HyperDiff：超图引导的 3D 人体姿态估计算法模型

Bing Han, Yuhua Huang*, and Pan Gao*
Nanjing University of Aeronautics and Astronautics, Nanjing, China
{icehan@nuaa.edu.cn, hyuhua2k@163.com, gaopan.1005@gmail.com}

摘要—单目三维人体姿态估计（HPE）在从 2D 到 3D 提升的过程中经常会遇到深度模糊和遮挡等挑战。此外，传统方法在利用骨骼结构信息时可能会忽略多尺度的骨架特征，这可能对姿态估计的准确性产生负面影响。为了解决这些问题，本文介绍了一种新颖的三维姿态估计方法 HyperDiff，该方法将扩散模型与 HyperGCN 相结合。扩散模型有效捕捉数据不确定性，缓解深度模糊和遮挡问题。同时，作为去噪器的 HyperGCN 采用多粒度结构准确建模关节之间的高阶相关性，这提高了模型处理复杂姿势时的去噪能力。实验结果表明，HyperDiff 在 Human3.6M 和 MPI-INF-3DHP 数据集上实现了最先进的性能，并且可以灵活适应不同的计算资源以平衡性能和效率。代码发布于 https://github.com/IHENL/HyperDiff

***Index Terms*—3D 人体姿态估计，扩散，图卷积网络**

I. 介绍

单目三维人体姿态估计（HPE）旨在从二维图像或视频序列中预测人类关节在三维空间中的位置。作为各种下游视觉应用的基本任务，它在虚拟现实 [1]–[3]、人机交互 [4]–[6] 和自动驾驶 [7] 等领域发挥着关键作用。传统方法通常将三维 HPE 过程分解为两个阶段：（1）二维关节估计，即现有的二维关键点检测器从 RGB 图像中识别出二维关键点 [8]–[11]，以及（2）二维到三维的提升，即将检测到的二维关键点映射到三维姿态。本文主要关注第二阶段——二维到三维的提升，旨在根据二维关键点估计准确的三维姿态。

准确估计从 2D 坐标到 3D 关节位置仍然面临挑战，由于固有的深度模糊和频繁的自我遮挡。此外，在单目数据中人体运动和身体构型的高度变异性在提升过程中引入了显著的不确定性。为了解决这些问题，越来越多的研究探索了使用扩散模型 [12], [13] 的应用。扩散模型 [14]–[16] 逐步向真实数据添加噪声并在生成过程中去除它，有效建模数据不确定性并缓解深度模糊和遮挡问题。然而，现有的基于扩散的方法通常在输入模型之前将 2D 姿态展平化，这限制了结构骨架信息的充分利用。此外，虽然人类骨架可以自然地使用图结构 [17]–[19] 表示，并且图卷积网络（GCNs）[20], [21] 适当地捕捉关节交互，大多数现有的基于 GCN 的方法 [17], [22]–[24] 仅关注关节层面的信息，忽略了更广泛的生理结构。

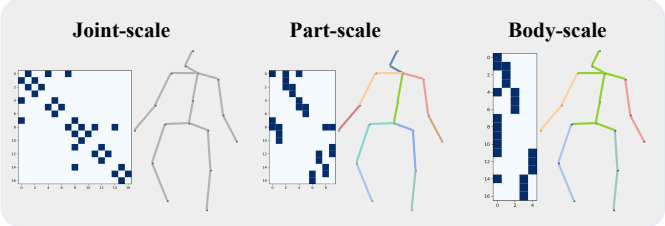

图 1. 基于图和超图的人体骨架表示。

本研究提出了一种新的扩散模型来建模数据不确定性，其中它协同使用 HyperGCN [25]–[28] 来增强模型捕捉和保存骨骼结构详细信息的能力。具体来说，我们训练了一个空间 HyperGCN 来去除受污染的 3D 姿势中的噪声，条件是基于 2D 关键点。结合重建目标，去噪器隐式地捕获了帧内关节之间的空间相关性。此外，我们通过引入更细粒度的空间尺度来增强去噪器，将人体骨骼划分为三个层次 [28], [29]：联合尺度图、部分规模超图和体规模超图，如图 1 所示。这种分割使得能够更精确地建模局部关节结构，从而提高去噪器的性能。这种新颖的基于扩散的人体姿态估计方法被称为**超差分**（**超距**图引导**差异**合模型）。通过建模高阶联合相关性而不破坏骨架，HyperDiff 生成更精确的三维姿态。总结来说，本文的主要贡献包括：

