HDFormer: High-order Directed Transformer for 3D Human Pose Estimation

Hanyuan Chen[1] , **Jun-Yan He**[1] , **Wangmeng Xiang**[1] , **Zhi-Qi Cheng**[2] ,
Wei Liu[1] , **Hanbing Liu**[3] , **Bin Luo**[1] , **Yifeng Geng**[1] **and Xuansong Xie**[1]
[1]DAMO Academy, Alibaba Group
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]Tsinghua University, Beijing, China
hanyuan.chy, leyuan.hjy, wangmeng.xwm, luwu.lb, cangyu.gyf@alibaba-inc.com
zhiqic@cs.cmu.edu, ustclwwx@gmail.com, liuhb21@mails.tsinghua.edu.cn, xingtong.xxs@taobao.com

Abstract

Human pose estimation is a challenging task due to its structured data sequence nature. Existing methods primarily focus on pair-wise interaction of body joints, which is insufficient for scenarios involving overlapping joints and rapidly changing poses. To overcome these issues, we introduce a novel approach, the ***H**igh-order **D**irected Transformer* (HDFormer), which leverages high-order bone and joint relationships for improved pose estimation. Specifically, HDFormer incorporates both self-attention and high-order attention to formulate a multi-order attention module. This module facilitates first-order "joint↔joint", second-order "bone↔joint", and high-order "hyperbone↔joint" interactions, effectively addressing issues in complex and occlusion-heavy situations. In addition, modern CNN techniques are integrated into the transformer-based architecture, balancing the trade-off between performance and efficiency. HDFormer significantly outperforms state-of-the-art (SOTA) models on Human3.6M and MPI-INF-3DHP datasets, requiring only 1/10 of the parameters and significantly lower computational costs. Moreover, HDFormer demonstrates broad real-world applicability, enabling real-time, accurate 3D pose estimation. The source code is at https://github.com/hyer/HDFormer.

1 Introduction

Despite significant strides in deep learning-based 3D pose estimation [Iskakov *et al.*, 2019; Qiu *et al.*, 2019; Pavllo *et al.*, 2018; Li *et al.*, 2020; Zhu *et al.*, 2021; Gong *et al.*, 2021; Ye *et al.*, 2022], achieving stable, accurate pose sequences remains elusive. The prevalent 3D pose estimation framework takes 2D pose detection results [Chen *et al.*, 2017; Sun *et al.*, 2019] as inputs and estimates depth information via end-to-end Graph Convolutional Networks (GCNs)[Cai *et al.*, 2019; Pavllo *et al.*, 2018] or Transformers[Zhang *et al.*, 2022]. However, complex scenarios involving overlapping keypoints, rapid pose changes, and varying scales pose challenges to the depth estimation of 3D keypoints.

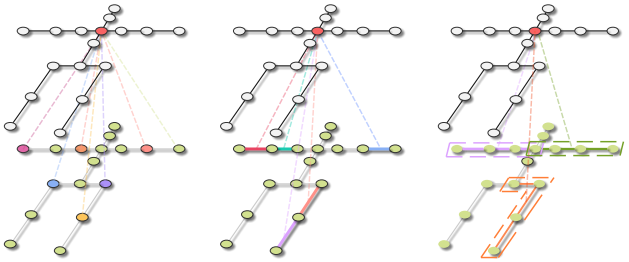

Figure 1: Illustration of first-order (joint↔ joint) attention, second-order (bone↔ joint) and high-order (hyperbone↔ joint) attention. First-order attention models connections between joints, while second-order attention focuses on the relationship between joints and bones. High-order attention delves into intricate relationships between joints and hyperbones.

Face with these challenges, existing methods mainly utilize first-order "joint ↔ joint" and second-order "bone ↔ joint" connections, often overlooking high-order interactions among joint sets (referred to as hyperbones). Different from first-order ("joint ↔ joint") and second-order ("bone ↔ joint") that focus on pair-wise joints/bones connections, high-order relations could describe complex motion dynamics. The high-order interactions contain rich semantic information in motions, as skeletons often move in specific patterns and involve multiple joints and bones simultaneously.

To Learn high-order information without expensive costs, we propose a novel framework named **H**igh-order **D**irected Transformer (HDFormer), which coherently exploits the multi-order information aggregation of skeleton structure for 3D pose estimation. Specifically, HDFormer leverages the first-order attention to learn the spatial semantics among "joint ↔ joint" relationships. Additionally, it integrates a robust high-order attention module to enhance 3D pose estimation accuracy by capturing both second-order and high-order information. To encode the hyperbone features, the hyperbone representation encoding module is employed under the constraints of a pre-defined directed human skeleton graph. HDFormer strikes a commendable balance between efficacy and efficiency. In summary, the key contributions of this paper are summarized as follows:

- We investigate high-order attention module to learn both the *"bone↔joint"* and *"hyperbone↔joint"* with an effective and efficient cross-attention mechanism. To the best of our knowledge, it is the first end-to-end model to utilize high-order information on a directed skeleton graph for 3D pose estimation.
- We propose a novel High-order Directed Transformer (HDFormer) for 3D pose estimation. It utilizes *"joint↔joint"*, *"bone↔joint"* and *"hyperbone↔joint"* information with a three-stage U-shape architecture design, which endows the network with the ability to handle more complex scenarios.
- HDFormer is evaluated on popular 3D pose estimation benchmarks Human3.6M and MPI-INF-3DHP with analysis of quantitative and qualitative results. Specifically, it achieves 21.6% (96 frames) on Human3.6M without using any extra data, which outperforms the existing SOTA work MixSTE [Zhang *et al.*, 2022] with only 1/10 parameters and a fraction of computational cost.

