相应尺度上估计深度分布图。为兼顾效率与特征丰富度，我们选择单个$H/8 \times W/8$ 分辨率特征图作为输入，并独立估计每个关节的深度分布。为简化连续深度值的预测，我们将深度范围$[-d_{min}, d_{max}]$ 划分为K_{bin} 个分箱，并将每个分箱视为一个独立类别，从而将深度估计转化为分类任务 [30, 1]。随后，将生成的深度分布图 $N_J \times H/8 \times W/8 \times K_{bin}$ 进行

通过插值放大，以匹配不同的图像分辨率，形成不同的$Dist_D$。此外，我们在深度网络输出的特征上使用单个 Transformer编码器层 [7]，以生成深度感知嵌入F_D，为三维锚点编码几何深度线索。

为了解决缺少真实稠密深度图时，生成与三维姿态一致的准确深度分布这一难题，我们引入将三维姿态标注用作稀疏监督的方法。因此，我们首先将真实三维姿态的深度值映射到对应的 bin 作为标签，然后针对每个关节深度图分别计算二元交叉熵损失 [6]。为降低学习复杂度并滤除无关噪声，损失计算仅限于二维关节周围的$r \times r$区域。

$$\mathcal{L}_{depth} = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{k=0}^{l_n-1} logP^k_{(n,0)} + \sum_{k=l_n}^{K_{bin}} logP^k_{(n,1)}\right) \tag{4}$$

其中，$N = N_J \times r \times r$ 表示逐关节深度监督像素的总数。

3.4 基于二维姿态先验的特征采样

对于三维人体姿态估计，基于注意力的模型通常会处理每个像素，这可能引入背景噪声并增加计算负载，却无法提升准确率。为解决这些问题，我们提出一种基于二维姿态先验的特征采样策略，以改进传统注意力模块的特征提取。我们利用二维姿态作为先验，有选择地关注关节周围$r \times r$ 区域内的特征，而非整幅图像（见图 7(b)）。这种方法减少了无关信息的干扰，同时提高了计算效率。对于多尺度特征，我们将该策略应用于前两个高分辨率层，采样半径与分辨率成正比（$r = H/16$），同时保留低分辨率层中的全局语义信息。此外，我们将图像特征通道归一化为相同的C_I。深度分布也采用类似的采样策略，以确保输入锚点特征交互解码器的N_F 个 token 实现像素对齐，其中N_F 是将多层采样特征图展平后汇总得到的像素 token 数量。随后，利用锚点-特征交互解码器对采样的图像特征$F_I \in \mathbb{R}^{N_F \times C'}$ 和深度分布$Dist_D \in \mathbb{R}^{N_F \times K^{bin}}$ 进行处理，使其在三维锚点空间内进一步交互，从而在存在遮挡的情况下实现稳健的关节定位。

3.5 三维锚点特征交互

在三维锚点空间中，我们提出利用可学习自适应三维锚点，使其与深度特征和视觉特征交互，从而增强对空间上下文的理解，解决姿态估计中准确捕捉空间关系和深度信息所面临的挑战。锚点特征交互解码器的结构如图 8所示。基于可学习自适应三维锚点，我们将其编码为可学习锚点查询向量 Q_{anchor}，通过锚点-特征交互解码器预测三维姿态。每个解码器层由一个深度交叉注意力层、一个锚点间自注意力层和一个三维可变形交叉注意力层组成。深度交叉注意力层用于捕捉潜在深度特征，使锚点能够根据深度引导区域自适应地理解空间上下文，增强对关节间深度关系的感知。在此过程中，我们进行线性变换

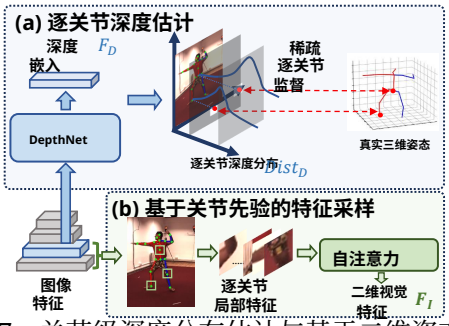


图 7：关节级深度分布估计与基于二维姿态先验的特征采样流程。

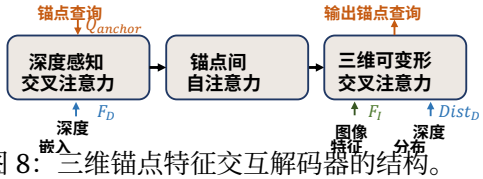


图 8：三维锚点特征交互解码器的结构。

将锚点查询和深度嵌入分别作为查询Q_D、键K_D和值V_D:

$$Q_D = Linear(Q_{anchor}), \quad K_D, V_D = Linear(F_D), \tag{5}$$

随后，将深度感知三维锚点查询向量输入锚点间自注意力层，以促进锚点之间的关节线索。最后，我们将展平的平面内特征F_I 提升至三维空间

	方法	发表场所	帧	参数（M）用于提升模块	MPJPE ↓	PA-MPJPE ↓
序列基于	PoseFormer [57]	ICCV 2021	81	9.5	44.3	34.6
	MHFormer [20]	CVPR'22	351	24.8	43.0	34.4
	MixSTE [50]	CVPR'22	243	33.6	40.9	32.6
	P-STMO [31]	ECCV'22	243	4.6	43.0	34.4
	STCFormer [35]	CVPR'23	243	18.9	41.0	32.0
	KTPFormer [28]	CVPR'24	243	35.2	40.1	31.9
图像基础	完整测试集					
	GraphSH [43]	CVPR'21	1	3.7	51.9	-
	HCSF [48]	ICCV 2021	1	-	47.9	39.0
	GraFormer [56]	CVPR'22	1	-	51.8	-
	Diffpose [10]	CVPR'23	1	1.9	49.7	-
	Zhou等人[59]	AAAI'24	1	-	46.4	-
	HiPART [58]	CVPR'25	1	2.4	42.0	-
	CA-PF [55]	NeurIPS'23	1	14.1	41.4	33.5
	PandaPose（我们的方法		1	15.2	**39.8** (1.6↓)	**32.7** (0.8↓)
	困难子集					
	CA-PF [55]	NeurIPS'23	1	14.1	82.4	82.0
	PandaPose（我们的方法）		1	15.2	**73.1** (9.3↓)	**69.9** (12.1↓)