- 我们提出了 HyperDiff 框架，该框架利用 Hyper-

*Corresponding author

GCN 作为扩散模型的去噪器。这有效地解决了三维姿态估计中的不确定性和深度模糊问题。

- 我们进一步引入了一种多粒度 HyperGCN 结构，该结构优化了高阶结构建模，并提高了去噪器的性能，从而增强了姿态精度和结构表现力。
- 超级差分在 Human3.6M 和 MPI-INF-3DHP 数据集上取得了最先进的结果。此外，它通过调整去噪和迭代步骤的数量，在不同的计算资源条件下平衡了性能和效率。

II. 方法

A. *HyperGCN* 战略

我们首先介绍本研究采用 HyperGCN 的动机。传统的扩散方法 [14] 直接将 3D 关节作为输入，缺乏先验的骨架结构信息。由于关节对之间的关系复杂且密集，建模它们之间的依赖性具有挑战性，这使得优化任务变得复杂。为了解决这些问题，我们认为人体骨架可以表示为图，并且 GCN 能够更有效地捕捉关节间的相互作用。然而，大多数现有的基于 GCN 的方法依赖于单尺度的关节信息来提取骨架特征，忽略了有价值的多尺度上下文信息。为此，我们根据人体的动态链结构构建了两种类型的超图：部分尺度的超图和身体尺度的超图，定义如下：

$$
\begin{aligned}
&p1 = \{hip, spine, thorax\} \quad p2 = \{thorax, neck, head\} \\
&p3 = \{hip, rhip, rknee\} \qquad p4 = \{rknee, rfoot\} \\
&p5 = \{hip, lhip, lknee\} \qquad p6 = \{lknee, lfoot\} \\
&p7 = \{relbow, rwrist\} \qquad p8 = \{lelbow, lwrist\} \\
&p9 = \{thorax, rshoulder, relbow\} \\
&p10 = \{thorax, lshoulder, lelbow\} \\
&b1 = \{hip, rhip, lhip, spine, thorax, neck, head, \\
&\quad lshoulder, rshoulder\} \\
&b2 = \{rhip, rknee, rfoot\} \quad b3 = \{lhip, lknee, lfoot\} \\
&b4 = \{rshoulder, relbow, rwrist\} \\
&b5 = \{lshoulder, lelbow, lwrist\}
\end{aligned}
\tag{1}
$$

这里，p_i 表示包含在部件级超图 H_{part} 的超边中的关节，而 b_i 表示包含在身体级超图 H_{body} 的超边中的关节。

具体来说，关节尺度图关注的是关节之间的依赖关系，有效建模它们的直接连接和局部依赖关系，从而捕捉关节间的局部关系。部分尺度超图通过定义不同身体部位之间的关系，进一步增强了对局部关节依赖性的建模。另一方面，全身尺度超图则考虑更广泛的全局关系，能够捕捉不同身体区域之间的相互依赖性，从而提供对人体解剖结构的更为全面的建模。这种分层划分使模型能够准确地在局部和全局尺度上建模关节间的依赖关系。它还通过捕捉更高阶和更复杂的关节交互来增强骨骼结构的表现力。因此，三维姿态估计的精度和优化得到了提升。

B. 基于扩散的估计策略

在本节中，我们概述了 HyperDiff 的整体扩散策略。扩散模型通过前向和后向过程来模拟数据分布。HyperDiff 定义的前向过程是逐渐将依赖时间步的高斯噪声 $\epsilon \sim \mathcal{N}(0, I)$ 添加到真实 3D 姿态 $y_0 \in \mathbb{R}^{J\times 3}$ 上。遵循 DDPM [12]，这个过程表示为：

$$
y_t = \sqrt{\bar{\alpha}_t} y_0 + \epsilon\sqrt{1-\bar{\alpha}_t} \tag{2}
$$

其中的时间步是 $t = 1, .., T$。$\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$ 和 $\alpha_t := 1 - \beta_t$。$\{\beta_t\}_{t=1}^{T}$ 是余弦噪声调度。

