Deep Semantic Graph Transformer for Multi-View 3D Human Pose Estimation

Lijun Zhang[1, 2], **Kangkang Zhou**[1, 2], **Feng Lu**[3, 4], **Xiang-Dong Zhou**[1, 2], **Yu Shi**[1, 2]

[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China
[2]Chongqing School, University of Chinese Academy of Sciences, Chongqing, China
[3]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4] Peng Cheng Laboratory, Shenzhen, China
{zhanglijun, zhouxiangdong, shiyu}@cigit.ac.cn, zhoukangkang21@mails.ucas.ac.cn, lf22@mails.tsinghua.edu.cn

Abstract

Most Graph Convolutional Networks based 3D human pose estimation (HPE) methods were involved in single-view 3D HPE and utilized certain spatial graphs, existing key problems such as depth ambiguity, insufficient feature representation, or limited receptive fields. To address these issues, we propose a multi-view 3D HPE framework based on deep semantic graph transformer, which adaptively learns and fuses multi-view significant semantic features of human nodes to improve 3D HPE performance. First, we propose a deep semantic graph transformer encoder to enrich spatial feature information. It deeply mines the position, spatial structure, and skeletal edge knowledge of joints and dynamically learns their correlations. Then, we build a progressive multi-view spatial-temporal feature fusion framework to mitigate joint depth uncertainty. To enhance the pose spatial representation, deep spatial semantic feature are interacted and fused across different viewpoints during monocular feature extraction. Furthermore, long-time relevant temporal dependencies are modeled and spatial-temporal information from all viewpoints is fused to intermediately supervise the depth. Extensive experiments on three 3D HPE benchmarks show that our method achieves state-of-the-art results. It can effectively enhance pose features, mitigate depth ambiguity in single-view 3D HPE, and improve 3D HPE performance without providing camera parameters. Codes and models are available at https://github.com/z0911k/SGraFormer.

Introduction

3D human pose estimation (HPE) is a popular research topic in computer vision. It is a crucial tool for analyzing human behavior since it is able to estimate human pose by predicting the locations of main human body joints in 3D space. As a result, it is the foundational technology for many human-assisted vision tasks, such as robotics, action recognition, pedestrian re-identification, and virtual/augmented reality.

With the advancement of deep learning techniques, 3D HPE methods based on Convolutional Neural Networks (CNNs) have risen to prominence. They are broadly characterized as direct estimation methods (Pavlakos et al. 2017; Luvizon, Tabia, and Picard 2019) and 2D-3D lifting methods (Tekin et al. 2017; Zhou et al. 2019). The latter performs better due to the intermediate supervision of 2D poses, which is the current mainstream. CNNs commonly concatenate the 2D joint coordinates directly as input features in the 2D-3D lifting approaches, ignoring the original spatial arrangement of human body joints. Since Graph Convolutional Networks (GCNs) have good performance when processing irregular graph data, researchers have introduced GCNs into 3D HPE and gained some achievements.

GCN-based 3D HPE methods are currently employed primarily in single-view 3D HPE (Cai et al. 2019; Zhao et al. 2019; Xu and Takano 2021; Zhang et al. 2022b), where graph features are commonly generated based on adjacency matrices of connected 2D joints. However, there is an inherent depth ambiguity problem with single-view 3D HPE, as a 2D pose may project multiple 3D poses. The utilization of multi-view information can mitigate the depth ambiguity problem, while few graph-based multi-view 3D HPE methods have evolved, hence this paper covers this direction.

Most GCN-based 3D HPE approaches (Pavllo et al. 2019; Zhao et al. 2019; Xu and Takano 2021; Liu et al. 2021) only consider the connections between joint points, without taking into account the original position and skeletal edge information of human joints, as well as their effective fusion. Additionally, the GCN has a limited receptive field, resulting in inadequate feature representation. Some work (Cai et al. 2019; Wang et al. 2020; Zeng et al. 2021; Zhang et al. 2022b) incorporated temporal information to enhance the feature and alleviate depth uncertainty, while it is difficult to build long-time dependencies. Transformer has addressed this issue and has been employed in several 3D HPE methods with decent results (Zheng et al. 2021; Shuai, Wu, and Liu 2022; Zhao et al. 2023; Li et al. 2023). However, these works directly transform and concatenate the coordinates of 2D points into input feature tokens, with no regard for node spatial structural information such as graphs. Only a few methods (Zhao, Wang, and Tian 2022; Ionescu et al. 2023) incorporated graph features into transformer, while they have limited model performance.

To address the above problems, we propose a multi-view 3D HPE method based on a deep semantic graph transformer. The network can dynamically learn deep semantic features and their correlations involving the position, spatial structure, and skeletal edge of all human joints. It progressively fuses significant spatial-temporal information across multiple viewpoints and successfully models the long-time

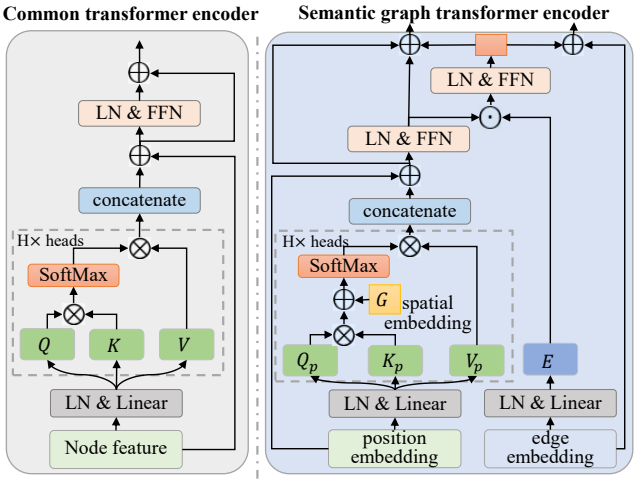

Figure 1: Our semantic graph transformer encoder over common transformer encoder, which dynamically learns the position, spatial structure, and skeletal edge knowledge of human joints, as well as their correlations.

dependencies of relative frames, with the goal of effectively alleviating the depth ambiguity problem of single-view 3D HPE and improving model performance.

Many graph-based 3D HPE methods get graph features by manually creating adjacency matrices, which are then fed into CNNs to predict the pose. These graphs mainly focus on the connections between joints while ignoring many significant details like joint locations or edges. The convolutional window has a limited receptive field, making it challenging to model long-time joint dependencies. Also, since its weights are independent of the input, there is no interaction between the graph features, which can be considered static. To deal with these, we improve the common transformer and propose a deep semantic graph transformer encoder by introducing several graph features conveying spatial information on joint position features, as shown in Figure 1. To extract deep hidden semantic knowledge, we simultaneously mine the position, spatial structure, and bone edge information associated with human nodes and generate a variety of relevant feature embeddings. Then, we establish a dynamic communication strategy among these features via a semantic attention mechanism, which fully exploits their correlations and performs effective fusion to enhance feature representation.