表 1：与 Human3.6M 上最先进方法的比较。MPJPE 和 PA-MPJPE 的单位为毫米。最佳结果以**粗体**显示。我们的方法不仅在完整测试集上取得领先性能，在困难子集上也有显著提升。

利用逐关节深度分布 $Dist_D$ 通过外积 $F_{3D} = Dist_D \otimes F_I$，并应用三维可变形交叉注意力（DCA）层，使三维锚点查询向量能够有效聚合三维场景的视觉特征。随后，以三维锚点作为参考点，对于位于位置 P_a的特定锚点a，我们通过三维可变形交叉注意力进行特征交互：

$$DCA(a) = \sum_{n \in N} W_n \phi(F_{3D}, P_a + \Delta S_n). \tag{6}$$

对于 [61], 每个锚点都关联着N 个位于P_A附近的自适应采样点，其偏移量ΔS_n 由输入锚点查询Q_{anchor} 通过线性层预测。这使 Transformer 能够以数据驱动的方式动态关注稀疏但语义丰富的局部区域，从而同时提升效率和准确性。术语 $\phi(F_{3D}、P_a + \Delta S_n)$ 表示通过三线性插值从扩展的三维特征图F_{3D}中采样特征。输出锚点查询在三维锚点空间中形成统一表示，融合多模态信息，并有助于解决深度歧义问题。

3.6 锚点到关节预测

利用解码器输出的锚点查询Q_{anchor}，我们使用多层感知机层提取偏移量O 和权重W，以表示锚点相对于所有关节的位置。关节位置以锚点到关节偏移量的加权和形式表示：

$$P_j^{3D} = \sum_{a \in A} \tilde{W}_{a,j}(P_a + O_{a,j}), \tag{7}$$

其中，P_j和P_a分别表示特定关节j和锚点a的三维位置。$O_{a,\ j}$表示从a 指向j的偏移~。$W_{a,\ j}$表示由 Softmax 得出的a 指向j的权重。

3.7 损失函数

我们以稀疏监督的方式训练逐关节深度分布图（第 3.3 节），记为$\mathcal{L}_{depth}$，并使用 MPJPE [27, 57] 监督三维姿态的训练，记为 $\mathcal{L}_{pose}$。总体损失函数表示为：

$$\mathcal{L} = \lambda_1 \mathcal{L}_{pose} + \lambda_2 \mathcal{L}_{depth}. \tag{8}$$

这里我们设置 $\lambda_1 = 2$和 $\lambda_2 = 0.1$，以平衡尺度。

4 实验

4.1 数据集和评估指标

方法	正确关键点百分比 ↑	曲线下面积 ↑	MPJPE ↓
完整测试集			
GraFormer [56]	79.0	43.8	-
Li 等人[18]	81.2	46.1	99.7
HCSF [48]	82.1	46.2	-
Zhou 等人[59]	88.2	59.3	-
CA-PF [55]	98.0	75.4	32.7
PandaPose（我们的方法）	**98.6** (0.6↑)	**75.8** (0.4↑)	**31.8** (0.9↓)
困难子集			
CA-PF [55]	84.5	53.2	66.6
PandaPose（我们的方法）	**94.3** (9.8↑)	**62.5** (9.3↑)	**51.8** (14.8↓)

表2：基于图像的方法在 MPI-INF-3DHP 上的比较。

方法	MPJPE	PA-MPJPE ↓
STRGCN [2]	112.9	69.6
VideoPose [27]	101.8	63.0
PoseFormer [57]	118.2	73.1
学习 [49]	91.1	54.3
PCT [9]	83.1	53.9
DiffPose[10]	82.7	53.8
Di²Pose [41]	79.3	50.1
HiPART [58]	77.2	48.8
PandaPose（我们的方法）	**74.9**(2.3↓)	**46.9**(1.9↓)

表 3：在 3DPW 上，我们的方法与最先进方法的跨数据集比较。

全局固定锚点	自适应局部锚点	MPJPE↓（完整）	MPJPE ↓（困难）
不含锚点的 PandaPose		42.1	81.9
✓		40.8 (1.3↓)	76.2 (5.0↓)
	✓	40.1 (2.1↓)	74.0 (7.2↓)
✓	✓	39.8 (2.3↓)	73.1 (8.1↓)

表4：锚点设置策略比较。

锚点特征	深度分布	MPJPE↓（完整）	MPJPE ↓（困难）
2D	-	40.9	80.8
3D	单一	40.3 (0.6↓)	75.9 (4.9↓)
3D	逐关节	39.8 (1.1↓)	73.1 (7.7↓)

表 5：逐关节三维特征提升的消融研究。

我们分别在两个常用的三维人体姿态数据集（即，**Human3.6M** [13]和 **MPI-INF-3DHP** [26]）上训练模型，以展示其效果 - PandaPose 的有效性。为更好地验证其泛化能力，我们在 Human3.6M 上训练模型，并在 **3DPW** [39]野外数据集上进行跨数据集评估。为更好地说明模型对遮挡和二维姿态不准确性的鲁棒性，我们从 Human3.6M 和 MPI-INF-3DHP 中选取预测的二维姿态与真实二维姿态之间误差为> 5 的数据样本，作为困难子集（≈ 5占数据集的 %）。关于各数据集和评估指标的更多一般信息见附录A.1。