2 Related Work

2.1 3D Human Pose Estimation

Despite progress in 2D human pose estimation, 3D versions face challenges due to depth ambiguity. While some methods leverage multi-view images/videos [Iskakov *et al.*, 2019; Qiu *et al.*, 2019; Ye *et al.*, 2022; He *et al.*, 2021; Zhao *et al.*, 2018; Huang *et al.*, 2021; Qiao *et al.*, 2022; He *et al.*, 2023], these setups are costly and complex. Hence, estimating 3D poses with monocular images/videos is more practical, with recent studies showcasing the efficacy of converting 2D joint locations to 3D [Pavllo *et al.*, 2018; Zhu *et al.*, 2021]. Strategies such as data augmentation have been used to generate diverse 2D-3D pose pairs, improving generalization [Gong *et al.*, 2021; Li *et al.*, 2020].

2.2 Graph ConvNet Based Methods

GCNs [Scarselli *et al.*, 2009; Gilmer *et al.*, 2017], extending conventional convolution operators to graphs, have been utilized in several human-related tasks like action recognition [Shi *et al.*, 2019; Shi *et al.*, 2018; Xiang *et al.*, 2022], action synthesis [Yan *et al.*, 2019], and 3D human pose estimation [Zhou *et al.*, 2022; Zhou *et al.*, 2023]. [Wang *et al.*, 2020] created a lightweight, efficient U-shaped model to capture temporal dependencies. Inspired by [Shi *et al.*, 2019], [Hu *et al.*, 2021] proposed a conditional directed graph convolution for adaptive graph topology, enhancing non-local dependence. Other methods, such as Semantic Graph Convolution (SemGConv) [Zhao *et al.*, 2019] and Modulated-GCN [Zou and Tang, 2021], also prioritize spatial joint relationships. However, they often overlook edge information, particularly high-order relationships. Unlike Skeletal-GNN [Zeng *et al.*, 2021] which utilizes GCN to capture action-specific poses, we establish joint and hyperbone relationships via cross-attention in the directed graph skeleton.

2.3 Transformer Based Methods

Transformers, first introduced by [Vaswani *et al.*, 2017], have been widely used in visual tasks [Zhou *et al.*, 2022; Tu *et al.*, 2023; Cheng *et al.*, 2022]. In the domain of 3D pose estimation, [Li *et al.*, 2021] proposed the Stride Transformer to lift a long 2D pose sequence to a single 3D pose. PoseFormer [Zheng *et al.*, 2021] created a model comprising Spatial-Temporal Transformer blocks to capture spatial and global dependencies across frames. Similarly, MixSTE [Zhang *et al.*, 2022] captured the temporal motion of body joints over long sequences. While CrossFormer [Wang *et al.*, 2021] and GraFormer [Zhao *et al.*, 2022] encoded dependencies between body joints, our method applies a cross-attention mechanism to integrate high-order relationships and explore the hyperbone-joint relationship. Although transformer-based methods effectively capture global spatial-temporal joint relationships, their computational costs significantly exceed those of GCN-based methods.

3 Proposed Method

3.1 Preliminaries

Directed Skeleton Graph

The directed skeleton graph $\mathcal{G}$ shown in the right part of Fig. 2 represents the human skeleton structure, where the nodes are human skeleton joints and the arrows are human skeleton bones. Generally, the human skeleton can be represented as a graph $\mathcal{G} = (\mathcal{V}; \mathcal{E})$, where the vertices are human skeleton joints and edges are physical connections between two joints. Here $\mathcal{V}$ is the set of N joints and $\mathcal{E}$ is characterized by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. The raw pose data, i.e., the joint keypoints vector, is a set of 2D coordinates. In this way, the pose data is transformed into a graph sequence and specifically represented as a tensor $\mathbf{X} \in \mathbb{R}^{T\times N\times C}$, where T, N, and C denote the temporal length, numbers of joints and channels, respectively. Directed graph is adopted, as it allows a convenient hyperbone definition.

Transformer

Our model's attention mechanism is built upon the original implementation of the classic Transformer [Vaswani *et al.*, 2017]. The attention computing with *query*, *key* and *value* matrix Q, K, V in each head are depicted as:

$$\text{Attn}(Q, K, V) = \text{Softmax}((QK^T + \mathbf{A} + \Psi)/\sqrt{d_m})V, \quad (1)$$

where $Q, K, V \in \mathbb{R}^{N\times d_m}$, N is the number of tokens, and d_m indicates the dimension of each token. The multi-head attention of S heads is defined as follows:

$$\hbar_i = \text{Attn}(Q_i, K_i, V_i), i \in \{1, \ldots, S\}, \quad (2)$$

$$\text{MSA} = \text{Concat}(\hbar_1, ..., \hbar_S)W_o, \quad (3)$$

where $\hbar$ is the attention calculation result for a single head, $W_o \in \mathbb{R}^{d_m\times d_m}$ is the linear projection weight. $\mathbf{A}$ is the adjacency matrix and Ψ is a learnable adjacency matrix. The matrix $\mathbf{A}$ is fixed and represents the predetermined connections between joints, while the learnable adjacency matrix Ψ adjusts the connection weights based on the input data, improving the capturing of spatial relationships between different joints. The ablation study on the impact of Ψ is presented in line 1 of Tab. 7.