In order to reduce depth ambiguity and improve model performance, we build a progressive spatial-temporal feature fusion framework across multiple viewpoints. To make the spatial feature more expressive, we perform cross-view spatial feature fusion during monocular semantic feature extraction using multi-head attention between features of different views. The mutual supervision and interaction of spatial semantic knowledge from different viewpoints are utilized to rich the joint features. To supplement the depth information, the spatial and temporal features across multiple viewpoints are progressively fused, and long-time dependencies of relevant frames are dynamically learned and adopted. Extensive experiments demonstrate the efficacy of our method. It significantly mitigates the depth ambiguity problem of single-view 3D HPE and improves the accuracy of 3D pose prediction with the proposed graph features and fusion framework.

The main contributions of this paper are:

- We propose a deep semantic graph transformer encoder, which effectively enhances pose feature representation through deeply mining the position, spatial structure, and skeletal edge information of human joints, as well as learning their correlations dynamically.
- We build a progressive multi-view spatial-temporal feature fusion framework. The depth uncertainty of human joints is greatly reduced by performing feature fusion from spatial to temporal and modeling long-time dependencies of relevant images across multiple viewpoints.
- Extensive experiments on three popular 3D HPE benchmarks reveal that our method can outperform several state-of-the-art 3D HPE approaches, significantly mitigates the depth ambiguity problem of single-view 3D HPE and improves 3D HPE performance.

Related Works

CNN-Based 3D HPE Methods

According to different frameworks, CNN-based 3D HPE methods can be classified into direct estimation and 2D-3D lifting approaches. The direct estimation methods (Luvizon, Picard, and Tabia 2018; Luvizon, Tabia, and Picard 2019; Xiang, Joo, and Sheikh 2019) design an end-to-end network to directly infer 3D pose from the input image. The 2D to 3D lifting methods (Zhou et al. 2019; Cai et al. 2019; Pavllo et al. 2019; Yeh, Hu, and Schwing 2019; Zeng et al. 2021; Liu et al. 2020b) first utilize a 2D pose estimator to obtain the 2D pose, then adopt the 2D-3D lifting network to acquire 3D pose, which usually performs better due to intermediate supervision of 2D pose. Our method adheres to the 2D-3D lifting line, but improves several constraints of CNNs-based methods by combining the graph with transformer.

It can also be divided into single-view and multi-view 3D HPE based on camera view number. Single-view 3D HPE methods (Luvizon, Tabia, and Picard 2019; Zheng et al. 2021; Li et al. 2022b; Zeng et al. 2021) predict 3D pose from monocular images, which is an ill-posed problem with depth ambiguity during 2D-3D pose mapping. Multi-view 3D HPE methods (Shuai, Wu, and Liu 2022; He et al. 2020; Ma et al. 2021) have evolved to address this, because knowledge from various views can supplement the missing joint depth, yielding superior results in complex scenes with occlusion or camera motion. Some jobs (He et al. 2020; Xie, Wang, and Wang 2022; Wang et al. 2021; Iskakov et al. 2019) used epipolar geometry or triangulation to integrate multi-view 2D heatmaps while neglecting considerable joint semantic knowledge and requiring pre-providing camera parameters. Some (Bouazizi et al. 2021; Gholami et al. 2022; Kim et al. 2022) fused multi-view features only at the deep network levels, ignoring useful information at the shallow and medium network layers. Some (Iqbal, Molchanov, and Kautz 2020; Zhang et al. 2020) used complex loss functions, making model training challenging. We present a progressive multi-view feature fusion framework from spatial

to temporal using basic L2 loss, with no extrinsic camera parameters required during implementation.

Graph-Based 3D HPE Methods

Current graph-based 3D HPE work mostly employs GCNs to acquire graph features of 2D poses and then predict 3D poses, which is mainly used in single-view 3D HPE (Cai et al. 2019; Zhao et al. 2019; Liu, Zou, and Tang 2020; Liu et al. 2020a; Zou et al. 2020, 2021; Xu and Takano 2021; Liu et al. 2021; Zhang et al. 2022b, 2023a). However, most of these methods (Pavllo et al. 2019; Zhao et al. 2019; Cai et al. 2019; Xu and Takano 2021; Liu et al. 2021; Zhang et al. 2022b, 2023a) only consider certain structural information in the pose graph, ignoring numerous significant signals such as joint locations and bone edges. Some (Cai et al. 2019; Wang et al. 2020; Zeng et al. 2021; Liu et al. 2021; Zhang et al. 2022b, 2023a) use temporal information of related images to help identify the joint depth of the target image, although modeling long-time relationships is challenging. In contrast to these efforts, we construct a deep semantic graph transformer encoder that dynamically and adaptively learns the location, spatial structure, and skeletal edge properties of human joints. We also create a multi-view information fusion network capable of mining the spatial-temporal dependencies of human nodes in long-time related images.

Transformer-Based 3D HPE Methods

Due to the superior performance of transformer (Vaswani et al. 2017) in modeling long-range dependencies, 3D HPE work utilizing transformer has increasingly emerged. Currently, its primary application is in single-view 3D HPE (Zheng et al. 2021; Li et al. 2022b,a; Zhao, Wang, and Tian 2022; Zhao et al. 2023; Li et al. 2023; Gong et al. 2023; Shan et al. 2023), and only a few jobs are about multi-view (He et al. 2020; Ma et al. 2021; Shuai, Wu, and Liu 2022; Zhang et al. 2023b; Zhou et al. 2023). The input feature tokens of most these works are generally converted by 2D joint positions that ignore much spatial structure information of human nodes. A few single-view works (Zhao, Wang, and Tian 2022; Ionescu et al. 2023) mix graphs with transformer together, but they only evaluate certain structural messages and have restricted performance. Our work combines graphs with transformer networks. It first proposes a semantic graph transformer encoder that learns the position, structure, and edge features of human joints adaptively to enhance node spatial feature representation. A multi-view spatial-temporal feature fusion framework is also developed to address the depth ambiguity issue of single-view 3D HPE and improve model performance.

Method

The framework of the proposed method is illustrated in Figure 2. For the input image sequence $\mathcal{I} = \{I_i\}_{i=1}^{V\times T}$ with T frames from V views, we first use an offline 2D pose estimator to detect the 2D pose $P_{2D} \in \mathcal{R}^{T\times J\times 2}$ of the human body in each frame, and then input these 2D poses into the subsequent 2D-3D lifting network to estimate the 3D pose $P_{3D} \in \mathcal{R}^{T\times J\times 3}$ of the target image I_i. In our network, we first propose a deep semantic graph transformer encoder to fully extract the position, structure, and skeletal edge features involved in human joints, and utilize the attention mechanism to mine their correlations and dependencies to enhance the representation of spatial features. On this basis, we build a hierarchical multi-view information fusion framework to fully fuse the spatial and temporal features from multiple views, mitigate depth ambiguity of single-view 3D HPE, and enhance 3D pose prediction accuracy.