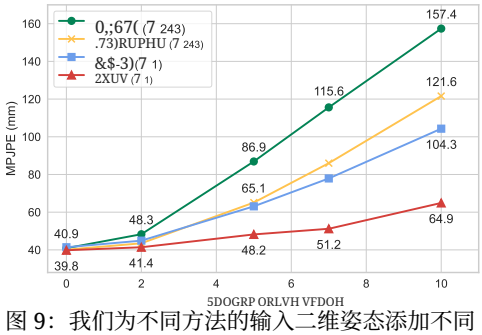


图 9：我们为不同方法的输入二维姿态添加不同尺度的高斯噪声，以测试其对噪声输入的鲁棒性。

4.2 实现细节

我们的模型使用 PyTorch 实现。根据 [55], 我们使用预训练且冻结的 HRNet-w32 [33]作为图像骨干网络，并仅提取金字塔特征图作为模型输入。为确保公平并与以往工作保持一致，我们在 Human3.6M 和 3DPW 中使用 CPN 检测的 [4] 2维姿态作为输入，在 MPI-INF-3DHP 中使用真实二维姿态。实验在 2 个 NVIDIA RTX 3090 GPU 上进行，使用 AdamW 优化器 [24]，学习率为 $4e^{-4}$，权重衰减为 0。98，共训练 30 个周期。所有消融研究均在 Human3.6M 上进行 [13]。

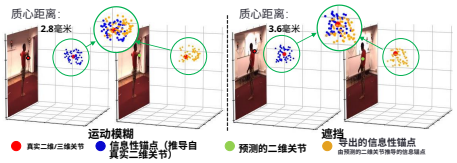


图 10：不准确的/真实二维姿态下的自适应三维锚点设置对比。我们计算两组锚点的质心，并求出它们之间的距离（即，质心距离 [36]）。由不准确二维姿态生成的自适应三维锚点与真实二维姿态分布高度吻合，质心距离较小。

4.3 与最先进方法的比较

Human3.6M 数据集。我们在表 1 中将本模型与 Human3.6M 上的最先进方法进行了评估，涵盖基于图像的方法和基于序列的方法。与最先进的基于图像的方法相比，本模型在 MPJPE 上取得了显著提升，将 MPJPE 从 41.4 mm 降至 39.8 mm（降低 1.6 mm），同时 PA-MPJPE 也降低了 0.8 mm。值得注意的是，我们的单帧模型无需时序上下文，便达到了最先进的基于序列的方法的性能，MPJPE 从 40.1 mm 降至 39.8 mm。为展示本方法的优越性，我们将其与困难子集上的 CA-PF [55] 进行了比较。本方法在 MPJPE 上显著提升了 9.3 mm，在 PA-MPJPE 上提升了 12.1 mm。困难子集上的可视化对比见图 11。我们还在附录 A.3 中对非困难场景进行了性能比较。

方法	MPJPE（完整）↓	MPJPE（挑战）↓
回归	41.6	77.2
使用16个分箱进行分类	40.7	74.8
使用64个分箱进行分类	39.8	73.1
使用128个分箱进行分类	40.2	74.7

表6：深度离散化策略比较。

方法	GPU显存（M）	MPJPE ↓
不使用特征采样	21670	40.0
随机采样	13784	45.6
二维姿态先验特征采样	13784	39.8

表 7：特征采样策略比较 .

MPI-INF-3DHP 数据集。我们在表 2 中评估了模型在 MPI-INF-3DHP 数据集上的性能。我们的 PandaPose 取得了最佳结果，优于现有最先进的基于图像的模型，正确关键点百分比提高了 0。6%，曲线下面积提高了 0。4%，MPJPE 降低了 0。9*mm* 。在困难子集上（例如图 11第 3 行），本方法取得了显著优势：正确关键点百分比提高了 9。8%，曲线下面积提高了 9。3%，MPJPE 降低了 14。8*mm* 。

3DPW 数据集。我们评估了在 Human3.6M 上预训练的模型在 3DPW 数据集上的表现，如表3所示。在相同的跨数据集设置下，我们的模型取得了最先进的性能，MPJPE 显著提升了 2。3毫米，PA-MPJPE 提升了 1。9毫米，展现出我们方法强大的泛化能力。

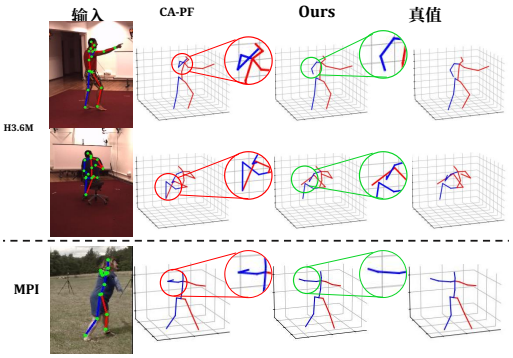


图 11：困难案例的视觉比较（例如，明显遮挡或二维姿态不准确）。圆圈标出了我们的方法预测效果更好的位置。

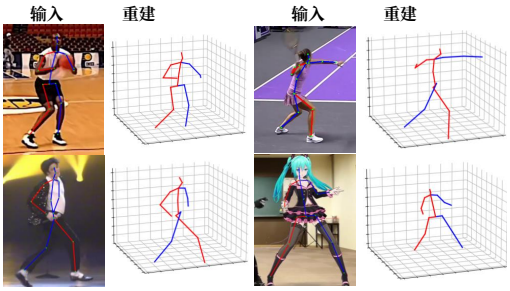


图 12：来自互联网数据集的样本可视化，涵盖各种场景和虚拟化身。

4.4 消融研究

三维锚点设置。为验证锚点到关节回归和我们提出的自适应三维锚点设置的有效性，我们进行了消融研究，如表4所示。基线模型使用多层感知机（MLP），直接根据解码器的输出特征回归三维姿态，其性能显著下降（与锚点到关节回归方式相比，完整测试集上的 MPJPE 增加了 2。3毫米，困难子集上增加了 8。1毫米）。此外，全局锚点和自适应局部锚点都能提升性能；由于自适应局部锚点与关节位置更为接近，因此精度更高。在困难子集中，自适应锚点相比全局锚点提升了 2。2毫米。将全局锚点和局部锚点结合起来，同时整合全局与局部上下文，可使模型达到最佳性能。这些发现表明，有效的回归方式和经过精心设计的三维锚点能够提升三维姿态估计的性能。