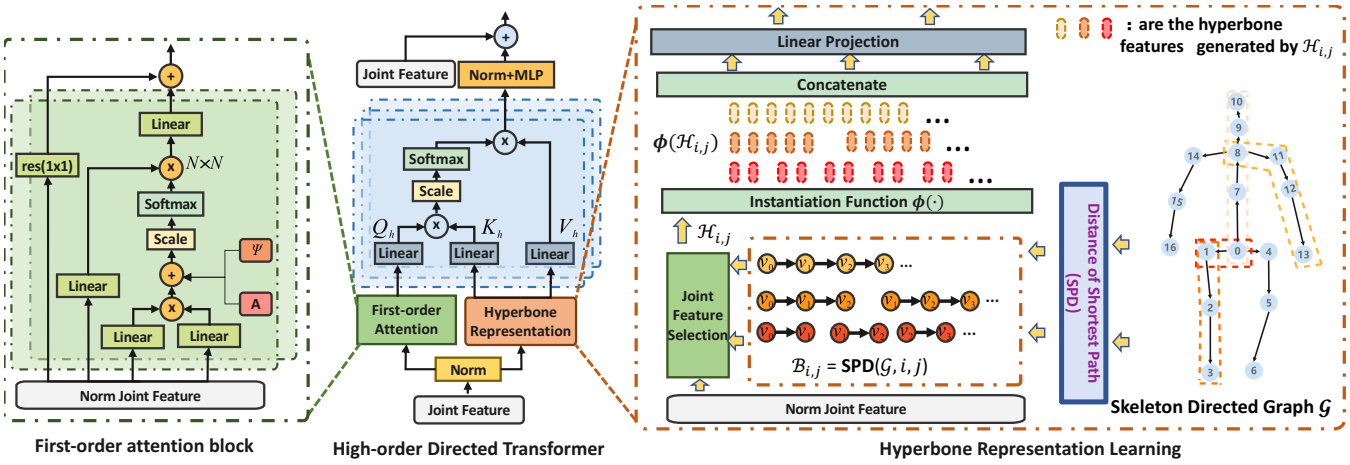

Figure 2: The illustration of High-order Direction Transformer (HDFormer) block. HDFormer block consists of three major parts: (a) First-order attention block to capture "*joint*↔ *joint*" spatial relationship; (b) Hyperbone representation learning module to encode hyperbone features; (c) High-order attention block to capture both second-order "*bone* ↔ *joint*" and high-order "*hyperbone*↔ *joint*" interactions.

3.2 High-order Directed Transformer

The spatial connections between "joint↔joint" and "joint↔bone" are referred to as first-order and second-order information in the 3D pose estimation, which is widely studied in the previous works [Zhang *et al.*, 2022]. Nevertheless, pairwise first-order and second-order information alone cannot fully describe the complex human skeleton dynamics in the 2D to 3D mappings. For example, human skeletons often move in specific patterns and involve multiple joints and bones at the same time. This observation leads us to further investigate the high-order information interaction of the human skeleton by integrating the high-order attention learning with directed graph and propose a High-order Direct Transformer (HDFormer) for 3D pose estimation.

First-order Attention Modeling

The joint sets of the skeleton describe the rough posture of the human body, and the global multi-head attention [Zhang *et al.*, 2022] has demonstrated its effectiveness in 3D pose estimation. Therefore, the multi-head attention scheme is adopted in this work for first-order attention modeling. As illustrated in Fig. 2(a), $A \in \mathbb{R}^{N\times N}$ is the adjacency matrix and Ψ is the learnable adjacency matrix which has the same dimension as A. Specifically, given the joint token set $\mathcal{Z} = \{z_1 \dots z_i\}$ where $i \in N$ denotes the index of skeleton joints, $z_i \in \mathbb{R}^C$, C represents the feature channel. The first-order self-attention modeling and feature of joints can be obtained by following Eq. 1, where *query*, *key*, and *value* matrices are generated by three linear layers, $Q = W_q\mathcal{Z}$, $K = W_k\mathcal{Z}$, and $V = W_v\mathcal{Z}$, respectively. S represents the number of heads and $\hbar_k$ is the output of each head. Unlike the traditional multi-head attention fuses the output of the attention module with concatenating (Eq. 3), we revamp the fusion scheme with a simple accumulation:

$$\hat{\mathcal{Z}} = \sum_{k=1}^{S} \hbar_k, \tag{4}$$

where $\hat{\mathcal{Z}}$ denote the final output of first-order attention.

Hyperbone Representation

A hyperbone is a series of joints and bones that are connected sequentially. The human skeleton can be represented as a special type of graph without loops, allowing for the unique determination of the shortest path between two joints. Given a starting and ending joint, the corresponding hyperbone can be identified using the *distance of the shortest path* (SPD). Specifically, as shown in Fig. 2, the human skeleton can be described as a directed graph $\mathcal{G}$. The "hip" joint is defined as the directed graph's root node. Given two joint nodes on the directed graph, we could follow the direction of edges to find the shortest path from starting joint i to j. For example, there is a shortest path from joint index 0 to joint index 3 by [0, 1, 2, 3], which is done by moving from index 0 to index 3 following the edges of the directed graph (bone for human skeleton). Formally, given the human skeleton-directed graph $\mathcal{G}$ and the $(start, end)$ joint indices (i, j), we could utilize the SPD algorithm to discover the joint set belonging to the hyperbone $\mathcal{B}_{i,j} = \{v_{h_i}, v_{h_{i+1}}, \dots, v_{h_j}\}$, where v_* represents the human joint, $|\mathcal{B}_{i,j}| = n$ represents the number of joints in hyperbone, and we call this the *order* of hyperbone, h_* is the joint index:

$$\mathcal{B}_{i,j} = \text{SPD}(\mathcal{G}, i, j), \tag{5}$$

To encode hyperbone features, we propose a novel hyperbone encoding method. Specifically, the feature of hyperbone can be obtained by a function $\phi(\cdot)$ that takes hyperbone joint set $\mathcal{B}_{i,j}$'s corresponding features $\mathcal{H}_{i,j} = \{z_{h_i}, z_{h_{i+1}}, \dots, z_{h_j}\}$ as input and generate hyperbone features, where z_{h_i} is the feature of joint v_{h_i}.

Instantiation

Previous works, e.g. Anatomy [Chen *et al.*, 2021], used a simple subtraction of joint features to construct bone features. In contrast, we propose a general process for constructing both bone and hyperbone features, and offer instantiation methods. Specifically, we investigate several instantiations of the function $\phi(\cdot)$.

Subtraction. $\phi(\cdot)$ can be defined as a subtraction operation. As we use a directed graph to represent the human skeleton when adopting subtraction for hyperbone representation, it is equivalent to the subtraction of start and end joints:

$$\phi(\mathcal{H}_{i,j}) = f(z_{h_i} - z_{h_j}), \tag{6}$$

where z is the joint feature, f is a linear mapping. This representation is easy to calculate and works fine for second-order bone representation, however, it loses information on bone sequence for hyperbone with higher order.