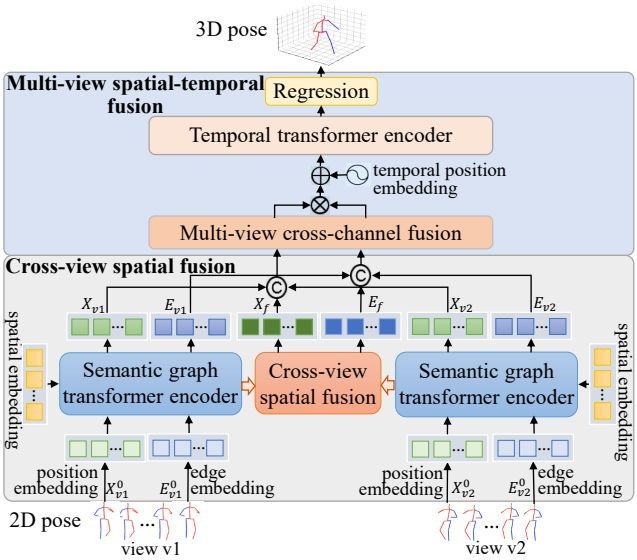

Figure 2: The architecture of our method. A deep semantic graph transformer encoder involving the position, spatial structure, and bone edge information of human joints is proposed to enhance spatial feature representation. A progressive multi-view spatial-temporal feature fusion framework is built to mitigate the depth ambiguity of single-view 3D HPE and improve 3D HPE performance.

Input Feature Embedding

Human body joints mainly involve the location of each joint, the spatial structure formed by all joints, and bones between connected joints. Most current graph-based 3D HPE work utilizes the adjacency matrices produced by the connections between joints to construct the graph feature, which only depicts part of the structural information of the human joints, while disregarding the influence of the position and bone edges. In order to enrich the spatial knowledge of pose features, we here consider the position, spatial structure, and skeleton edge feature embeddings of human nodes at the same time, and try to dynamically learn and mine their correlations. The following are the specifics of these features:

Node position embedding. The transformed features of the 2D joint coordinates obtained by the 2D pose estimator are defined as the node position embedding:

$$X = \varphi\left(\|_{i=1,j=1}^{T,J}\{P_{ij}\}\right), \tag{1}$$

where $P_{ij} = (x_{ij}, y_{ij})$, φ is a feature conversion function.

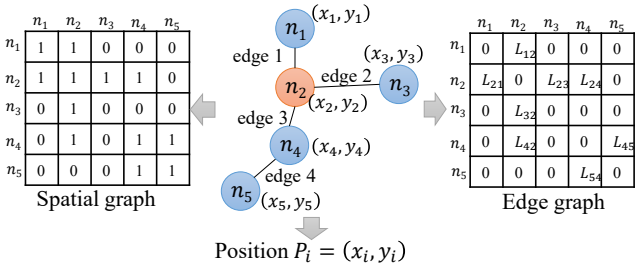

Figure 3: The position, spatial structure, and bone edge feature embeddings of five connected nodes.

It describes the initial spatial position information of each node ignored by many GCN-based 3D HPE approaches.

Spatial graph embedding. Based on connections of human joints, we construct multi-order graph features from global to local to describe the spatial structure of human joints. The global graph denotes the connected relationships between all nodes, whereas the local graph depicts specific special relationships such as similarity. For a 1-order graph adjacency matrix A, if there is a connection between node i and j, then its element $a_{ij} = 1$, otherwise $a_{ij} = 0$. These graph features serve as the node spatial embedding G and characterize their spatial arrangements, denoted as:

$$G = \|_{k=1}^{K} \sigma\left(W_k X \tilde{A}_k\right), \tag{2}$$

where $\tilde{A} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is the symmetric matrix of $\hat{A}$. $\hat{A} = A + I$, A is the interconnecting adjacency matrix, and I means self-connections. $\hat{D}$ is the normalized diagonal matrix of $\hat{A}$. $W = \{w_{ij}\}$ is a learnable weight matrix. σ denotes a nonlinear activation function, $\|$ is the concatenation of K kinds of global to local graph features.

Edge graph embedding. The bone edge connecting two nodes describes their spatial information as well. We build the edge graph adjacency matrix B based on whether there is a connected bone between two points, and use their distance as the corresponding element of this matrix. This feature is used as the node edge embedding E, depicting the skeletal connection and bone length between two connected joints.

$$E = \sigma\left(W X \tilde{B}\right), \tag{3}$$

where b_{ij} is the element of matrix B at position (i, j). If i and j is connected, $b_{ij} = L_{ij} = \|P_i - P_j\|^2$, else $b_{ij} = 0$.

Semantic Graph Transformer Encoder

We improve the common transformer encoder (Vaswani et al. 2017) and propose a semantic graph transformer encoder, as shown in Figure 1. In order to enhance the pose spatial feature representation, it deeply mines significant semantic information hidden in the position, spatial, and edge embeddings of all human joints, and dynamically builds an adaptive communication and fusion bridge between them.

The attention matrix can be analogous to a row-normalized adjacency matrix of a directed weighted complete graph. Unlike a static input graph, it aggregates input features dynamically using the attention mechanism. When producing the attention matrix, however, there is no direct way to merge the input spatial features. To solve the problem, we incorporate spatial embedding G into position embedding X and propose the semantic attention (SA), which is described as:

$$Q_p = \mathrm{LN}(\psi_1(X)), K_p = \mathrm{LN}(\psi_2(X)), V_p = \mathrm{LN}(\psi_3(X)), \tag{4}$$

$$\mathrm{SA}(Q_p, K_p, V_p) = Softmax\left(Q_p K_p^{\top}/\sqrt{d} + \sigma(G)\right) V_p, \tag{5}$$

where ψ is a linear layer, and d is the feature dimension. $\sigma(G)$ serves as a bias term to position features, indicating the combination of joint spatial and location knowledge. The multi-head semantic attention (MHSA) is also utilized to further enhance the feature, denoted as:

$$\mathrm{MHSA}(X) = \|_{h=1}^{H} \mathrm{SA}(Q_p^h, K_p^h, V_p^h), \tag{6}$$

where $\|$ is the concatenation of H attention heads. To make the feature more expressive, we further merge skeletal edge embeddings associated with each pair of connected joints into the position and spatial features. We transform their outputs passing through the MHSA and perform element-wise product with the layer-normalized and linear converted edge features. Residual connection is utilized to help in network training. Node features from layer $l-1$ to l are changed as:

$$X^{l\prime} = \mathrm{MHSA}(X^{l-1}) + X^{l-1}, \tag{7}$$

$$X^{l} = X^{l\prime} + \mu_p^l(X^{l\prime}) + \tau_p^l\left(\mu_p^l(X^{l\prime}) \odot \mathrm{LN}(\psi_p(E^{l-1}))\right), \tag{8}$$

$$E^{l} = E^{l-1} + \tau_e^l\left(\mu_e^l(X^{l\prime}) \odot \mathrm{LN}(\psi_e(E^{l-1}))\right), \tag{9}$$

where $\odot$ denotes element-wise product. μ and τ are different feature transformation functions in the feedforward network.