在逆过程中，带有噪声的 3D 姿态 $y_t \in \mathbb{R}^{J\times 3}$ 被输入到一个去噪器 $\mathfrak{D}$ 中，该去噪器基于 2D 姿态 $x \in \mathbb{R}^{J\times 2}$ 和时间步长 t 生成去噪后的 3D 姿态 $\hat{y}_0$：

$$
\hat{y}_0 = \mathfrak{D}(y_t, x, t) \tag{3}
$$

在推理过程中，我们从高斯噪声 $\mathcal{N}(0, 1)$ 中采样 H 个初始噪声姿态 $y_{0:H,t}$。这些姿态使用训练好的 $\mathfrak{D}$ 进行去噪，以获得 $\hat{y}_{0:H,0}$。遵循 DDIM [13]，我们从 $\hat{y}_{0:H,0}$ 生成下一迭代的带噪样本 $y_{0:H,t'}$，这些样本用作时间步长 t' 处去噪器的输入。这可以形式化为：

$$
y_{0:H,t'} = \sqrt{\bar{\alpha}_{t'}}\hat{y}_{0:H,0} + \epsilon_t\sqrt{1+\bar{\alpha}_t-\sigma_t^2} + \sigma_t\epsilon \tag{4}
$$

其中 t 和 t' 分别是当前和下一个时间步，t 的范围是从 T 到 1。$\epsilon \sim \mathcal{N}(0, I)$ 是与 $y_{0:H,t}$ 独立的标准高斯噪声，

$$
\begin{aligned}
\epsilon_t &= (y_{0:H,t} - \sqrt{\bar{\alpha}_t}\hat{y}_{0:H,0})/\sqrt{1-\bar{\alpha}_t} \\
\sigma_t &= \sqrt{(1-\bar{\alpha}_{t'})/(1-\bar{\alpha}_t)} \cdot \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t'}}
\end{aligned}
\tag{5}
$$