逐关节三维特征提升。为评估将二维特征提升为深度感知三维特征的效果，我们在表5中对关键设计进行了消融研究。首先，我们移除整个深度分支，仅使用三维锚点和平面内二维特征进行交互，作为基线。接着，我们预测单幅深度图，并加入三维可变形交叉注意力和深度交叉注意力，以促进特征提升。性能有所提升，凸显了空间上下文对三维姿态估计的重要性。随后，我们预测逐关节深度分布，而非仅预测单幅深度图。结果显示，性能有所提升，尤其是在困难子集上，MPJPE显著降低了7.7毫米，这表明关节级的精细深度信息有助于提高三维姿态估计的准确性，尤其是在存在自遮挡的情况下。

深度离散化策略 如表6所示，我们将分类深度头替换为回归头进行比较。分类方法的准确率更高，最佳分箱数量为 64。性能提升可归因于任务复杂度与轻量级分类头之间更好的匹配，这有助于缓解拟合困难。尤其是在存在严重遮挡的困难情况下，预测的 MPJPE 提升了 4.1 毫米 （77.2 - 73.1）。我们还发现，分箱数量过多会导致性能下降。

对噪声二维姿态的鲁棒性。我们向输入的二维姿态添加随机高斯噪声，以比较不同噪声水平下的性能，如图9所示。随着二维姿态估计质量下降，我们的方法仍比其他方法保持更高的准确率，表明其对二维姿态估计器稳定性的敏感度更低，实际应用灵活性更强。如图10所示，我们的自适应锚点在困难场景下表现出更强的泛化能力，即使二维姿态不准确，其分布仍与真值接近。

基于二维姿态先验的特征采样。我们在表7中验证了基于二维姿态先验的特征采样的有效性。在批量大小为 176 时，与使用完整多尺度特征图相比，我们的策略仅使用近一半的 GPU显存，同时实现了稳定的性能。相比之下，对等量特征进行随机采样会导致性能显著下降，表明我们的方法能够在降低环境噪声的同时，高效捕获关键特征。

4.5 定性分析

为直观展示我们方法的优越性和泛化能力，我们在数据集之外的样本上（即 Human3.6M 和 MPI-INF-3DHP）评估了我们的方法，如图12所示。为确保多样性，我们选取了不同的事件、视角和服装，并且还在虚拟化身上测试了我们的模型。我们的方法展现出良好的结果，揭示了其在实际应用中的潜力。关于效率和失败案例的更多可视化与分析见附录。A

5 结论

本文提出了 PandaPose，这是一种新颖的三维人体姿态提升方法，通过将二维姿态先验传播至三维锚点空间，将其作为统一中间表示。三维锚点空间包括：规范坐标系中的逐关节三维锚点，为减轻二维姿态估计误差提供准确且鲁棒的先验；深度感知的逐关节特征提升，以分层方式整合深度信息，解决由自遮挡引起的歧义；以及锚点-特征交互解码器，将三维锚点与提升后的特征相结合，生成统一的锚点查询，其中封装了逐关节三维锚点集合、视觉线索和几何深度信息。随后，利用这些锚点查询促进锚点到关节集成预测。在 Human3.6M、MPI-INF-3DHP 和 3DPW 数据集上的实验表明，PandaPose 不仅解决了上述挑战，还实现了最先进的性能，尤其是在困难场景下。局限性与更广泛的影响将在附录 A.7 中讨论。

致谢。
本工作得到国家自然科学基金委员会（项目编号：62271221）和华中科技大学太湖未来技术创新基金（项目编号：2023-B-8）的支持。本研究还得到 Joey Tianyi Zhou 的 A*STAR SERC 中央研究基金（应用导向的基础研究）支持。

参考文献

[1] Shariq Farooq Bhat、Ibraheem Alhashim 和 Peter Wonka。Adabins：使用自适应区间进行深度估计。在*IEEE*计算机视觉与模式识别会议，第4009–4018页，2021年。5[2] Yujun Cai、Liuhao Ge、Jun Liu、Jianfei Cai、Tat-Jen Cham、Junsong Yuan 和 Nadia Magnenat Thalmann。利用时空关系，通过图卷积网络进行三维姿态估计。在*IEEE*会议论文集/*CVF*国际计算机视觉会议，第2272–2281页，2019年。8[3] Liang-Chieh Chen、George Papandreou、I asonas Kokkinos、Kevin Murphy 和 Alan L Yuille。DeepLab：使用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割。*IEEE*模式分析与机器智能汇刊，40(4)：834–848，2017年。14[4] Yilun Chen、Zhicheng Wang、Yuxiang Peng、Zhiqiang Zhang、Gang Yu 和 Jian Sun。用于多人姿态估计的级联金字塔网络。在*IEEE*计算机视觉与模式识别会议，第7103–7112页，2018年1，3，8，15–。[5] Seunggeun Chi、Hyung-gun Chi、Qixing Huang 和 Karthik Ramani。Infogcn++：通过预测未来学习表示，用于在线基于骨架的动作识别。*IEEE*模式分析与机器智能汇刊，2024年。1