Summation/Multiplication. $\phi(\cdot)$ can also be defined as element-wise summation or multiplication for joints:

$$\phi(\mathcal{H}_{i,j}) = \sum_{z \in \mathcal{H}_{i,j}} f(z)/n, \tag{7}$$

$$\phi(\mathcal{H}_{i,j}) = \prod_{z \in \mathcal{H}_{i,j}} f(z), \tag{8}$$

where n is the number of joints, f is a linear mapping.

Concatenation. $\phi(\cdot)$ can be defined with concatenation and linear mapping:

$$\phi(\mathcal{H}_{i,j}) = f([z_{h_1}, \ldots, z_{h_n}]), \tag{9}$$

where the operator $[\cdot]$ represents the concatenation of features in the shortest path, f maps the concatenated feature to the same dimension as the joint feature.

Sub-Concat. To overcome the sequence information loss issue in subtraction, we combine subtraction and concatenation for a mixed function for $\phi(\cdot)$:

$$\phi(\mathcal{H}_{i,j}) = f([z_{h_1} - z_{h_2}, \ldots, z_{h_{n-1}} - z_{h_n}]), \tag{10}$$

where the second-order bone feature is calculated with subtraction and the high-order hyperbone feature is obtained with concatenation and linear mapping.

High-order Directed Transformer

Fig. 2(b) illustrates the architecture of our proposed High-order Directed Transformer block, which consists of three components: first-order attention block, hyperbone encoding block, and high-order attention block. The cross-attention fusion involves joint features $\hat{\mathcal{Z}}$ from first-order attention modeling block and hyperbone feature $H = [Y_2, ..Y_o, ..Y_n]$, where Y_o represent hyperbone features with order o from hyperbone encoding block. Formally, the cross-attention fusion can be expressed as:

$$\begin{aligned} Y_o &= [\phi(\mathcal{H}_{i,j})], |\mathcal{H}_{i,j}| = o, \\ H &= [Y_2, \ldots, Y_n], \\ Q_h = W_{q_h}\hat{\mathcal{Z}}, K_h &= W_{k_h}H, V_h = W_{v_h}H, \\ \text{CrossAttn}(Q_h, K_h, V_h) &= \text{Softmax}(Q_h K_h^T/\sqrt{d_m})V_h, \end{aligned} \tag{11}$$

where $W_{q_h}, W_{k_h}, W_{v_h}$ are learnable parameters. Since we only use the joint feature in the query, the computation and memory complexity of generating the cross-attention map in cross-attention are linear rather than quadratic as in all-attention, making the entire process more efficient. Similar to MHSA [Vaswani *et al.*, 2017], we also adopt a multi-head attention design and add an MLP layer after the attention layer.

Loss Function

We adopted the loss function similar to UGCN [Wang *et al.*, 2020], which was formulated as follow:

$$\mathcal{L} = \mathcal{L}_p + \lambda \mathcal{L}_m, \tag{12}$$

where $\mathcal{L}_p$ is the 3D joint coordinates loss, which is defined as the mean per joint position error (MPJPE), and $\mathcal{L}_m$ is motion loss introduced by [Wang *et al.*, 2020], λ is a hyperparameter for balancing two objectives and is set to 0.1. Motion loss allows our model to capture more natural movement patterns of the keypoints in the prediction, since Minkowski Distance loss does not consider the similarity of temporal structure between the estimated pose sequence and ground truth.

3.3 Network Architecture

As illustrated in Fig. 3, the proposed network architecture contains three stages: 1) Downsampling stage collects long-time range information by temporal pooling. The temporal downsampling block has an inside temporal convolution's stride set to 2 and kernel size set to 5. It is used to downsample the temporal resolution; 2) Upsampling stage recovers the temporal resolution, and skip connections are adopted between the downsampling stage and the upsampling stage to integrate the low-level features. The temporal upsampling block is the conventional bilinear interpolation along the temporal axis to recover higher temporal resolution. 3) Merging stage transforms the feature maps at different temporal scales in the upsampling stage, and fuses them to obtain the final embedding. Finally, the 3D coordinate for each keypoint is regressed by a fully connected layer.

4 Experiment

4.1 Datasets and Metric

Experiments are conducted on the 3D pose estimation benchmark dataset Human3.6M[Ionescu *et al.*, 2014] and MPI-INF-3DHP [Mehta *et al.*, 2016a]. Human3.6M is the most widely used evaluation benchmark, containing 3.6 million video frames captured from four synchronized cameras with different locations and poses at 50 Hz. 11 subjects are performing 15 kinds of actions. MPI-INF-3DHP is a 3D human body pose estimation dataset consisting of both constrained indoor and complex outdoor scenes. It consists of 1.3M frames captured from the 14 cameras. For fair comparisons, the evaluation metric MPJPE is adopted in this work, which follows the setting of the previous works [Hu *et al.*, 2021; Cai *et al.*, 2019; Zhang *et al.*, 2022; Zhao *et al.*, 2022].

4.2 Implementation Details

We optimized the model by the AdaMod optimizer [Ding *et al.*, 2019] for 110 epochs with a batch size of 256, and the base learning rate is 5×10^{-3} with decayed by 0.1 at 80, 90, and 100 epochs. To avoid over-fitting, we set the weight decay factor to 10^{-5} for parameters of convolution layers and the dropout rate to 0.3 at part of the layers. Besides, we followed UGCN [Wang *et al.*, 2020] to apply the sliding window algorithm with a step length of 5 to estimate a variable-length pose sequence with fixed input length at inference time.