Progressive Multi-View Feature Fusion

Single-view 3D HPE suffers from severe depth ambiguity. Since a monocular image cannot determine the depth of human joints, a 2D pose may map several different 3D poses, making single-view 2D-3D lifting challenging. To address this issue, we take full advantage of the intermediate supervision of information from multiple viewpoints and design a progressive multi-view feature fusion framework from spatial to temporal, which alleviates depth uncertainty and improves 3D pose prediction accuracy.

Cross-view Spatial Fusion (CSF). To improve the spatial feature representation, we perform cross-view spatial feature fusion during the spatial semantic feature extraction of each individual viewpoint. The node and edge features are fused individually to mine richer unique information. Assume the output node features of layer l of view $v1$ and $v2$ are X_{v1} and X_{v2}, respectively. We first transform them using linear layers, and then feed them into the general transformer encoder for interaction and fusion. The converted X_{v1} is as the Q and K of the multi-head attention, while the converted X_{v2} is as the V. The fusion node feature is generated as:

$$X' = \mathrm{MHA}\left(\eta_1\left(X_{v1}\right), \eta_2\left(X_{v1}\right), \eta_3\left(X_{v2}\right)\right), \tag{10}$$

$$X'' = X' + X_{v1} + X_{v2}, \tag{11}$$

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Single-view methods																
(Ionescu et al. 2023)	47.9	50.0	47.1	51.3	51.2	59.5	48.7	46.9	56.0	61.9	51.1	48.9	54.3	40.0	42.9	50.5
(Zeng et al. 2021)	43.1	50.4	43.9	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	37.3	47.9
(Geng et al. 2023)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47.8
(Zhao et al. 2023)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
(Liu et al. 2020b)	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
(Zheng et al. 2021)	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
(Li et al. 2023)	39.1	42.7	38.7	40.3	44.1	50.0	41.4	38.7	53.9	61.6	43.6	40.8	42.5	29.6	30.6	42.5
(Zhang et al. 2022a)	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
(Gong et al. 2023)	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	**24.4**	**24.1**	36.9
(Ci et al. 2023)	**31.7**	**35.4**	**31.7**	**32.3**	**36.4**	**42.4**	**32.7**	**31.5**	**41.2**	**52.7**	**36.5**	**34.0**	**36.2**	29.5	30.2	**35.6**
Multi-view methods (camera parameters are **given**)																
(Kadkho. et al. 2021)	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
(Luvizon et al. 2022)(+)	31.0	33.0	41.0	34.0	41.0	37.0	37.0	51.0	56.0	43.0	44.0	37.0	33.0	42.0	32.0	39.0
(Bultmann and Behnke 2021)	27.1	29.9	27.0	26.5	31.3	28.9	27.1	29.8	36.5	36.0	30.8	29.3	29.7	27.3	26.3	29.8
(Bartol et al. 2022)	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1
(He et al. 2020)	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	28.8	31.7	28.2	26.4	23.6	28.3	23.5	26.9
(Qiu et al. 2019) (+)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	**26.9**	31.0	25.6	25.0	28.1	24.4	26.2
(Iskakov et al. 2019)	**19.9**	**20.0**	**18.9**	**18.5**	**20.5**	**19.4**	**18.4**	**22.1**	**22.5**	28.7	**21.2**	**20.8**	**19.7**	**22.1**	**20.2**	**20.8**
Multi-view methods (camera parameters are **not given**)																
(Luvizon et al. 2022)(+)	40.0	36.0	44.0	39.0	44.0	42.0	41.0	66.0	70.0	46.0	49.0	43.0	34.0	46.0	34.0	45.0
(Huang et al. 2020)	26.8	32.0	25.6	52.1	33.3	42.3	25.8	**25.9**	40.5	76.6	39.1	54.5	35.9	**25.1**	**24.2**	37.5
(Iskakov et al. 2019)	27.6	30.3	29.0	29.4	33.1	36.5	27.4	34.8	39.1	54.0	34.4	30.7	36.2	26.2	28.4	33.1
(Remelli et al. 2020)	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	**31.7**	31.2	29.9	**26.9**	33.7	30.4	30.2
(Gordon et al. 2022)	**22.0**	**23.6**	24.9	26.7	30.6	35.7	**25.1**	32.9	29.5	32.5	32.6	26.5	34.7	26.0	27.7	30.2
Ours (CPN, T=27)	26.5	28.3	**23.0**	**25.9**	**27.2**	**31.0**	25.4	27.2	**28.6**	33.8	**28.6**	**25.6**	30.1	27.1	26.5	**27.6**

Table 1: Comparisons with state-of-the-art 3D HPE methods on Human3.6M with P1 (mm) using the detected 2D poses. Our results are given when the temporal receptive field is under 27. (+) means using extra data. Best in bold.

$$X_f = \text{MLP}\left(\text{LN}\left(X''\right)\right) + X'', \tag{12}$$

where η is a linear layer. The X_f are concatenated with X_{v1} and X_{v2} to generate the final cross-view spatial features X_F, which is fed into the deeper network.

$$X_F = Concat(X_{v1}, X_f, X_{v1}). \tag{13}$$

The cross-view fusion edge features E_F can be obtained in the same way. This approach not only preserves the distinctive characteristics of each viewpoint, but also fully embeds fusion features across various viewpoints, resulting in richer hidden information extraction.

Multi-view Spatial-Temporal Fusion (MSTF). To further mitigate the depth ambiguity of 3D HPE, we develop a multi-view spatial-temporal fusion module. It profoundly integrates temporal knowledge of related images with spatial data from multiple viewpoints to supplement the missing depth message of human joints in the target image. For the cross-view fused features X_F and E_F, we first utilize a multi-view cross-channel fusion block (consisting of a batch normalization, 1x1 convolution layer, and layer normalization) to better preserve and refine the original spatial information, and convert these features to Y_X and Y_E. These two features are then multiplied and embedded with frame temporal position encoding E_{TPos} before being fed into the temporal transformer encoder for further spatial-temporal fusion. Finally, an MLP layer is utilized to regress the final pose features and predict the 3D pose. The spatial-temporal fusion feature Z is defined as:

$$Y' = \rho(Y_X \otimes Y_E) + E_{TPos}, \tag{14}$$

$$Y'' = \text{MHA}\left(\text{LN}\left(Y'\right)\right) + Y', \tag{15}$$

$$Z = \text{MLP}\left(\text{LN}\left(Y''\right)\right) + Y'', \tag{16}$$

in which ρ is a feature transformation function, and $\otimes$ is the dot product operation.