其中 ϵ_t 表示时间步 t 的噪声，而 σ_t 是控制扩散过程中随机性程度的参数。

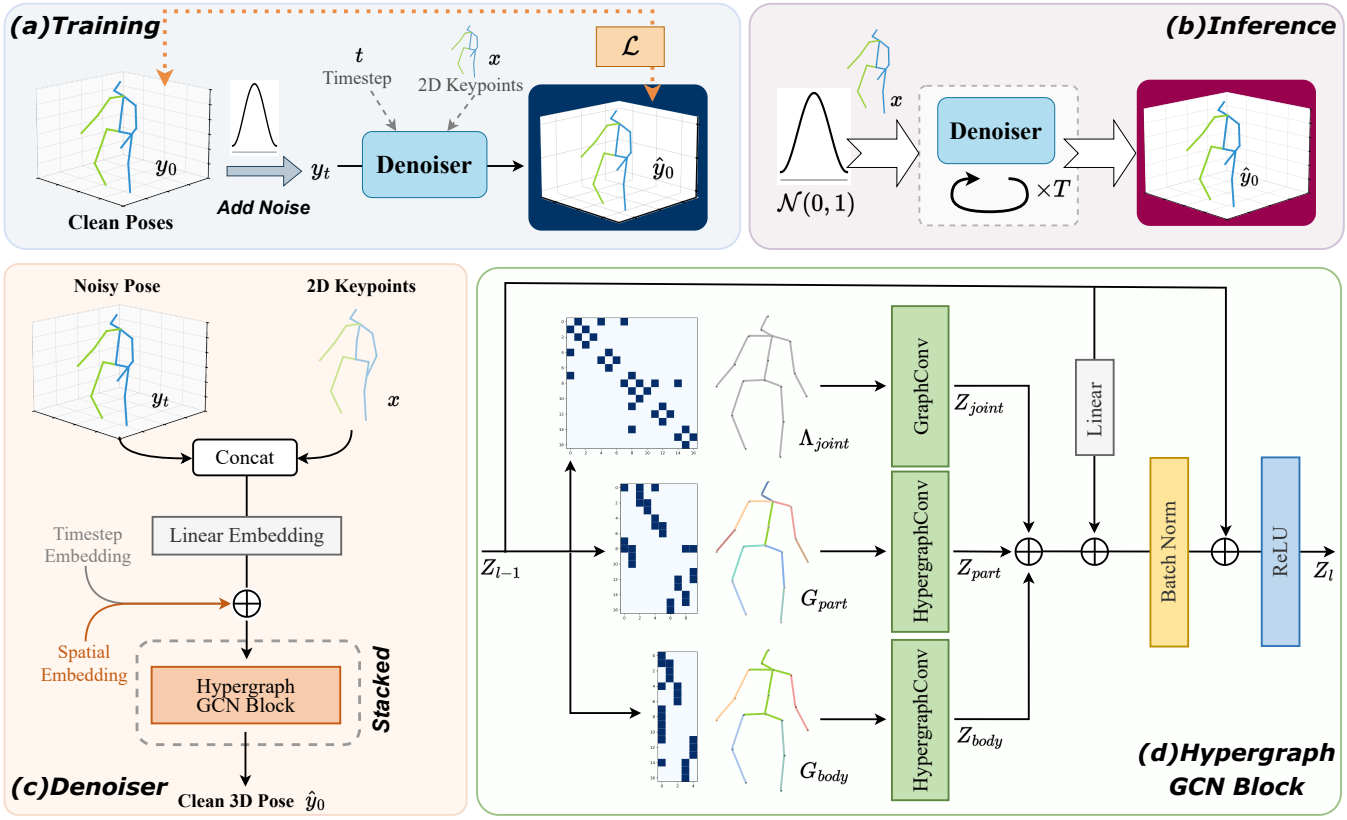

图 2. 提出的 HyperDiff 框架概述。(a) 和 (b) 展示了训练和推理过程。(c) 显示了去噪器的架构，而 (d) 描绘了所提出的超图 GCN 块结构。

迭代过程重复 K 次，从时间步长 T 开始。每次迭代的时间步长由 $t = T \cdot \left(1 - \frac{k}{K}\right)$，$k \in [0, K)$ 决定。

C. 去噪器架构设计

本节介绍了基于超图卷积的去噪器架构。如图 2(c) 所示，给定输入的噪声姿态 $y_t \in \mathbb{R}^{J\times 3}$，我们将其与对应的 2D 姿态 $x \in \mathbb{R}^{J\times 2}$ 拼接起来，结果为 $y'_t \in \mathbb{R}^{J\times(3+2)}$。然后应用线性嵌入将特征维度投影到 d_m，并添加空间和时间步长嵌入以获得嵌入标记 $Z \in \mathbb{R}^{J\times d_m}$。接下来，将 Z 输入堆叠的超图 GCN 块中学习多级空间表示。最后，使用一个投影头将特征转换为干净的 3D 姿态 $\hat{y}_0 \in \mathbb{R}^{J\times 3}$。

最近的研究表明，基于 GCN 的架构在从 2D 关键点推断准确的 3D 关节位置的空间表示建模中非常有效（ [17], [24], [30]）。标准图卷积操作定义为：

$$GCN(Z) = \sigma(D^{-\frac{1}{2}}\widetilde{A}D^{-\frac{1}{2}}ZW) = \sigma(\Lambda ZW) \quad (6)$$

其中 σ 是激活函数，$\widetilde{A} = A + I$ 是有自环的邻接矩阵，D 是度矩阵，而 W 是可学习权重矩阵。为了简化符号，我们定义 $\Lambda = D^{-\frac{1}{2}}\widetilde{A}D^{-\frac{1}{2}}$ 为图卷积核。

随着图卷积的发展，研究人员将这种操作扩展到了超图 [26], [31]。给定在 II-A 中定义的超图，超图卷积可以在多个尺度上聚合骨架信息。超图卷积操作定义如下：

$$HGCN(Z) = \sigma(D_v^{-\frac{1}{2}}HMD_e^{-1}H^TD_v^{-\frac{1}{2}}ZW) = \sigma(GZW)) \quad (7)$$

其中，H 是超图的邻接矩阵，D_e 是超边度数的对角矩阵，D_v 是顶点度数的对角矩阵，W 是可学习的节点嵌入权重矩阵，M 是可学习的超边权重矩阵，初始设置为单位矩阵（即所有超边权重相等）。为了简化，我们使用 $G = D_v^{-\frac{1}{2}}HMD_e^{-1}H^TD_v^{-\frac{1}{2}}$ 作为超图卷积核。

如图 2(d) 所示，超图 GCN 模块首先将输入特征 $Z_{l-1} \in \mathbb{R}^{J\times d_m}$ 通过三个卷积分支：

$$\begin{aligned} Z_{joint} &= \sigma(\Lambda_{joint}Z_{l-1}W_{joint}) \\ Z_{part} &= \sigma(G_{part}Z_{l-1}W_{part}) \\ Z_{body} &= \sigma(G_{body}Z_{l-1}W_{body}) \end{aligned} \quad (8)$$

其中 Z_{joint}, Z_{part} 和 Z_{body} 代表在不同尺度下获得的骨架特征，G_{part} 和 G_{body} 是部分尺度和身体尺度下的超