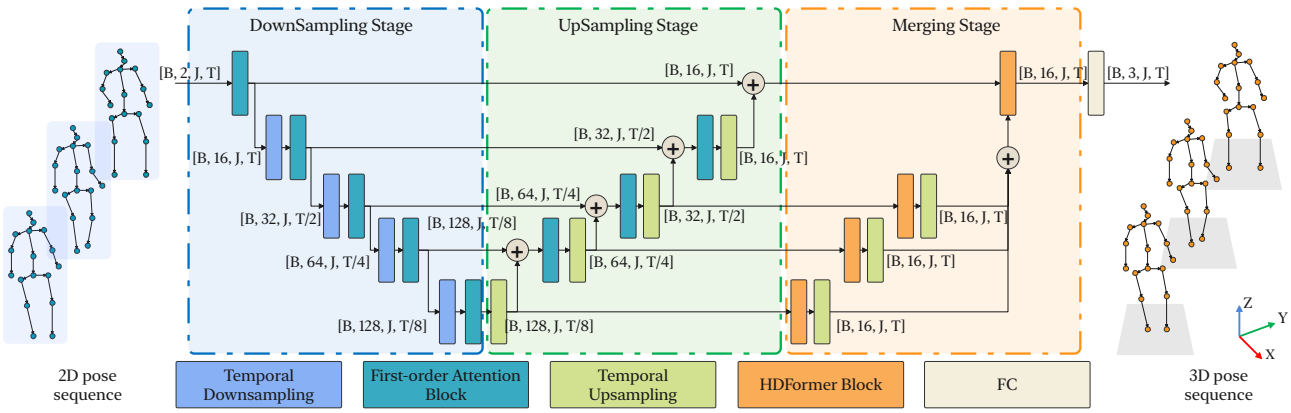

Figure 3: Overview of our framework: A High-order Directed Transformer with a U-shaped design for 3D human pose estimation. The framework includes downsampling, upsampling, and merging stages, incorporating high-order attention and multi-scale temporal information. The batch size, the number of nodes, and the sequence length are represented by symbols B, J, and T, respectively.

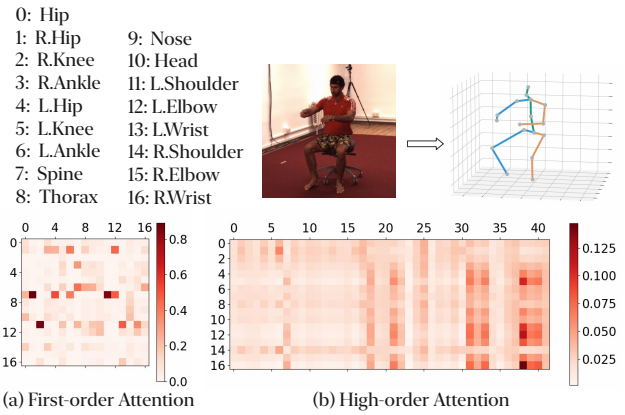

Figure 4: Visualization of first-order attentions and high-order attention between body joints and hyperbone.

4.3 Quantitative Evaluation

Results on Human3.6M. The proposed approach is compared with the state-of-the-art methods to evaluate the performance. In this subsection, the reported performance in their original paper is directly copied as their results. The performance comparison with the state-of-the-art works on Human3.6M [Ionescu *et al.*, 2014] is listed in Tab. 1, including graph ConvNet-based and Transformer-based methods. For fair comparisons with other SOTA methods, we consider not only the effectiveness of the model but also the scale of parameters and latency in Tab. 2, which can comprehensively demonstrate the real-world performance of the model. To our knowledge, this is the first comprehensive comparison experiment on the benchmark dataset Human3.6M. The current SOTA method MixSTE [Zhang *et al.*, 2022], a transformer-based model, achieves 25.9% and 21.6% MPJPE with the input of 81 and 243-frame sequences, respectively. Compared to MixSTE, the proposed HDFormer achieves 21.6% MPJPE with the input of only 96 frames. More importantly, our model has only a 1/10 scale of 3.7 M vs. 33.8 M and six times the speed. The graph ConvNet-based SOTA method U-CondDGCN has a very small scale of parameters, latency, and ideal performance. However, the proposed HDFormer is a transformer-based method that achieves significant improvement compared to U-CondDGCN with a very close scale of parameters and same-level latency.

Results on MPI-INF-3DHP. In Tab. 3, we compared our method with state-of-the-art methods on the MPI-INF-3DHP benchmark to evaluate the generalization ability of the proposed HDFormer. We take the ground-truth 2D poses as model input. Our method achieves the same trends as the results on Human3.6M, which is also the SOTA performance under the metric of PCK, AUC, and MPJPE.

4.4 Qualitative Results

As shown in Fig. 4, we further conduct visualization on the First-order attention and High-order attention. The selected action (Eating of test set S9) is applied for visualization. For the First-order attention map in Fig. 4(a), the horizontal and vertical axes are all joint indexes, and it can be easily observed that the dependency between the **spine** node and left/right **elbow** nodes are significant for the "eating" sequence. Besides, the left **shoulder** node also plays an important role in the spatial relationship with the left **ankle** node when eating in the sitting pose. Furthermore, to demonstrate the effect of the proposed high-order attention block, we further visualize the high-order attention map for the action of eating from the test set S9 in Fig. 4(b), where the vertical axes were the index of the joint while the horizontal axes were the index of hyperbones. In our experiments, the maximum SPD length was 4. As a result, the hyperbone sequence has 42 bone features in the horizontal axes. From the attention map, we can find that the hyperbone feature has an impact on different joints. The left/right **elbows** and left/right **wrist** have a relatively large response to hyperbone sequence index from 38 to 41, which corresponds to the higher order bone feature. The hyperbone 38 to 41 corresponding to joint