Loss Function. We train our model using only the basic Mean Squared Error (MSE) loss function without any bells and whistles, which minimizes the $L2$ distance error between the estimated human joint points and the corresponding ground-truth joint points, denoted as

$$\mathcal{L} = \sum_{i=1}^{T}\sum_{j=1}^{J}\left\|\hat{P}_{i,j}^{3D} - P_{i,j}^{3D}\right\|_2, \tag{17}$$

where $\hat{P}_{i,j}^{3D}$ and $P_{i,j}^{3D}$ denotes predicted and ground-truth 3D coordinates of the j-th node in the i-th frame, respectively.

Experiments

Datasets and Protocols

Human3.6M. (Ionescu et al. 2013) is the largest and most popular 3D HPE benchmark. It contains 3.6 million 3D human pose images and corresponding annotations captured by 4 synchronized cameras at 50Hz with different viewpoints in a controlled indoor environment. These data involves 15

Methods	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
(Gordon et al. 2022)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22.9
(Shuai, Wu, and Liu 2022)	15.5	17.1	13.7	15.5	14.0	16.2	15.8	16.5	15.8	16.1	14.5	14.5	16.9	14.3	13.7	15.3
Ours	**11.7**	**13.0**	**10.1**	**12.1**	**10.7**	**13.0**	**12.1**	**10.7**	**10.8**	**11.9**	**11.0**	**11.6**	**12.8**	**11.1**	**12.0**	**11.7**

Table 2: Comparisons with state-of-the-art multi-view 3D HPE methods on Human3.6M with P1 (mm) using the ground-truth 2D poses. Our results are given when the temporal receptive field is under 27. Best in bold.

action scenes performed by 11 professional actors, with 17 human nodes annotated in each image. Following previous works (Zheng et al. 2021; Li et al. 2022b, 2023), we use subjects S1, S5, S6, S7, and S8 for model training, S9 and S11 for model testing. The Protocol 1 (P1) and Protocol 2 (P2) are used to evaluate the validity of our models. P1 calculates the mean per joint positioning error (MPJPE) in millimeters, which is the Euclidean distance between the ground-truth and the predicted joint points. P2 indicates the Procrustes-MPJPE (P-MPJPE) error in millimeters, i.e., the MPJPE error between the predicted and ground-truth nodes after rigid alignment in terms of translation, rotation, and scale.

MPI-INF-3DHP. (Mehta et al. 2017) is a large-scale 3D human pose dataset from both indoor and outdoor scenes. It consists of more than 1.3 million images captured by 14 synchronized cameras from different viewpoints, recording 8 types of activities of 8 participants. 17 nodes of each image are annotated. Four chest views of S1-S6 are used for training, S7 and S8 are for testing. The Protocol 1, Protocol 2, Percentage of Correct Keypoints (PCK) with a threshold of 150 mm, and corresponding Area Under Curve (AUC) are used to evaluate the model.

Ski-Pose PTZ-Camera. (Fasel et al. 2016) is a smaller dataset with challenging in-the-wild images of alpine skiers performing giant slalom runs. It contains images of 6 subjects captured from 6 camera viewpoints. Following the official implementations, we use the subject 1-5 for model training, and subject 6 for model testing. The Protocol 1 and Protocol 2 are used for model evaluation.

Implementation Details

Our experiments are conducted on the PyTorch platform with 4 GeForce RTX 1080Ti GPUs. The Amsgrad optimizer is used with a weight decay of 0.1. For model training, the initial learning rate is 0.0002. The learning shrink factor after each epoch is $\alpha = 0.98$. When training the model, we set the maximum epoch and batch size to 50 and 1024, respectively. Four-order global-to-local spatial embedding graph features are considered. Four cascaded spatial and temporal transformer encoder layers are used in our framework, respectively. When using the detected 2D pose to obtain the 3D pose, we adopt the Cascaded Pyramid Network (CPN) (Chen et al. 2018) as the 2D pose detector. For all the three datasets, our models are trained just using the dataset themselves, without any other additional training data.

Comparison With State-of-the-art Methods

Results on Human3.6M. Table 1 shows our comparisons with state-of-the-art (SOTA) single-view and multi-view 3D HPE algorithms on Human3.6M. Our method outperforms all SOTA single-view 3D HPE methods, with MPJPE reduced by 8.0mm (22.5%) to (Ci et al. 2023). This finding suggests that intermediate supervision of multi-view information is helpful to reduce depth ambiguity of single-view 3D HPE and effectively enhances model performance. Our model surpasses several SOTA multi-view 3D HPE techniques that require camera calibration but performs somewhat worse than (Qiu et al. 2019) and (Iskakov et al. 2019). It indicates that while our method is competitive, methods using camera calibration still dominate in model performance. However, they are difficult in adjusting to various scenes because they are too reliant on camera settings. When compared with multi-view 3D HPE methods without pre-providing camera parameters, we achieve superior results, with MPJPE decreasing by 2.6mm (8.6%) compared to (Gordon et al. 2022). When given ground-truth 2D pose, as shown in Table 2, our model performance improves, with MPJPE 15.9mm (57.6%) and 3.6mm (23.5%) lower than Ours (CPN) and (Shuai, Wu, and Liu 2022), respectively. This implies that 2D pose is essential to the 2D-3D lifting and that better 2D poses facilitate the model. It is worth noting that our model was trained without any extra training data, using only the basic L2 loss function. These demonstrate how effective our method is.

Methods	Trainset	PCK	AUC	P1(mm)	P2(mm)
(Chen et al. 2021)	H36M	64.3	31.6	-	-
(Luvizon et al. 2022)	H36M+	80.6	42.1	112.1	-
(Iqbal et al.2020)	H36M+	80.2	-	110.8	-
(Kocabas et al. 2019)	3DHP	77.5	-	109.0	-
(Gholami et al. 2022)	3DHP	-	-	101.5	76.5
(Wandt et al. 2021)	3DHP	77.0	-	104.0	70.3
(Kocabas et al. 2019)	H36M	-	-	76.6	67.5
Ours	3DHP	98.7	90.2	16.9	12.1
	H36M	**99.9**	**91.7**	**10.6**	**7.6**

Table 3: Comparison results on 3DHP dataset. Best in bold.