[6] Xiaomeng Chu、Jiajun Deng、Yuan Zhao、Jianmin Ji、Yu Zhang、Houqiang Li 和 Yanyong Zhang。Oa-bev：通过引入目标感知，为多相机三维目标检测构建鸟瞰视图表示。arXiv预印本 arXiv:2301.05711，2023年。2、3、6[7] Alexey Dosovitskiy。一幅图像胜过1.6万字：用于大规模图像识别的Transformer。arXiv预印本 arXiv:2010.11929，2020年。6[8] Linpu Fang、Xingyan Liu、Li Liu、Hang Xu 和 Wenxiong Kang。Jgr-p2o：基于联合图推理的像素到偏移预测网络，用于从单幅深度图进行三维手部姿态估计。在欧洲计算机视觉会议，第120–137页，2020年。3[9] Zigang Geng、Chunyu Wang、Yixuan Wei、Ze Liu、Houqiang Li 和 Han Hu。作为组合式标记的人体姿态。在IEEE/CVF计算机视觉与模式识别会议论文集，第660–671页，2023年。8[10] Ji a Gong、Lin Geng Foo、Zhipeng Fan、Qiuhong Ke、Hossein Rahmani 和 Jun Liu。Diffpose：迈向更可靠的三维姿态估计。在IEEE计算机视觉与模式识别会议，第13041–13051页，2023年。7、8[11] Nate Hagbi、Oriel Bergig、Jihad El-Sana 和 Mark Billinghurst。移动增强现实中的形状识别与姿态估计。IEEE可视化与计算机图形学汇刊，第1369–1379页，2011年10月。1[12] Weiting Huang、Pengfei Ren、Jingyu Wang、Qi Qi 和 Haifeng Sun。Awr：用于三维手部姿态估计的自适应加权回归。在AAAI，第34卷，第11061–11068页，2020年。3[13] Catalin Ionescu、Dragos Papava、Vlad Olaru 和 Cristian Sminchisescu。Human3. 6m：用于自然环境中三维人体感知的大规模数据集与预测方法。IEEE模式分析与机器智能汇刊，36(7)：1325–1339，2013年。3、4、8、14、15[14] Changlong Jiang、Yang Xiao、Cunlin Wu、Mingyang Zhang、Jinghong Zheng、Zhiguo Cao 和 Joey Tianyi Zhou。A2j-transformer：用于从单幅RGB图像进行三维交互手部姿态估计的锚点到关节Transformer网络。在IEEE计算机视觉与模式识别会议，第8846–8855页，2023年。2、3、4[15] Woo-Shik Kim、Antonio Ortega、PoLin Lai、Dong Tian 和 Cristina Gomila。用于视图渲染和深度编码的深度图失真分析。在IEEE国际图像处理会议，第721–724页，2009年。2[16] Branislav Kisa˘canin、Vladimir Pavlovic 和 ThomasS. Huang。人机交互实时视觉。2005年1月。1[17] Hongyang Li、Hao Zhang、Zhaoyang Zeng、Shilong Liu、Feng Li、Tianhe Ren 和 Lei Zhang。Dfa3d：用于二维到三维特征提升的三维可变形注意力。在国际计算机视觉会议，第6684–6693页，2023年。2、3[18] Shichao Li、Lei Ke、Kevin Pratama、Yu-Wing Tai、Chi-Keung Tang 和 Kwang-Ting Cheng。结合进化式训练数据的级联深度单目三维人体姿态估计。在IEEE计算机视觉与模式识别会议，第6173–6183页，2020年。8[19] Wenhao Li、Hong Liu、Runwei Ding、Mengyuan Liu、Pichao Wang 和 Wenming Yang。利用步幅Transformer和时间上下文进行三维人体姿态估计。IEEE多媒体汇刊，25：1282–1293，2022年。3[20] Wenhao Li、Hong Liu、Hao Tang、Pichao Wang 和 Luc Van Gool。Mhformer：用于三维人体姿态估计的多假设Transformer。在IEEE计算机视觉与模式识别会议，第13147–13156页，2022年。1、3、7、15、16[21] Wenhao Li、Mengyuan Liu、Hong Liu、Tianyu Guo、Ti Wang、Hao Tang 和 Nicu Sebe。Graphmlp：用于三维人体姿态估计的类图MLP架构。模式识别，158：110925，2025年。3[22] Ruixu Liu、Ju Shen、He Wang、Chen Chen、Sen-ching Cheung 和 Vijayan Asari。注意力机制利用时间上下文：实时三维人体姿态重建。在IEEE计算机视觉与模式识别会议，第5064–5073页，2020年。3[23] Shilong Liu、Feng Li、Hao Zhang、Xiao Yang、Xianbiao Qi、Hang Su、Jun Zhu 和 Lei Zhang。Dab-detr：动态锚框是更好的DETR查询向量。arXiv预印本 arXiv:2201.12329，2022年。3[24] I Loshchilov。解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101，2017年。8[25] Jameel Malik、Ibrahim Abdelaziz、Ahmed Elhayek、Soshi Shimada、Sk Aziz Ali、Vladislav Golyanik、Christian Theobalt 和 Didier Stricker。Handvoxnet：用于从单幅深度图进行三维手部形状与姿态估计的深度体素网络。在IEEE计算机视觉与模式识别会议，第7113–7122页，2020年。3–。