Protocol #1	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Cai [Cai *et al.*, 2019] (CPN, T=7)	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Pavllo [Pavllo *et al.*, 2018] (CPN, T=243)	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Xu [Xu *et al.*, 2020] (CPN, T=9)	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6
Liu [Liu *et al.*, 2020](CPN, T=243)	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
Wang [Wang *et al.*, 2020] (CPN, T=96)	41.3	43.9	44.0	42.2	48.0	57.1	42.2	43.2	57.3	61.3	47.0	43.5	47.0	32.6	31.8	45.6
Hu [Hu *et al.*, 2021] (CPN, T=96)	38.0	43.3	39.1	39.4	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Zhang [Zhang *et al.*, 2022] (CPN, T=81)	39.8	43.0	38.6	40.1	43.4	50.6	40.6	41.4	52.2	56.7	43.8	40.8	43.9	29.4	30.3	42.4
Wang [Wang *et al.*, 2020] (HR-Net, T=96)	38.2	41.0	45.9	39.7	41.4	51.4	41.6	41.4	52.0	57.4	41.8	44.4	41.6	33.1	30.0	42.6
Hu [Hu *et al.*, 2021] (HR-Net, T=96)	35.5	41.3	36.6	39.1	42.4	49.0	39.9	37.0	51.9	63.3	40.9	41.3	40.3	29.8	28.9	41.1
Zhang [Zhang *et al.*, 2022] (HR-Net, T=243)	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
HDFormer(CPN, T=96)	38.1	43.1	39.3	39.4	44.3	49.1	41.3	40.8	53.1	62.1	43.3	41.8	43.1	31.0	29.7	42.6
HDFormer (HR-Net, T=96)	34.7	41.7	36.0	38.4	41.1	45.3	39.6	37.4	49.0	63.1	39.8	38.9	40.2	29.3	29.1	40.3
Protocol #2	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Cai [Cai *et al.*, 2019] (CPN, T=7)	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
Pavllo [Pavllo *et al.*, 2018] (CPN, T=243)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Xu [Xu *et al.*, 2020] (CPN, T=9)	31.0	34.8	34.7	34.4	36.2	43.9	31.6	33.5	42.3	49.0	37.1	33.0	39.1	26.9	31.9	36.2
Liu [Liu *et al.*, 2020](CPN, T=243)	32.3	35.2	33.3	35.8	35.9	41.5	33.2	32.7	44.6	50.9	37.0	32.4	37.0	25.2	27.2	35.6
Wang [Wang *et al.*, 2020] (CPN, T=96)	32.9	35.2	35.6	34.4	36.4	42.7	31.2	32.5	45.6	50.2	37.3	32.8	36.3	26.0	23.9	35.5
Hu [Hu *et al.*, 2021] (CPN, T=96)	29.8	34.4	31.9	31.5	35.1	40.0	30.3	30.8	42.6	49.0	35.9	31.8	35.0	25.7	23.6	33.8
Zhang [Zhang *et al.*, 2022] (CPN, T=81)	32.0	34.2	31.7	33.7	34.4	39.2	32.0	31.8	42.9	46.9	35.5	32.0	34.4	23.6	25.2	33.9
Wang [Wang *et al.*, 2020] (HR-Net, T=96)	28.4	32.5	34.4	32.3	32.5	40.9	30.4	29.3	42.6	45.2	33.0	32.0	33.2	24.2	22.9	32.7
Hu [Hu *et al.*, 2021] (HR-Net, T=96)	27.7	32.7	29.4	31.3	32.5	37.2	29.3	28.5	39.2	50.9	32.9	31.4	32.1	23.6	22.8	32.1
Zhang [Zhang *et al.*, 2022] (HR-Net, T=243)	28.0	30.9	28.6	30.7	30.4	34.6	28.6	28.1	37.1	47.3	30.5	29.7	30.5	21.6	20.0	30.6
HDFormer (CPN, T=96)	29.6	33.8	31.7	31.3	33.7	37.7	30.6	31.0	41.4	47.6	35.0	30.9	33.7	25.3	23.6	33.1
HDFormer (HR-Net, T=96)	27.9	32.8	29.7	30.6	32.5	35.0	28.9	29.2	38.3	50.0	32.9	30.1	31.8	23.6	22.8	31.7

Table 1: Quantitative comparisons with state-of-the-art methods on Human3.6M under protocol #1 and protocol #2, where methods marked with † are video-based; T denotes the number of input frames; and CPN and HR-Net denote the input 2D poses are estimated by [Chen *et al.*, 2017] and [Sun *et al.*, 2019], respectively. The best results of CPN and HR-Net are marked in red and blue, respectively.

Methods	MPJPE[↓]	Params	Latency	Frames
U-CondDGCN [Hu *et al.*, 2021]	22.7	3.4 M	0.6 ms	96
Cai [Cai *et al.*, 2019]	37.2	5.04M	11.6ms	7
Pavllo [Pavllo *et al.*, 2018]	37.2	17.0M	-	243
Liu [Liu *et al.*, 2020]	34.7	11.25M	9.9ms	243
Wang [Wang *et al.*, 2020]	25.6	1.69M	-	96
MixSTE [Zhang *et al.*, 2022]	25.9	33.7M	2.6ms	81
MixSTE ‡[Zhang *et al.*, 2022]	21.6	33.8M	8.0 ms	243
HDFormer*	22.1	2.8 M	0.9 ms	96
HDFormer	**21.6**	3.7 M	1.3 ms	96

Table 2: Results on Human3.6M with ground-truth 2D poses as input. Our method with subtraction feature representation is marked with *. The latency is measured with batch size = 1.

Methods	PCK[↑]	AUC[↑]	MPJPE[↓]
CNN [Mehta *et al.*, 2016b]	75.7	39.3	-
VNect(ResNet101) [Mehta *et al.*, 2017]	79.4	41.6	-
TrajectoryPose3D [Lin and Lee, 2019]	83.6	51.4	79.8
UGCN [Wang *et al.*, 2020]	86.9	62.1	68.1
U-CondDGCN [Hu *et al.*, 2021]	97.9	69.5	42.5
MixSTE [Zhang *et al.*, 2022](T=27)	94.4	66.5	54.9
HDFormer(T=32)	96.8	64.0	51.5
HDFormer(T=96)	**98.7**	**72.9**	**37.2**

Table 3: Results on MPI-INF-3DHP with three metrics.

Methods	Order	MPJPE[↓]
HDFormer	1	25.0
	2	23.6
	3	22.8
	4	**21.6**
	5	22.7

Table 4: Ablation study of the order number. We compared the results of different orders involved in the high-order attention transformer block in Human3.6M with ground truth as input.

sets (0,7,8,9,10), (0,7,8,11,12), (0,7,8,14,15), (7,8,11,12,13), which maps to the upper body parts and head. Besides, the **knee** joints are crooked when eating in a sitting pose as Fig. 4 (b). We also evaluate the visual result of estimated poses in Fig. 5. It can be seen that HDFormer estimates more accurate poses in cluttered and self-occlusion hands and feet compared to MixSTE [Zhang *et al.*, 2022].