Results on 3DHP. Table 3 compares our methods with relevant SOTA approaches on 3DHP. Two alternative scenarios have been explored. The first involves training and testing the model both on 3DHP. The other is finetuning the model trained on Human3.6M and testing it on 3DHP. Results depict that our method outperforms others in both scenarios and works better in the second case, with MPJPE and P-MPJPE decreasing by 66.0mm (86.2%) and 59.9mm (88.7%), respectively, in comparison to (Kocabas, Karagoz, and Akbas 2019). Because the Human3.6M dataset is larger

Methods	Trainset	P1(mm)	P2(mm)
(Chen et al. 2021)	H36M	130.2	108.7
(Wandt et al. 2021)	Ski	128.1	89.6
(Chen et al. 2021)	H36M+	99.4	74.7
(Rhodin et al. 2018)	Ski	85.0	-
(Gordon et al. 2022)	Ski	65.5	-
Ours	Ski	63.2	48.5
	H36M	**45.3**	**31.4**

Table 4: Comparisons on Ski-Pose dataset. Best in bold.

position	spatial	edge	P1(mm)	P2(mm)
✓	✓	✓	**27.6**	**21.8**
✓	✓	×	28.2	21.8
✓	×	✓	28.4	21.9
✓	×	×	29.4	22.7

Table 5: Impact of various features. Best in bold.

and contains more action categories, the second scenario allows the model to learn more about data types, actions, and scenes, bringing in better results. These show that our approach works for both indoor and outdoor datasets.

Results on Ski-Pose. Table 4 compares how well our algorithm performs against related methods on Ski-Pose. Scenarios of both training and testing on Ski-Pose, as well as finetuning the model trained on Human3.6M and testing on Ski-Pose, are studied. Results show that our method outperforms other approaches in both scenarios. In the second scenario, our MPJPE is 20.2mm (30.8%) lower than (Gordon et al. 2022), proving that the model benefits from more training data categories. In the first scenario, our method is also highly competitive, with MPJPE decreasing by 2.3mm (3.5%) compared to (Gordon et al. 2022). These show efficacy of our method in handling challenging in-the-wild data.

We present some qualitative results of our method on the three datasets in Figure 5, demonstrating the intuitive effectiveness of our approach in predicting 3D poses. It can be observed that our method performs well even for severe self-occluded poses and challenging complex poses.

Ablation Study

Impact of various features. Table 5 shows how the employed node position, structure, and edge feature embeddings affect the model. When all three features are utilized, the model performs best. The spatial embeddings, which provide main spatial structural knowledge of human joints, have a greater impact on the model than edge embeddings, while edge features also contribute. When we solely use joint position features, the model performs the worst, proving that the introduction of related spatial graph features can enrich joint information and improve feature representation. We have also shown attention maps of the three feature embeddings, i.e., position, spatial, and edge in Figure 4, which demonstrates how the pose information gradually becomes richer and more meaningful as more feature types are embedded, indicating the ability of our model to learn and utilize significant semantic information.

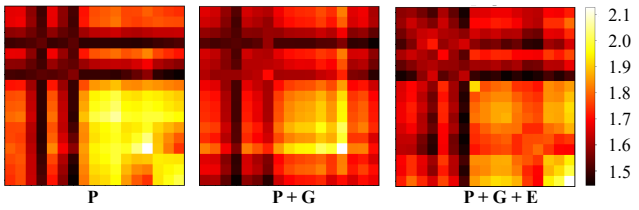

Figure 4: The attention map of different feature embeddings.

CSF	MSTF	Params(M)	FLOPs (G)	P1(mm)	P2(mm)
✓	✓	11.42	0.37	**27.6**	**21.8**
×	✓	11.39	0.35	29.0	23.1
✓	×	6.92	0.12	31.7	24.0

Table 6: Impact of fusion modules. Best in bold.

view number	Params(M)	FLOPs (G)	P1(mm)	P2(mm)
1	7.25	0.28	47.8	37.2
2	9.64	0.31	32.0	25.1
3	10.02	0.33	31.1	24.7
4	11.42	0.37	**27.6**	**21.8**

Table 7: Impacts of different viewpoints. Best in bold.

Impact of fusion modules. Table 6 displays the impacts of our proposed feature fusion modules, CSF and MSTF, on 3D HPE performance. Results reveal that removing any of the fusion modules worsens the model, emphasizing the importance of progressive multi-view feature fusion in improving 3D pose prediction. When MSTF is removed, MLP is used for feature conversion, and the model performs worse than when CSF is removed, indicating that spatial-temporal feature fusion has a greater influence on model improvement than simple spatial feature fusion, and that temporal knowledge of relevant frames is crucial for reducing depth ambiguity and improving model performance.

Impact of views. The effects of various input numbers of viewpoint information on model performance are depicted in Table 7. Our model improves steadily as the number of viewpoint increases, and it performs best when the viewpoint number is 4 (since the Human3.6M data contains a maximum of 4 viewpoints, we set the maximum number of viewpoints to 4 here). It suggests that intermediate supervision of multi-view data can successfully compensate for missing joint depth in single-view 3D HPE and boost the model. Furthermore, it illustrates that our framework is applicable for multi-view data fusion, capable of accepting an unlimited number of viewpoints.

Impact of temporal receptive fields (TRFs). We explore the impact of various TRFs on the model in Figure 6. It can be observed that when TRF grows, the model gradually improves. When TRF reaches 81, the model is essentially saturated, and further increases will not result in significant performance advancements. It implies that while temporal data promotes the model, utilizing a large TRF necessitates more