[26] Dushyant Mehta、Helge Rhodin、Dan Casas、Pascal Fua、Oleksandr Sotnychenko、Weipeng Xu 和 Christian Theobalt。利用改进的卷积神经网络监督，在野外环境中进行单目三维人体姿态估计。载于《2017年国际三维视觉会议》，第506–516页。IEEE，2017年。3, 8, 14[27] Dario Pavllo、Christoph Feichtenhofer、David Grangier 和 Michael Auli。利用时间卷积和半监督训练进行视频中的三维人体姿态估计。载于 IEEE 计算机视觉与模式识别会议，第7753–7762页，2019年。1, 3, 7, 8[28] Jihua Peng、Yanghong Zhou 和 PY Mok。KTPFormer：用于三维人体姿态估计的运动学与轨迹先验知识增强型 Transformer。载于 IEEE 计算机视觉与模式识别会议，第1123–1132页，2024年。2, 3, 7, 15[29] Jonah Philion 和 Sanja Fidler。Lift, Splat, Shoot：通过隐式反投影到三维空间，对任意相机系统采集的图像进行编码。载于欧洲计算机视觉会议，第194–210页。Springer，2020年。3[30] Cody Reading、Ali Harakeh、Julia Chae 和 Steven L Waslander。用于单目三维目标检测的类别深度分布网络。载于 IEEE 计算机视觉与模式识别会议，第8555–8564页，2021年。5[31] Wenkang Shan、Zhenhua Liu、Xinfeng Zhang、Shanshe Wang、Siwei Ma 和 Wen Gao。P-STMO：用于三维人体姿态估计的预训练时空多对一模型。载于欧洲计算机视觉会议，第461–478页。Springer，2022年。3, 7[32] Vinkle Srivastav、Keqi Chen 和 Nicolas Padoy。SelfPose3D：自监督多人多视角三维姿态估计。载于 IEEE 计算机视觉与模式识别会议，第2502–2512页，2024年。3[33] Ke Sun、Bin Xiao、Dong Liu 和 Jingdong Wang。用于人体姿态估计的深度高分辨率表征学习。载于 IEEE 计算机视觉与模式识别会议，第5693–5703页，2019年。1, 3, 4, 8, 15[34] M. Svenstrup、S. Tranberg、H.J. Andersen 和 T. Bak。面向人机交互的姿态估计与自适应机器人行为。载于《2009年IEEE国际机器人与自动化会议》，2009年5月。1[35] Zhenhua Tang、Zhaofan Qiu、Yanbin Hao、Richang Hong 和 Ting Yao。利用时空交叉注意力进行三维人体姿态估计。载于 IEEE 计算机视觉与模式识别会议，第4790–4799页，2023年。7[36] H. Teng、D. Chatziparaschis、X. Kan、A. Roy-Chowdhury 和 K. Karydis。彩色点云的质心距离关键点检测器。2023年。8[37] Hanyue Tu、Chunyu Wang 和 Wenjun Zeng。VoxelPose：面向野外环境中的多相机三维人体姿态估计。载于欧洲计算机视觉会议，第197–212页，2020年。3[38] A Vaswani。Attention Is All You Need。载于《神经信息处理系统进展》，2017年。3[39] Timo Von Marcard、Roberto Henschel、Michael J Black、Bodo Rosenhahn 和 Gerard Pons-Moll。利用惯性测量单元和移动相机，在野外环境中恢复精确的三维人体姿态。载于《欧洲计算机视觉会议论文集》（ECCV），第601–617页，2018年。3, 8, 14[40] Jingbo Wang、Sijie Yan、Yuanjun Xiong 和 Dahua Lin。利用视频中的运动引导进行三维姿态估计。载于欧洲计算机视觉会议，第764–780页。Springer，2020年。3[41] Weiquan Wang、Jun Xiao、Chunping Wang、Wei Liu、Zhao Wang 和 Long Chen。DiPose：用于遮挡情况下三维人体姿态估计的离散扩散模型。CoRR，2024年。8[42] Fu Xiong、Boshen Zhang、Yang Xiao、Zhiguo Cao、Taidong Yu、Joey Tianyi Zhou 和 Junsong Yuan。A2J：用于从单幅深度图像估计三维关节姿态的锚点到关节回归网络。载于国际计算机视觉会议，第793–802页，2019年。2, 3, 4[43] Tianhan Xu 和 Wataru Takano。用于三维人体姿态估计的图堆叠沙漏网络。载于 IEEE 计算机视觉与模式识别会议，第16105–16114页，2021年。3, 7[44] Tingbing Yan、Wenzheng Zeng、Yang Xiao、Xingyu Tong、Bo Tan、Zhiwen Fang、Zhiguo Cao 和 Joey Tianyi Zhou。CrossGLG：大语言模型以跨层级方式指导单样本骨架三维动作识别。arXiv 预印本 arXiv:2403.10082，2024年。1[45] Hang Ye、Wentao Zhu、Chunyu Wang、Rujie Wu 和 Yizhou Wang。Faster VoxelPose：通过正交投影实现实时三维人体姿态估计。载于欧洲计算机视觉会议，第142–159页，2022年。3

[46] Zhu Yu、Runmin Zhang、Jiacheng Ying、Junchen Yu、Xiaohai Hu、Lun Luo、Si-Yuan Cao 和 Hui-Liang Shen。用于语义场景补全的上下文与几何感知体素 Transformer。arXiv 预印本 arXiv:2405.13675，2024年。3[47] Ailing Zeng、Xiao Sun、Fuyang Huang、Minhao Liu、Qiang Xu 和 Stephen Lin。SRNet：通过分割-重组方法提升三维人体姿态估计的泛化能力。载于欧洲计算机视觉会议，第507–523页。Springer，2020年。3[48] Ailing Zeng、Xiao Sun、Lei Yang、Nanxuan Zhao、Minhao Liu 和 Qiang Xu。面向困难三维姿态估计的骨架图神经网络学习。载于国际计算机视觉会议，第11436–11445页，2021年。1, 3, 7, 8[49] Jianfeng Zhang、Kehong Gong、Xinchao Wang 和 Jiashi Feng。学习增强姿态，以进行图像和视频中的三维人体姿态估计。IEEE《模式分析与机器智能汇刊》，45(8):10012–10026，2023年。8[50] Jinlu Zhang、Zhigang Tu、Jianyu Yang、Yujin Chen 和 Junsong Yuan。MixSTE：用于视频中三维人体姿态估计的序列到序列混合时空编码器。载于 IEEE 计算机视觉与模式识别会议，第13232–13242页，2022年。1, 2, 3, 7, 15, 16[51] Renrui Zhang、Han Qiu、Tai Wang、Ziyu Guo、Ziteng Cui、Yu Qiao、Hongsheng Li 和 Peng Gao。MonoDETR：深度引导的单目三维目标检测 Transformer。载于国际计算机视觉会议，第9155–9166页，2023年。2, 3[52] Si Zhang、Hanghang Tong、Jiejun Xu 和 Ross Maciejewski。图卷积网络：全面综述。《计算社会网络》，6(1):1–23，2019年。3[53] Siqi Zhang、Chaofang Wang、Wenlong Dong 和 Bin Fan。三维人体姿态估计中深度歧义的综述。《应用科学》，12(20):10591，2022年。2[54] Long Zhao、Xi Peng、Yu Tian、Mubbasir Kapadia 和 Dimitris N Metaxas。用于三维人体姿态回归的语义图卷积网络。载于 IEEE 计算机视觉与模式识别会议，第 – 页，2019年。1, 2, 3[55] Qitao Zhao、Ce Zheng、Mengyuan Liu 和 Chen Chen。一个带上下文的二维姿态，对于三维人体姿态估计而言抵得上数百个姿态。载于《神经信息处理系统进展》，第36卷，2024年。1, 2, 3, 4, 7, 8, 14, 15, 16[56] Weixi Zhao、Weiqiang Wang 和 Yunjie Tian。GraFormer：面向三维姿态估计的图导向 Transformer。载于 IEEE 计算机视觉与模式识别会议，第20438–20447页，2022年。3, 7, 8[57] Ce Zheng、Sijie Zhu、Matias Mendieta、Taojiannan Yang、Chen Chen 和 Zhengming Ding。利用空间和时间 Transformer 进行三维人体姿态估计。载于国际计算机视觉会议，第11656–11665页，2021年。1, 3, 7, 8[58] Hongwei Zheng、Han Li、Wenrui Dai、Ziyang Zheng、Chenglin Li、Junni Zou 和 Hongkai Xiong。HiPART：用于遮挡情况下三维人体姿态估计的分层姿态自回归 Transformer。arXiv 预印本 arXiv:2503.23331，2025年。7, 8[59] Feng Zhou、Jianqin Yin 和 Peiyang Li。通过图像提升——利用图像线索实现精确的三维人体姿态估计。载于 AAAI，第38卷，第7632–7640页，2024年。2, 3, 7, 8[60] Wentao Zhu、Xiaoxuan Ma、Zhaoyang Liu、Libin Liu、Wayne Wu 和 Yizhou Wang。MotionBERT：学习人体运动表征的统一视角。载于国际计算机视觉会议，第15085–15099页，2023年。1[61] Xizhou Zhu、Weijie Su、Lewei Lu、Bin Li、Xiaogang Wang 和 Jifeng Dai。Deformable DETR：用于端到端目标检测的可变形 Transformer。arXiv 预印本 arXiv:2010.04159，2020年。7