4.5 Ablation Study

Impact of Multi-Order Attention. We evaluated the influence of our multi-order attention schema by conducting ablation studies on various order combinations, as shown in Tab. 4. The experiments, with Human3.6M 2D ground truth key points as input, indicated an improvement in the Mean Per Joint Position Error (MPJPE) as the order number increased, with the best performance observed at $order = 4$. This affirms the value of high-order attention in capturing complex skeletal information through "hyperbone↔joint" feature interaction.

Effectiveness of HDformer Block at Different Stages. We conducted experiments by adopting HDFormer block at various stages in Tab. 5. Compared with adopting HDFormer at the downsampling stage (yields 6.9mm decrease), adopting it at all stages (yields 3.3mm decrease) and adopting it at the merge stage get the best performance (yields 4.0 mm improvement). We found that adopting HDFormer block at the merging stage achieves a better result than other stages. This could be due to the complex skeleton dynamics of hyperbones can not be learned at early stages (i.e., downsam-

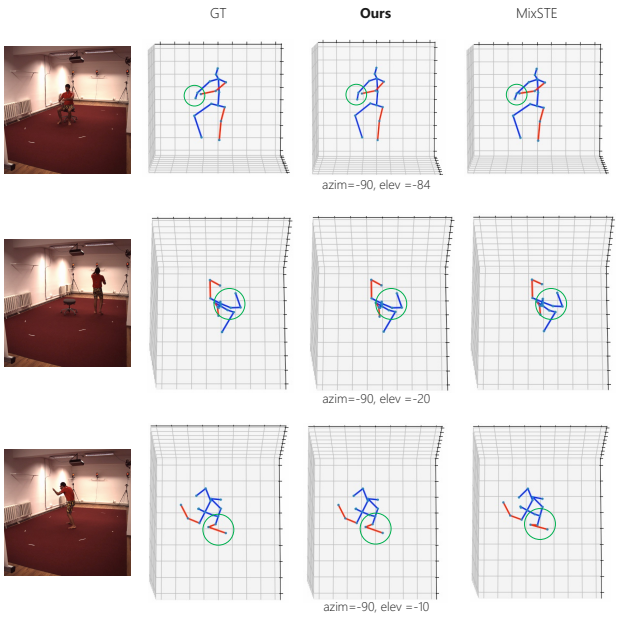

Figure 5: Qualitative comparison between our method (HDFormer) and [Zhang *et al.*, 2022] with the Eating (first and second row) and Photo (third row) actions on Human3.6M. The green circle highlights locations where our method has better results.

Methods	High-order Attention	MPJPE[↓]	Δ
Baseline	-	25.6	-
HDFormer	Upsampling stage	24.1	1.5
	Downsampling stage	32.5	-6.9
	Merging stage	**21.6**	**4.0**
	All stage	29.9	-3.3

Table 5: Ablation study of the effectiveness of HDFormer at different stages. We compared the results of the baseline (UGCN [Wang *et al.*, 2020]), our HDFormer, and different configurations for our HDFormer on Human3.6M. The Δ denotes the improvements compared with the baseline.

pling stage), therefore, leads to inferior performance. While in the merging stage, HDFormer block can fuse complex information from previous stages and by high-order attention.

Exploration of Hyperbone Representation. The hyperbone representation is a vital factor for the graph skeleton structure, and we exploit the way of the hyperbone feature representation. We conducted experiments by adopting 4^{th} order HDFormer block with different instantiations modes, which can be seen in Tab. 6. We found that all the hyperbone representation methods outperform the baseline, as they utilize high-order information. Among them, subtraction + concatenation boosts the performance over baseline by 4.0mm. It shows that bone feature concatenation with shortest path aggregation is effective for hyperbone feature representation.

Role of Position Encoding and Multi-Head Attention. We observed from our experimental results (line 2 of Tab. 7) that incorporating absolute positional encoding led to a decrease in performance. Besides, we have conducted an ablation study on the use of concatenation and summation in the multi-head attention module, and we found that summation resulted in better performance, which can be shown in line 3 of Tab. 7. Consequently, we adopted the summation in the multi-head attention in our proposed model. We also found extending the input frame numbers to 243 led to a slight decline compared to 96 frames as shown in line 4 of 7. The reason might be the small scale of our model (1/10 compared to [Zhang *et al.*, 2022]) can not well capture temporal redundancy and noise in dense sequence.

Methods	hyperbone representation	MPJPE[↓]	Δ
Baseline	-	25.6	-
HDFormer	summation	23.5	2.1
	multiplication	23.1	2.5
	concatenation	23.5	2.1
	subtraction + concatenation	**21.6**	4.0

Table 6: Ablation study of hyperbone representation. We compared the results of different edge feature representations for hyperbone encoding in Human3.6M with ground truth as input.

Line	Methods	MPJPE[↓]	Params	Frames
1	HDFormer (w/o Ψ)	27.4	3.7 M	96
2	HDFormer (with pos encoding)	22.1	3.8 M	96
3	HDFormer (multi-head concat)	22.9	3.7 M	96
4	HDFormer (T=243)	21.8	4.7 M	96
5	HDFormer (proposed)	**21.6**	3.7 M	96

Table 7: More results on Human3.6M with GT 2D poses as input.

5 Conclusion

In this work, we propose a novel model named High-order Directed Transformer (HDFormer), which considers both "*joint*↔*joint*", "*bone*↔*joint*" and "*hyperbone*↔*joint*" connections. Specifically, we propose a hyperbone representation learning module and a high-order attention module to model complicated semantic relations between hyperbone and joint. We conduct extensive experiments to provide both quantitative and qualitative analysis. Our proposed method achieves state-of-the-art performances with only 1/10 parameters and a fraction of computational cost compared to recently published SOTA.

Acknowledgments

Zhi-Qi Cheng's research in this project was supported by the US Department of Transportation, Office of the Assistant Secretary for Research and Technology, under the University Transportation Center Program (Federal Grant Number 69A3551747111), as well as Intel and IBM Fellowships.