图卷积核，而 W_{part} 和 W_{body} 则是可学习的权重矩阵。这三个分支的特征然后通过元素级求和融合：

$$Z = \alpha_{joint} \cdot Z_{joint} + \alpha_{part} \cdot Z_{part} + \alpha_{body} \cdot Z_{body} \quad (9)$$

其中 $\alpha_{joint}, \alpha_{part}$ 和 α_{body} 是可学习权重。这种加权融合允许模型全面整合来自局部、部分和全局尺度的信息，从而更有效地捕捉联合依赖关系。最后，$\mathrm{ReLU}(Z_{l-1} + \mathrm{BN}(Z + \mathrm{Linear}(Z_{l-1})))$ 产生 Z_l，它包含丰富的骨架特征信息，作为超图 GCN 块的输出用于后续操作。

D. 训练目标

为了确保空间信息的有效学习，我们使用真实值和估计姿态之间的均方误差（MSE）损失来监督框架，可以表示为：

$$\mathcal{L} = \|y_0 - \hat{y}_0\|_2 \quad (10)$$

其中 y_0 和 $\hat{y}_0$ 分别是真实值和估计的 3D 姿态。

III. 实验

A. 数据集和度量标准

我们在两个广泛使用的数据集 **Human3.6M (H36M)** [42] 和 **MPI-INF-3DHP（3DHP）** 上评估了我们的方法。H36M 是一个最常用的大型室内 3D HPE 数据集之一。它包含 360 万帧视频，并涵盖由 11 名专业被试者进行的 15 种不同活动，通过四台同步和校准的相机以每秒 50 帧的速度捕捉。与先前的工作 [14], [15] 类似，我们的模型在五个被试者（S1, S5, S6, S7, S8）上进行训练，并在两个被试者（S9, S11）上进行评估。我们在该基准测试中报告了平均每个关节位置误差 (MPJPE) 和 Procrustes 对齐的平均每个关节位置误差 (P-MPJPE)。3DHP 也是一个公开的大型数据集。此数据集包含训练子集中 8 名演员在 8 种活动中以及评估子集中 7 种活动下的三种不同设置：绿幕、非绿幕和户外环境。根据 [40], [43], [44]，我们计算 MPJPE，在 150 毫米阈值下的正确关键点百分比 (PCK) 和曲线下的面积 (AUC)。

B. 定量结果

1) ***H36M****:* 如表 I 所示，我们在 Human3.6M 数据集上将提出的 HyperDiff 方法与其他最先进的方法进行了比较。当使用检测到的 2D 姿势和真实的 2D 姿势作为输入时，我们的方法分别达到了 46.8 毫米和 28.6 毫米（$H = 1, K = 1$），仅排名第二，仅次于整合了额外视觉信息的 LiftingByImage [36]。此外，当应用来自 D3DP 的多假设策略 [14] 时，我们的模型在准确性方面超过了 LiftingByImage（$H = 10, K = 5$）。这些结果表明，将扩散模型与 HyperGCN 结合可以有效地捕捉关节之间的高阶依赖关系，从而显著提高姿态估计的准确性。

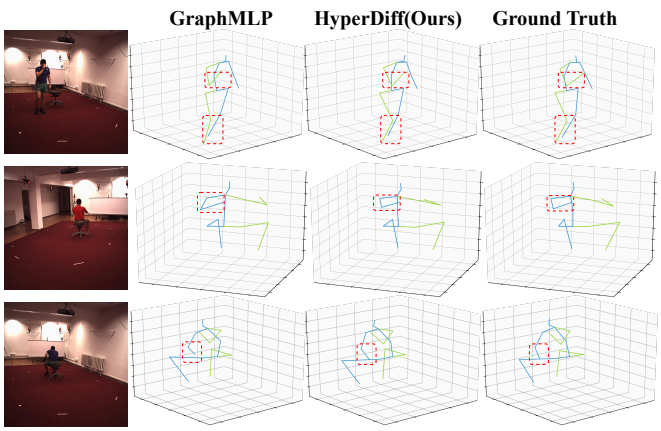

图 3. 定性结果在 Human3.6M 上。蓝色/绿色线表示左侧/右侧骨架的估计姿态序列。

2) **三维人体姿势***:* 如表 II 所示，我们提出的方法在 MPI-INF-3DHP 数据集上优于现有的最先进的方法，特别是在 PCK 和 AUC 指标上。采用 $H = 20, K = 10$ 配置时，我们的模型达到 88.4%PCK 和 58.7%AUC，超越了 LiftingByImage 及其他比较方法。这些结果表明更高的准确性和更强的泛化能力，显示出在多种场景下的稳定性能。