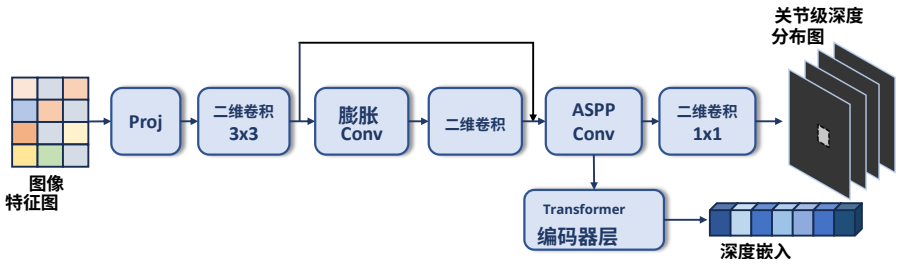


图13：轻量级深度网络示意图。

方法	MPJPE（困难子集）27377 个样本	MPJPE（非挑战子集）515967 个样本	MPJPE（完整测试集）543344 个样本
CA-PF [55] PandaPose（我们的方法）	82.4 73.1 (9.3↓)	39.9 38.8 (1.1↓)	41.4 39.8 (1.6↓)

Tabl图 8：Human3.6M 测试集上不同子集的性能比较。非挑战子集 g
sub集指从完整测试集中排除挑战子集后剩余的样本。

A 技术附录与补充材料

A.1 数据集和评估指标

Human3.6M [13] 是广泛使用的三维人体姿态估计基准。遵循以往的协议，我们使用 5 名受试者（S1、S5、S6、S7、S8）的数据训练模型，并使用 2 名受试者（S9、S11）的数据进行评估。我们报告 Human3.6M 上的 MPJPE 和 PA-MPJPE。我们选择预测的二维姿态与真实二维姿态之间误差 >5的样本作为困难子集（27.4k 个样本，约占测试集的 5%）。这些二维姿态不准确的样本通常存在严重遮挡或容易混淆的姿态模式，如图11所示。

MPI-INF-3DHP [26] 也是广泛使用的三维人体姿态估计基准，数据采集环境既包括室内，也包括具有挑战性的室外环境。我们报告 150 mm 范围内的正确关键点百分比（PCK）、曲线下面积（AUC）和 MPJPE 作为评估指标。遵循上述相同原则，我们选取了 111 个样本（约占 4%）作为困难子集。

3DPW [39] 是一个具有挑战性的野外数据集。我们在 Human3.6M 上训练模型，并在 3DPW 上进行测试，以评估模型的泛化能力。我们报告了在 3DPW 上的 MPJPE 和 PA-MPJPE。

A.2 深度网络结构

我们采用的轻量级深度网络如图所示。13。该网络约有 $5M$ 个参数，由多个卷积层组成。它以单层图像特征图作为输入，随后通过特征投影层进行处理。接着，卷积层提取深层特征，并通过空洞空间金字塔池化图[3]（空洞空间金字塔池化）模块进一步细化，以捕获多层级信息。通过这一过程，得到两个输出：

- 深度特征经过 1x1 卷积，为每个关节生成深度分布图。
- 由自注意力机制构成的单层 Transformer 编码器对深度特征进行处理，以生成深度嵌入。

A.3 不同子集上的性能分析

在表1中，我们在完整测试集以及困难子集上，将 PandaPose 与其他最先进方法的性能进行了比较，以展示其优越性，尤其是在应对遮挡和二维姿态误差方面的鲁棒性。此外，人们可能会担心，在困难样本上的提升可能会以更简单样本上性能变弱为代价。为了解决这一疑虑，我们在表8 中开展了额外实验，排除困难子集后