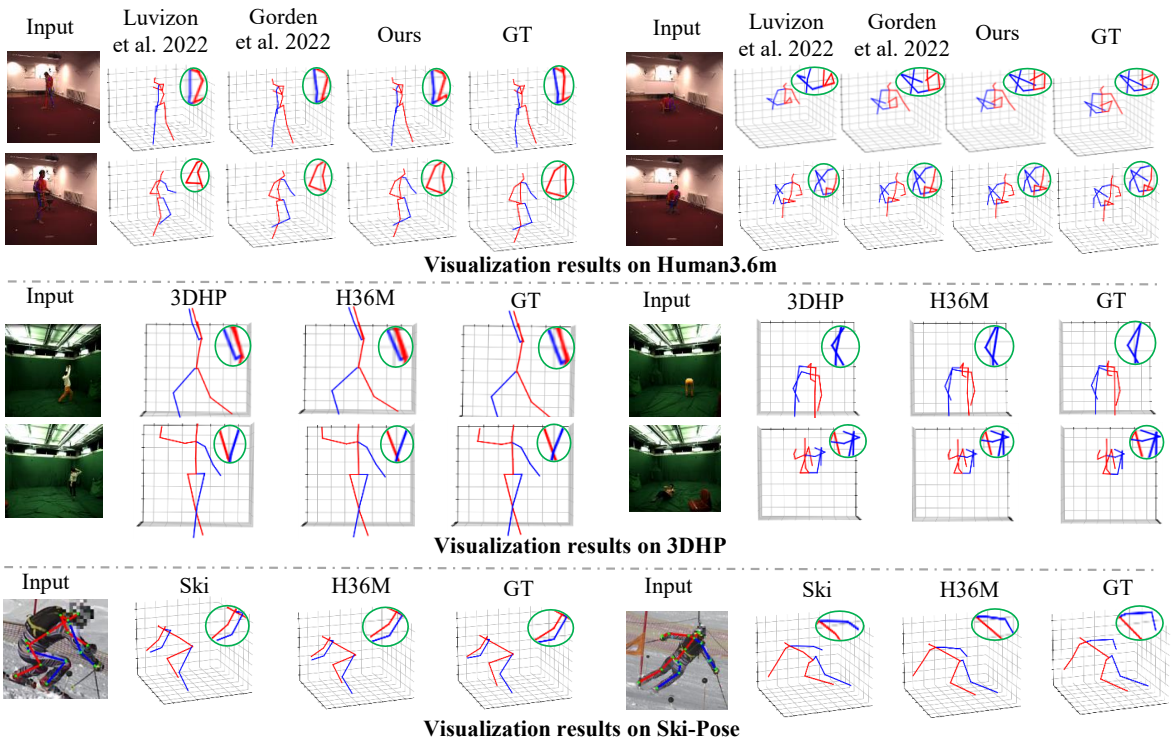

Figure 5: The qualitative results on three datasets.

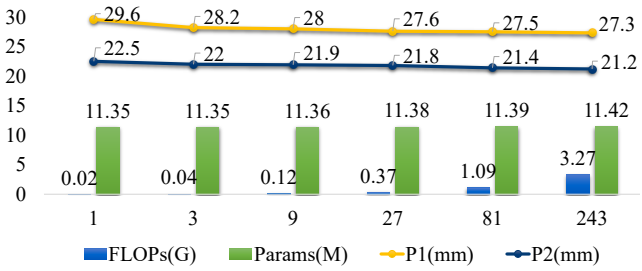

Figure 6: Impacts of temporal receptive fields (TRFs).

data to train a better model. When TRF=1, our model performs well, showing that our method can extract sufficiently rich pose semantic information and somewhat reduce depth ambiguity even in the absence of significant temporal information. As TRF grows, the model parameters almost remain constant, proving that our method is not sensitive to TRF and has no burden to process multi-frame data. FLOPs rise slowly with TRF but remain small even at TRF=243.

Computational complexity. Table 8 compares the computational complexity of our model with relevant methods. Even though our TRF is larger than (Luvizon, Picard, and Tabia 2022), our model has fewer parameters and yields better performance. When compared with (Gordon et al. 2022) using the same TRF, our model performs better with fewer parameters and FLOPs. These indicate that our approach strikes a balance between performance and efficiency, demonstrating strong practicality.

Methods	TRF	Params(M)	FLOPs(G)	P1(mm)
(Luvizon et al. 2022)	1	23.3	-	45.0
(Gordon et al. 2022)	27	70.4	8.5	30.2
Ours	27	**11.4**	**0.4**	**27.6**

Table 8: Computation complexity comparison. Best in bold.

Conclusion

In this paper, we developed a deep semantic graph transformer-based multi-view 3D HPE structure, which improved 3D pose prediction performance by adaptively learning and fusing various significant pose semantic features. First, we developed a deep semantic graph transformer encoder, which dynamically mined the position, spatial structure, and skeletal edge feature embeddings and their correlations of human joints, greatly enhancing the spatial feature representation. Then, we constructed a progressive multi-view spatial-temporal feature fusion framework, successfully merging the spatial-temporal distinguishing and consistent features across multiple viewpoints using various feature fusion modules. Extensive experiments on three 3D HPE benchmarks demonstrated how effective our approach is. It effectively increases the expressiveness of the pose feature, and its spatial-temporal feature fusion strategy is fairly beneficial in reducing depth ambiguity in single-view 3D HPE and significantly enhancing 3D HPE performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62106247 and Grant 62371438.

References

Bartol, K.; Bojanić, D.; Petković, T.; and Pribanić, T. 2022. Generalizable Human Pose Triangulation. In *CVPR*, 11028–11037.

Bouazizi, A.; Wiederer, J.; Kressel, U.; and Belagiannis, V. 2021. Self-Supervised 3D Human Pose Estimation with Multiple-View Geometry. In *FG*.

Bultmann, S.; and Behnke, S. 2021. Real-Time Multi-View 3D Human Pose Estimation using Semantic Feedback to Smart Edge Sensors. In *RSS*.

Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.; Yuan, J.; and Thalmann, N. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, 2272–2281.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 7103–7112.

Ci, H.; Wu, M.; Zhu, W.; Ma, X.; Dong, H.; Zhong, F.; and Wang, Y. 2023. GFPose: Learning 3D Human Pose Prior With Gradient Fields. In *CVPR*, 4800–4810.

Fasel, B.; Spörri, J.; Gilgien, M.; Boffi, G.; Chardonnens, J.; Müller, E.; and Aminian, K. 2016. Three-dimensional body and centre of mass kinematics in alpine ski racing using differential gnss and inertial sensors. *Remote Sensing*, 8(8): 617.

Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; and Hu, H. 2023. Human Pose as Compositional Tokens. In *CVPR*.

Gholami, M.; Rezaei, A.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022. Self-supervised 3D human pose estimation from video. *Neurocomputing*, 488: 97–106.

Gong, J.; Foo, L. G.; Fan, Z.; Ke, Q.; Rahmani, H.; and Liu, J. 2023. DiffPose: Toward More Reliable 3D Pose Estimation. In *CVPR*, 13041–13051.

Gordon, B.; Raab, S.; Azov, G.; Giryes, R.; and Cohen-Or, D. 2022. Flex: Parameter-free multi-view 3d human motion reconstruction. In *ECCV*.

He, Y.; Yan, R.; Fragkiadaki, K.; and Yu, S.-I. 2020. Epipolar transformers. In *CVPR*, 7779–7788.

Huang, F.; Zeng, A.; Liu, M.; and Lai, Q. 2020. DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image. In *WACV*, 429–438.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3.6M: Large scale datasets and predic- tive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1325–1339.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2023. Pose-Oriented Transformer with Uncertainty-Guided Refinement for 2D-to-3D Human Pose Estimation. *AAAI*, 37(1): 1296–1304.

Iqbal, U.; Molchanov, P.; and Kautz, J. 2020. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 5242–5251.

Iskakov, K.; Burkov, E.; Lempitsky, V.; and Malkov, Y. 2019. Learnable triangulation of human pose. In *ICCV*, 7717–7726.