3) **效率***:* 如表 III 所示，我们的方法表现出更优的计算复杂度和效率。通过使用 $H = 1, K = 1$，并且模型参数和浮点运算次数少于 D3DP，我们的模型实现了每秒 30289 帧的更快推理速度，同时保持 MPJPE 为 46.8 毫米。虽然在更大配置下速度有所下降，但我们的方法在精度方面仍然优于其他方法。因此，它有效地平衡了准确性和效率，使其适合适应实时应用。

C. 定性结果

图 3 显示了 HyperDiff 与 GraphMLP [17] 在 Human3.6M 数据集上的定性比较。利用多尺度骨架结构，我们的方法更有效地处理复杂的场景，如自我遮挡。此外，图 4 展示了 HyperDiff 在更具挑战性的野外图像上的定性结果，以评估我们模型的泛化能力。需要注意的是，这些自然视频中的动作在训练集中相对罕见或不存

表 I

HUMAN3.6M 上所提方法与最先进方法的 MPJPE 比较。顶部组使用检测到的 2D 姿态作为输入，底部组使用 GROUND TRUTH 的 2D 姿态作为输入。最佳和次佳结果分别标记为红色 和蓝色 。T, H, K: 输入帧数、假设数量和 D3DP [14] 的迭代次数。(‡) 表示使用视觉线索。

MPJPE	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Pur	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	Avg
MGCN [30]	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46	57.5	63	49.7	46.6	52.2	38.9	40.8	49.4
Graformer [24]	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	39.7	43.1	51.8
POT [32]	47.9	50	47.1	51.3	51.2	59.5	48.7	46.9	56	61.9	51.1	48.9	54.3	40	42.9	50.5
DiffPose [33]	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52	37.6	40.2	49.7
RS-Net [34]	44.7	48.4	44.8	49.7	49.6	58.2	47.4	44.8	55.2	59.7	49.3	46.4	51.4	38.6	40.6	48.6
Di2Pose [35]	41.9	47.8	45.0	49.0	51.5	62.2	45.7	45.6	57.6	67.1	50.1	45.3	51.4	37.3	40.9	49.2
LiftingByImage [36]‡	44.9	46.4	42.4	44.9	48.7	40.1	44.3	55	58.9	47.1	48.2	42.6	36.9	48.8	40.1	46.4
GraphMLP [17]	43.7	49.3	45.5	47.9	50.5	56.0	46.3	44.1	55.9	59.0	48.4	45.7	51.2	37.1	39.1	48.0
Ours ($H=1, K=1$)	44.1	44.8	41.5	48.6	45.4	53.2	44.2	45.2	48.2	58.5	46.0	45.5	53.1	41.0	43.1	46.8
Ours ($H=10, K=5$)	42.8	44.0	40.8	47.8	44.3	52.0	43.4	44.4	47.4	57.9	45.2	44.8	52.1	40.4	41.8	46.0
GraphSH [37]	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
HGN [38]	35.4	40.2	31.1	38.2	38.3	41.1	36.1	32.7	42.1	48.4	37.1	36.9	37.1	30.5	32.4	37.2
PHGANet [39]	32.4	36.5	30.1	33.3	36.3	43.5	36.1	30.5	37.5	45.3	33.8	35.1	35.3	27.5	30.2	34.9
DiffPose [33]	28.8	32.7	27.8	30.9	32.8	38.9	32.2	28.3	33.3	41.0	31.0	32.1	31.5	25.9	27.5	31.6
LiftingByImage [36]‡	29.5	30.1	25.0	29.0	28.5	28.6	26.9	30.5	31.1	27.7	32.4	27.7	24.8	30.0	25.9	28.6
GraphMLP [17]	32.2	38.2	29.3	33.4	33.5	38.1	38.2	31.7	37.3	38.5	34.2	36.1	35.5	28.0	29.3	34.2
PoseFormer [40]($T=81$)	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
MHFormer [41]($T=351$)	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
Ours ($H=1, K=1$)	29.9	31.9	24.0	29.1	27.4	29.5	33.5	28.4	27.4	29.1	27.5	30.1	29.6	24.8	26.6	28.6
Ours ($H=10, K=5$)	29.7	31.3	23.5	28.6	27.1	28.9	33.0	27.8	27.1	28.3	27.2	29.4	29.2	24.6	26.3	28.1