Contribution Statement

Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, and Zhi-Qi Cheng contributed equally to this work as co-first authors, with random ordering. They were involved in study design, experiment execution, manuscript writing, and discussions. Wangmeng Xiang is the corresponding author.

References

[Cai *et al.*, 2019] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat-Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. *IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019.

[Chen *et al.*, 2017] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2017.

[Chen *et al.*, 2021] Tianlang Chen, Chengjie Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:198–209, 2021.

[Cheng *et al.*, 2022] Zhi-Qi Cheng, Qi Dai, Siyao Li, Teruko Mitamura, and Alexander Hauptmann. Gsrformer: Grounded situation recognition transformer with alternate semantic attention refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3272–3281, 2022.

[Ding *et al.*, 2019] Jianbang Ding, Xuancheng Ren, Ruixuan Luo, and Xu Sun. An adaptive and momental bound method for stochastic learning. *ArXiv*, abs/1910.12249, 2019.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *ArXiv*, abs/1704.01212, 2017.

[Gong *et al.*, 2021] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2021.

[He *et al.*, 2021] Jun-Yan He, Xiao Wu, Zhi-Qi Cheng, Zhaoquan Yuan, and Yu-Gang Jiang. Db-lstm: Densely-connected bi-directional lstm for human action recognition. *Neurocomputing*, 444:319–331, 2021.

[He *et al.*, 2023] Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Wangmeng Xiang, Binghui Chen, Bin Luo, Yifeng Geng, and Xuansong Xie. Damo-streamnet: Optimizing streaming perception in autonomous driving. *arXiv preprint arXiv:2303.17144*, 2023.

[Hu *et al.*, 2021] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3d human pose estimation. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[Huang *et al.*, 2021] Siyu Huang, Haoyi Xiong, Zhi-Qi Cheng, Qingzhong Wang, Xingran Zhou, Bihan Wen, Jun Huang, and Dejing Dou. Generating person images with appearance-aware pose stylizer. In *International Joint Conference on Artificial Intelligence*, pages 623–629. International Joint Conferences on Artificial Intelligence, 2021.

[Ionescu *et al.*, 2014] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014.

[Iskakov *et al.*, 2019] Karim Iskakov, Egor Burkov, Victor S. Lempitsky, and Yury Malkov. Learnable triangulation of human pose. *IEEE/CVF International Conference on Computer Vision*, pages 7717–7726, 2019.

[Li *et al.*, 2020] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6172–6182, 2020.

[Li *et al.*, 2021] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2021.

[Lin and Lee, 2019] Jiahao Lin and Gim Hee Lee. Trajectory space factorization for deep video-based 3d human pose estimation. *ArXiv*, abs/1908.08289, 2019.

[Liu *et al.*, 2020] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen ching S. Cheung, and Vijayan K. Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5063–5072, 2020.

[Mehta *et al.*, 2016a] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *International Conference on 3D Vision*, pages 506–516, 2016.

[Mehta *et al.*, 2016b] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *International Conference on 3D Vision*, pages 506–516, 2016.

[Mehta *et al.*, 2017] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.*, 36:44:1–44:14, 2017.

[Pavllo *et al.*, 2018] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7745–7754, 2018.

[Qiao *et al.*, 2022] Jian-Jun Qiao, Zhi-Qi Cheng, Xiao Wu, Wei Li, and Ji Zhang. Real-time semantic segmentation

with parallel multiple views feature augmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6300–6308, 2022.

[Qiu *et al.*, 2019] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. *IEEE/CVF International Conference on Computer Vision*, pages 4341–4350, 2019.

[Scarselli *et al.*, 2009] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20:61–80, 2009.

[Shi *et al.*, 2018] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12018–12027, 2018.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7904–7913, 2019.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5686–5696, 2019.

[Tu *et al.*, 2023] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. *ArXiv*, abs/2304.10465, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[Wang *et al.*, 2020] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. *ArXiv*, abs/2004.13985, 2020.

[Wang *et al.*, 2021] Wenxiao Wang, Lulian Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. *ArXiv*, abs/2303.06908, 2021.

[Xiang *et al.*, 2022] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Language supervised training for skeleton-based action recognition. *ArXiv*, abs/2208.05318, 2022.

[Xu *et al.*, 2020] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 896–905, 2020.

[Yan *et al.*, 2019] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. *IEEE/CVF International Conference on Computer Vision*, pages 4393–4401, 2019.

[Ye *et al.*, 2022] Hang Ye, Wentao Zhu, Chun yu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision*, 2022.

[Zeng *et al.*, 2021] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. *IEEE/CVF International Conference on Computer Vision*, pages 11416–11425, 2021.

[Zhang *et al.*, 2022] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13222–13232, 2022.

[Zhao *et al.*, 2018] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 383–391, 2018.

[Zhao *et al.*, 2019] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3420–3430, 2019.

[Zhao *et al.*, 2022] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20415, 2022.

[Zheng *et al.*, 2021] Ce Zheng, Sijie Zhu, Mat'ias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *IEEE/CVF International Conference on Computer Vision*, pages 11636–11645, 2021.

[Zhou *et al.*, 2022] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *ArXiv*, abs/2211.09590, 2022.

[Zhou *et al.*, 2023] Yuxuan Zhou, Zhi-Qi Cheng, Jun-Yan He, Bin Luo, Yifeng Geng, Xuansong Xie, and Margret Keuper. Overcoming topology agnosticism: Enhancing skeleton-based action recognition through redefined skeletal topology awareness. *arXiv preprint arXiv:2305.11468*, 2023.

[Zhu *et al.*, 2021] Yiran Zhu, Xing Xu, Fumin Shen, Yanli Ji, Lianli Gao, and Heng Tao Shen. Posegtac: Graph transformer encoder-decoder with atrous convolution for 3d human pose estimation. In *International Joint Conference on Artificial Intelligence*, 2021.

[Zou and Tang, 2021] Zhiming Zou and Wei Tang. Modulated graph convolutional network for 3d human pose estimation. *IEEE/CVF International Conference on Computer Vision*, pages 11457–11467, 2021.