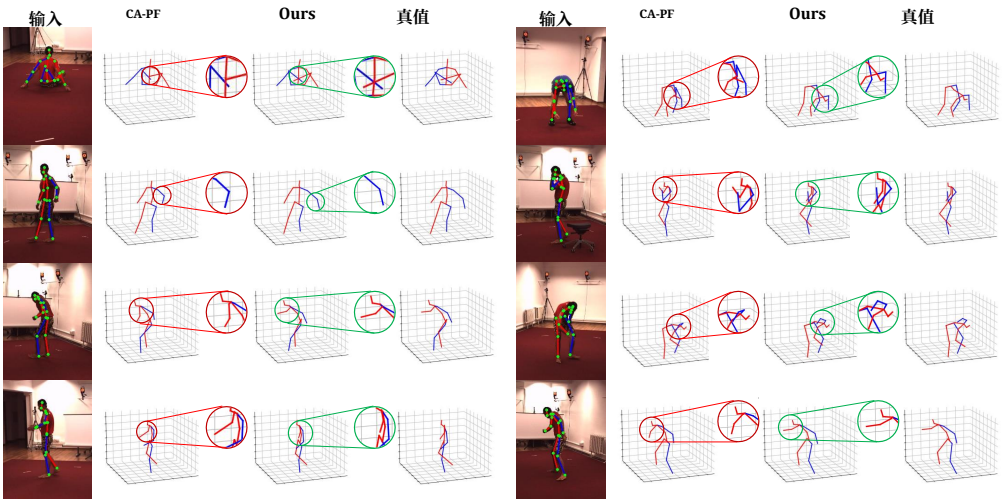


图 14：困难案例的可视化比较。圆圈标出了我们的方法预测效果更好的位置。

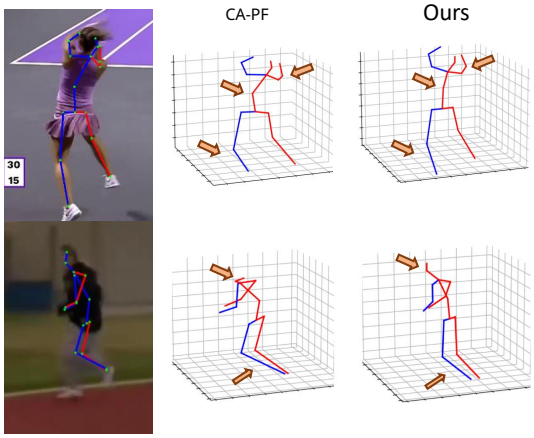


图 15：运动模糊或自遮挡情况下的失败案例可视化。

完整测试集，并将其定义为非挑战子集，并将其与 CA-PF 进行比较 [55]。即使在相对常见的情况下，我们的方法也展现出显著的改进。

A.4 不同二维姿态估计器的性能分析

我们选取三种广泛使用的二维姿态估计器（即 CPN [4],HRNet-w32 [33],HRNet-w48 [33]）进行消融实验，以考察预测三维姿态质量如何随输入二维精度变化。结果列于表 10。PandaPose 的精度相对于二维姿态估计器的误差保持相对稳定。与 CPN 和 HR-Net-w48 相比，二维精度（平均精度均值）下降了 5.5%，但 PandaPose 的误差仅增加了 1.5%。同时，在正文图 9 中，我们逐步向二维姿态添加高斯噪声，以研究方法对噪声水平的容忍度，我们的方法保持了最佳抗噪声能力。当像素级噪声尺度从 0 到 5 时，PandaPose 的误差仅增加 16%（从 39.8 到 48.2），而基于图像的当前最优方法 CA-PF [55]增加了 56%（从 41.4 到 64.9），基于序列的当前最优方法 KTPFormer [28] 增加了 112%（从 40.1 到 86.9）。抗噪性可归因于不同的三维锚点设置和锚点到关节集成预测机制。

A.5 运行效率比较

表9 展示了我们提出的方法（PandaPose）、最先进的基于图像的方法（CA-PF [55]）与基于序列的方法（MixSTE [50] 和 MHFormer [20]）在 Human3.6M [13]上的效率比较。所有指标均在 NVIDIA RTX3090 GPU 上评估。

方法	FPS	提升模块参数（M）	MPJPE ↓（完整）	MPJPE ↓（困难）
基于序列				
MHFormer（T=27）[20]	-	24.8	45.9	-
MHFormer（T=351）[20]	-	24.8	43.0	-
MixSTE(T=243) [50]	-	33.6	40.9	-
基于图像				
CA-PF [55]	19	14.1	41.4	82.4
PandaPose，使用 3 个解码器	16	15.2	**39.8**	**73.1**
PandaPose，使用 2 个解码器	18	12.7	**40.3**	**75.6**

表 9：运行效率比较。

二维姿态估计器	COCO 上的平均精度均值↑	Human 3.6M 上的 MPJPE ↓
CPN	68.6	40.1
HR-Net-w32	74.4	39.8
HR-Net-w48	76.3	39.5

表 10：COCO 和 Human 3.6M 数据集上二维姿态估计器的比较。

由于实验设置存在差异，我们省略基于序列的方法的帧率指标。与基于序列的方法相比，PandaPose 在保持性能相当的同时，使用的参数更少。为展示 PandaPose 的优势，我们将解码器的层数从 3 层减少到 2 层。此调整使参数量减少了 2.5M，在完整测试集上的性能仅下降 0.5mm，在困难子集上仅下降 2.5mm。与 CA-PF [55], 相比，PandaPose 在保持运行效率相当的同时，参数量减少了 1.4M。值得注意的是，在困难子集上，PandaPose 的 MPJPE 比 CA-PF 低 6.8mm。这也表明，PandaPose 在进一步提升效率的同时，仍具有实现优异性能的巨大潜力。

A.6 更多可视化结果

我们在图 14中展示了更多可视化案例，并将其与最先进的基于图像的方法（CA-PF [55]）和真值进行比较。在处理严重自遮挡或含噪二维姿态输入时，我们的方法能够更出色地处理关节之间的相对深度关系，从而生成更准确的三维姿态预测。

此外，图 15 展示了一些具有代表性的失败案例。在存在严重运动模糊或自遮挡的场景中，人体可能会与背景混淆，从而导致二维姿态预测不准确。因此，三维姿态预测的质量将受到影响。由于显式建模了关节之间的相对深度关系，并融合了锚点到关节集成预测，我们的方法能够在此类场景中预测出相对更加合理的三维人体姿态。

A.7 局限性与更广泛的影响

尽管我们的方法在基于图像的三维人体姿态估计方面取得了成果，但仍存在一些有待进一步改进的局限性。作为一种单帧方法，与利用相邻帧信息的基于序列的方法相比，PandaPose 缺乏时间平滑性。这可能导致连续动作中的姿态估计出现抖动，尤其是在动作快速或复杂的情况下。此外，PandaPose 引入了图像特征处理，以及包括特征提升和三维锚点空间中的集成预测在内的复杂操作。虽然这些步骤能够提高准确性，但不可避免地会带来计算资源开销。未来的研究将致力于开发性能相近但计算和内存成本更低的轻量化方法，并探索高效的特征提升与优化方法。我们的方法在训练过程中仅使用公开可用的数据集，因此不会产生广泛的社会影响，不涉及人工智能伦理，也不涉及任何隐私敏感数据。