表 II

与最先进的单帧方法在 MPI-INF-3DHP 上的性能比较。

Method	PCK↑	AUC↑	MPJPE↓
Simple [45]	82.6	50.2	88.6
Cascaded [46]	81.2	46.1	99.7
MGCN [30]	86.1	53.7	-
POT [32]	84.1	53.7	-
LiftingByImage [36]‡	88.2	59.3	68.9
GraphMLP	87.0	54.3	-
Ours ($H=1, K=1$)	87.6	57.0	69.2
Ours ($H=20, K=10$)	88.4	58.7	68.5

表 III

计算复杂性和效率的比较。

Method	H	K	MPJPE	Param (M)	FLOPs (G)	FPS
Graformer [24]	1	1	51.8	0.65	0.702	21588
GraphMLP [17]	1	1	48.0	9.49	0.348	41143
D3DP [14]	1	1	48.9	34.71	1.152	15514
Ours	1	1	46.8	13.07	0.443	30289
Ours	5	5	46.1	13.07	11.053	866
Ours	10	5	46.0	13.07	22.102	412

表 IV

不同粒度图组合的消融研究。

配置	MPJPE↓
Baseline	48.9
Joint-scale	49.5(+0.6)
+Part-scale	48.2(-0.7)
+Body-scale	47.6(-1.3)
+Part-scale+Body-scale	**46.8**(-2.1)

表 V

特征融合策略的消融研究。

融合策略	MPJPE↓
Concate Fusion	49.4
Product Fusion	47.1
Weighted Fusion	**46.8**

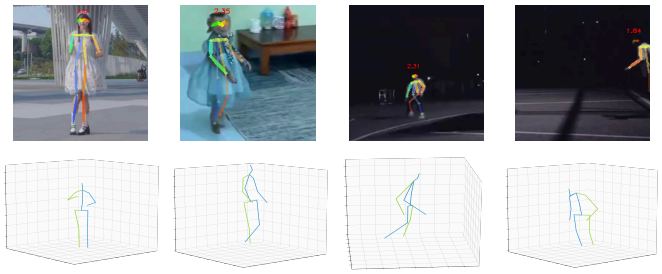

图 4. 我们的方法在野外视频中具有 3D 人体姿态的定性结果，涵盖多样且具挑战性的场景，包括深度模糊、自我遮挡、错误的 2D 检测和不完整的 2D 姿态。

在。例如，第一个实例展示了手部严重的深度模糊，第二个涉及自我遮挡，最后两个则显示了来自 2D 检测器的检测错误和不完整的检测。尽管存在这些挑战，我们的方法仍表现出强大的泛化能力，通过利用学习到的骨架特征准确预测三维姿态。

D. 消融研究

1) **不同图尺度的组合***:* 如表 IV 所示，在不同层级中引入图结构显著提高了模型的性能。与基线（D3DP 在 $T=1$）相比，仅使用关节尺度图将 MPJPE 提高到了 49.5mm，表明单独使用关节级图对性能的影响有限。当添加部分级和全身级图时，MPJPE 分别减少了 0.7

毫米和 1.3 毫米，这表明更高层次的结构更好地建模了关节间的关系。值得注意的是，当同时使用部分级和全身级图时，MPJPE 达到了最低值 46.8 毫米。这表明结合局部和全局结构可以更全面地捕捉复杂的关节依赖关系，显著提升了姿态估计的准确性。

2) **特征融合策略***:* 表 V 说明了不同特征融合策略对模型性能的影响。连接融合产生的结果最差，因为基本的连接无法有效整合多层级图特征。相比之下，加权融合通过根据特征的重要性分配权重，实现了最佳性能，使模型能够更好地利用多层级图的优势。虽然乘积融合也有效地捕捉到了特征交互，但在融合多层级特征方面略逊于加权融合。

IV. 结论

本研究提出了 HyperDiff 方法，该方法利用扩散模型与 HyperGCN 相结合来解决 3D 人体姿态估计中的深度歧义和自遮挡问题。通过引入多粒度超图 GCN，HyperDiff 增强了对关节之间高阶依赖关系的建模，从而提高了姿态估计精度。实验结果表明，HyperDiff 在多个标准数据集上取得了优异性能。此外，它有效地平衡了计算效率与性能，在实时应用中表现出适用性。

参考文献

[1] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, "So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6961–6970, 2019.

[2] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan, "Model-based 3d hand reconstruction via self-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10451–10460, 2021.