Kim, H.-W.; Lee, G.-H.; Oh, M.-S.; and Lee, S.-W. 2022. Cross-View Self-Fusion for Self-Supervised 3D Human Pose Estimation in the Wild. In *ACCV*, 1385–1402.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 1077–1086.

Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; and Yang, W. 2022a. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25: 1282–1293.

Li, W.; Liu, H.; Tang, H.; and Wang, P. 2023. Multi-Hypothesis Representation Learning for Transformer-Based 3D Human. *Pattern Recognition*, 141.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Gool, L. V. 2022b. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, 13147–13156.

Liu, J.; Rojas, J.; Li, Y.; Liang, Z.; Guan, Y.; Xi, N.; and Zhu, H. 2021. A Graph Attention Spatio-temporal Convolutional Networks for 3D Human Pose Estimation in Video. In *ICRA*, 3374–3380.

Liu, K.; Ding, R.; Zou, Z.; Wang, L.; and Tang, W. 2020a. A comprehensive study of weight sharing in graph networks for 3D human pose estimation. In *ECCV*, 318–334.

Liu, K.; Zou, Z.; and Tang, W. 2020. Learning Global Pose Features in Graph Convolutional Networks for 3D Human Pose Estimation. In *ACCV*, 89–105.

Liu, R.; Shen, J.; Wang, H.; Chen, C.; Cheung, S.; and Asari, V. 2020b. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, 5064–5073.

Luvizon, D.; Picard, D.; and Tabia, H. 2018. 2D/3D pose estimation and action recognition using multitask deep learning. In *CVPR*, 5137–5146.

Luvizon, D.; Tabia, H.; and Picard, D. 2019. Human pose regression by combining indirect part detection and contextual information. *Computers and Graphics*, 85: 15–22.

Luvizon, D. C.; Picard, D.; and Tabia, H. 2022. Consensus-Based Optimization for 3D Human Pose Estimation in Camera Coordinates. *International Journal of Computer Vision*, 130: 869–882.

Ma, H.; Chen, L.; Kong, D.; Wang, Z.; Liu, X.; Tang, H.; Yan, X.; Xie, Y.; Lin, S.-Y.; and Xie, X. 2021. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. In *BMVC*.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*.

Pavlakos, G.; Zhou, X.; Derpanis, K.; and Daniilidis, K. 2017. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 1263–1272.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 7745–7754.

Qiu, H.; Wang, C.; Wang, J.; Wang, N.; and Zeng, W. 2019. Cross View Fusion for 3D Human Pose Estimation. In *ICCV*, 4342–4351.

Remelli, E.; Han, S.; Honari, S.; Fua, P.; and Wang, R. 2020. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *CVPR*.

Rhodin, H.; Meyer, F.; Spörri, J.; Müller, E.; Constantin, V.; Fua, P.; Katircioglu, I.; and Salzmann, M. 2018. Learning Monocular 3D Human Pose Estimation From Multi-View Images. In *CVPR*, 8437–8446.

Shan, W.; Liu, Z.; Zhang, X.; Wang, Z.; Han, K.; Wang, S.; Ma, S.; and Gao, W. 2023. Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation. arXiv:2303.11579.

Shuai, H.; Wu, L.; and Liu, Q. 2022. Adaptive Multi-view and Temporal Fusing Transformer for 3D Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8): 1–14.

Tekin, B.; Marquez-Neila, P.; Salzmann, M.; and Fua, P. 2017. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *ICCV*, 3961–3970.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *NeurIPS*.

Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; and Rosenhahn, B. 2021. CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild. In *CVPR*, 13294–13304.

Wang, C.; Qiu, W.; Qin, W.; and Zeng, W. 2021. AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild. *International Journal of Computer Vision*, 129: 703–718.

Wang, J.; Yan, S.; Xiong, Y.; and Lin, D. 2020. Motion Guided 3D Pose Estimation from Videos. In *ECCV*, 764–780.

Xiang, D.; Joo, H.; and Sheikh, Y. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 10957–10966.

Xie, R.; Wang, C.; and Wang, Y. 2022. Metafuse: A pre-trained fusion model for human pose estimation. In *CVPR*.

Xu, T.; and Takano, W. 2021. Graph Stacked Hourglass Networks for 3D Human Pose Estimation. In *CVPR*, 16105–16114.

Yeh, R.; Hu, Y.; and Schwing, A. 2019. Chirality nets for human pose regression. *NeurIPS*, 32: 8163–8173.

Zeng, A.; Sun, X.; Yang, L.; Zhao, N.; Liu, M.; and Xu, Q. 2021. Learning skeletal graph neural networks for hard 3d pose estimation. In *ICCV*, 11436–11445.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022a. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In *CVPR*, 13232–13242.

Zhang, L.; Lu, F.; Zhou, K.; Zhou, X.-D.; and Shi, Y. 2023a. Hierarchical Spatial-temporal Adaptive Graph Fusion for Monocular 3D Human Pose Estimation. *IEEE Signal Processing Letters*, 1–5.

Zhang, L.; Shao, X.; Li, Z.; Zhou, X.-D.; and Shi, Y. 2022b. Spatio-temporal Attention Graph for Monocular 3D Human Pose Estimation. In *ICIP*, 1231–1235.

Zhang, L.; Zhou, K.; Liu, L.; Li, Z.; Zhao, X.; Zhou, X.-D.; and Shi, Y. 2023b. Progressive Multi-view Fusion for 3D Human Pose Estimation. In *ICIP*, 1600–1604.

Zhang, Y.; An, L.; Yu, T.; Li, X.; Li, K.; and Liu, Y. 2020. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, 1321–1330.

Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; and Metaxas, D. 2019. Semantic graph convolutional networks for 3D human pose regression. In *ICCV*, 3425–3435.

Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; and Chen, C. 2023. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In *CVPR*.

Zhao, W.; Wang, W.; and Tian, Y. 2022. GraFormer: Graph-Oriented Transformer for 3D Pose Estimation. In *CVPR*, 20438–20447.

Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; and Ding, Z. 2021. 3d human pose estimation with spatial and temporal transformers. In *CVPR*, 11656–11665.

Zhou, K.; Han, X.; Jiang, N.; Jia, K.; and Lu, J. 2019. HEMlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In *ICCV*, 2344–2353.

Zhou, K.; Zhang, L.; Lu, F.; Zhou, X.-D.; and Shi, Y. 2023. Efficient Hierarchical Multi-view Fusion Transformer for 3D Human Pose Estimation. In *ACMMM*, 7512–7520.

Zou, L.; Huang, Z.; Gu, N.; Wang, F.; Yang, Z.; and Wang, G. 2021. GMDN: A lightweight graph-based mixture density network for 3D human pose regression. *Computers and Graphics*, 95: 115–122.

Zou, Z.; Liu, K.; Wang, L.; and Tang, W. 2020. High-order graph convolutional networks for 3D human pose estimation. In *BMVC*.