[3] Y. Chen, Z. Tu, D. Kang, R. Chen, L. Bao, Z. Zhang, and J. Yuan, "Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 4008–4021, 2021.

[4] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang, "Synthesizing long-term 3d human motion and interaction in 3d scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9401–9411, 2021.

[5] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3d scenes by learning human-scene interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14708–14718, 2021.

[6] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2me: Inferring body pose in egocentric video via first and second person interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9890–9900, 2020.

[7] Q. Lu, W. Han, J. Ling, M. Wang, H. Chen, B. Varadarajan, and P. Covington, "Kemp: Keyframe-based hierarchical end-to-end deep model for long-term trajectory prediction," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 646–652, IEEE, 2022.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 483–499, Springer, 2016.

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.

[10] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7103–7112, 2018.

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[14] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, "Diffusion-based 3d human pose estimation with multi-hypothesis aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14761–14771, 2023.

[15] J. Xu, Y. Guo, and Y. Peng, "Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 561–570, 2024.

[16] Q. Cai, X. Hu, S. Hou, L. Yao, and Y. Huang, "Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 882–890, 2024.

[17] W. Li, M. Liu, H. Liu, T. Guo, T. Wang, H. Tang, and N. Sebe, "Graphmlp: A graph mlp-like architecture for 3d human pose estimation," *Pattern Recognition*, vol. 158, p. 110925, 2025.

[18] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, 2019.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[20] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[22] W. Hu, C. Zhang, F. Zhan, L. Zhang, and T.-T. Wong, "Conditional directed graph convolution for 3d human pose estimation," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 602–611, 2021.

[23] B. X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, and C. W. Chen, "Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8818–8829, 2023.

[24] W. Zhao, W. Wang, and Y. Tian, "Graformer: Graph-oriented transformer for 3d pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20438–20447, 2022.

[25] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[26] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis, and P. Talukdar, "Hypergcn: A new method for training graph convolutional networks on hypergraphs," *Advances in neural information processing systems*, vol. 32, 2019.

[27] J. Wei, Y. Wang, M. Guo, P. Lv, X. Yang, and M. Xu, "Dynamic hypergraph convolutional networks for skeleton-based action recognition," 2021.

[28] Y. Zhu, G. Huang, X. Xu, Y. Ji, and F. Shen, "Selective hypergraph convolutional networks for skeleton-based action recognition," in *Proceedings of the 2022 international conference on multimedia retrieval*, pp. 518–526, 2022.

[29] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 3558–3565, 2019.

[30] Z. Zou and W. Tang, "Modulated graph convolutional network for 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11477–11487, 2021.

[31] S. Bai, F. Zhang, and P. H. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognition*, vol. 110, p. 107637, 2021.

[32] H. Li, B. Shi, W. Dai, H. Zheng, B. Wang, Y. Sun, M. Guo, C. Li, J. Zou, and H. Xiong, "Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 1296–1304, 2023.

[33] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Diffpose: Toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13041–13051, 2023.

[34] M. T. Hassan and A. B. Hamza, "Regular splitting graph network for 3d human pose estimation," *IEEE Transactions on Image Processing*, vol. 32, pp. 4212–4222, 2023.

[35] W. Wang, J. Xiao, C. Wang, W. Liu, Z. Wang, and L. Chen, "Di^{2}Pose: Discrete diffusion model for occluded 3d human pose estimation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 98717–98741, 2024.

[36] F. Zhou, J. Yin, and P. Li, "Lifting by image–leveraging image cues for accurate 3d human pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 7632–7640, 2024.

[37] T. Xu and W. Takano, "Graph stacked hourglass networks for 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16105–16114, 2021.

[38] H. Li, B. Shi, W. Dai, Y. Chen, B. Wang, Y. Sun, M. Guo, C. Li, J. Zou, and H. Xiong, "Hierarchical graph networks for 3d human pose estimation," *arXiv preprint arXiv:2111.11927*, 2021.

[39] S. Zhang, C. Wang, L. Nie, H. Yao, Q. Huang, and Q. Tian, "Learning enriched hop-aware correlation for robust 3d human pose estimation," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1566–1583, 2023.

[40] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11656–11665, 2021.

[41] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13147–13156, 2022.

[42] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[43] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, "3d human pose estimation with spatio-temporal criss-cross attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4790–4799, 2023.

[44] J. Peng, Y. Zhou, and P. Mok, "Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1123–1132, 2024.

[45] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2640–2649, 2017.

[46] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3d human pose estimation with evolutionary training data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6173–6183, 2020.