applied sciences

MDPI

Article

SCGFormer: Semantic Chebyshev Graph Convolution Transformer for 3D Human Pose Estimation

Jiayao Liang [1] and Mengxiao Yin [1,2,*]

[1] School of Computer and Electronic Information, Guangxi University, Nanning 530004, China; 2113391024@st.gxu.edu.cn
[2] Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004, China
* Correspondence: ymx@gxu.edu.cn

Abstract: With the rapid advancement of deep learning, 3D human pose estimation has largely freed itself from reliance on manually annotated methods. The effective utilization of joint features has become significant. Utilizing 2D human joint information to predict 3D human skeletons is of paramount importance. Effectively leveraging 2D joint data can improve the accuracy of 3D human skeleton prediction. In this paper, we propose the SCGFormer model to reduce the error in predicting human skeletal poses in three-dimensional space. The network architecture of SCGFormer encompasses Transformer and two distinct types of graph convolution, organized into two interconnected modules: SGraAttention and AcChebGconv. SGraAttention extracts global feature information from each 2D human joint, thereby augmenting local feature learning by integrating prior knowledge of human joint relationships. Simultaneously, AcChebGconv broadens the receptive field for graph structure information and constructs implicit joint relationships to aggregate more valuable adjacent features. SCGraFormer is tested on widely recognized benchmark datasets such as Human3.6M and MPI-INF-3DHP and achieves excellent results. In particular, on Human3.6M, our method achieves the best results in 9 actions (out of a total of 15 actions), with an overall average error reduction of about 1.5 points compared to state-of-the-art methods, demonstrating the excellent performance of SCGFormer.

Keywords: human pose estimation; graph convolution; adjacency matrix; transformer

Citation: Liang, J.; Yin, M. SCGFormer: Semantic Chebyshev Graph Convolution Transformer for 3D Human Pose Estimation. *Appl. Sci.* **2024**, *14*, 1646. https://doi.org/10.3390/app14041646

Academic Editor: Antonio Fernández-Caballero

Received: 18 January 2024
Revised: 10 February 2024
Accepted: 10 February 2024
Published: 18 February 2024

1. Introduction

Traditional deep learning models, such as LSTM [1] and CNN [2], have demonstrated notable performance in Euclidean space data (language, images, videos, etc.). However, they exhibit certain limitations when processing non-Euclidean space data, such as social networks and information networks. Graph convolutional networks (GCNs) [3,4], as a recently emerged class of generalized neural network structures based on graph structures, have garnered significant attention and research due to their unique computational capabilities. Scholars have introduced the abstract concept of graphs from graph theory to represent non-Euclidean structured data and leveraged graph convolutional networks to process graph data, delving deeper into its features and patterns. The human skeleton, representing joints as nodes and bones as edges, can be conceptualized as a graph structure, providing a new direction for advancements in human pose estimation tasks.

In recent years, 3D human pose estimation has gained significant attention in computer vision, with applications spanning virtual reality [5,6], motion recognition [7–10], motion tracking [11], and more. This field has seen advancements in both single-stage [12–18] and multi-stage regression approaches [19–24]. Single-stage regression directly extracts features from images to reconstruct 3D skeleton coordinates but is sensitive to external factors such as background and lighting. In contrast, multi-stage regression initially predicts the planar coordinates of the skeleton and subsequently estimates 3D features using 2D

joint information. The latter approach has reached maturity, benefiting from a plethora of methods for detecting 2D skeletons in RGB images.

Nonetheless, 3D pose estimation encounters challenges including depth ambiguity, self-obscuration, and a scarcity of outdoor datasets. Graph structures, which represent skeletal connections among 2D joints, have been harnessed by graph convolutional neural networks (GCNs) to acquire effective 3D human body representations [25–28]. However, prior GCN-based approaches [29] exhibited certain limitations. They often relied on first-order neighborhood matrices, constraining their ability to capture information from distant joints, and employed the same kernel matrix for all edges, resulting in underutilization of relationships between adjacent nodes. SemConv [27] improved the operation of graph convolution. Its key idea is to learn channel-wise weights for edges as priors implied in the graph and then combine them with kernel matrices. This significantly improves the power of graph convolutions. The work [28] introduces a graph-oriented Transformer structure that incorporates a self-attention mechanism to better capture the relationships between joint points and a graph convolutional structure to infer pose information in the graph structure.

In order to improve the utilization of joint features, we propose SCGFormer, a semantic Chebyshev graph convolutional Transformer. SCGFormer harnesses the strengths of Transformer and graph convolutions, comprising two pivotal components: semantic graph—attention (SGraAttention) and Chebyshev graph convolution with combined adjacency matrix information (AcChebGconv). SGraAttention amalgamates Transformer and semantic graph convolution, facilitating the extraction of global information through self-attention while preserving human kinematic constraints. Concurrently, AcChebGconv improves Chebyshev graph convolution by fully exploiting 2D joint features and reinforcing local connections. We assess the performance of SCGFormer on the Human3.6M [30,31] and MPI-INF-3HP [12] datasets, achieving outstanding results across various action categories. The task achieved the best results on 9 actions (from a total of 15 actions), with an overall average error decrease of approximately 1.5 points compared to the baseline.

In summary, our contributions are as follows:

- We propose SCGFormer, a novel network incorporating the SGraAttention and AcChebGconv modules, which improve the effectiveness of the network by applying effective human body structural constraints.
- We amplify the correlations between joints and their remote neighbors by blending first-order and second-order adjacency matrices in the AcChebGconv module, thereby maximizing the utilization of 2D joint features in our approach.
- We conducted experiments on well-established benchmarks to showcase the robustness and precision of the SCGraFormer in the domain of 3D pose estimation.

2. Related Work

In this section, we will present an overview of research pertaining to 3D human pose estimation. Following this, we will introduce approaches that are particularly pertinent to the subject of this paper, encompassing both Transformer-based and graph-convolution-based methods.

2.1. 3D Human Pose Estimation

Three-dimensional (3D) human pose estimation represents an important research area within computer vision [32,33]. Its fundamental aim is to ascertain the spatial coordinates of human joints in three-dimensional space, primarily utilizing RGB images or video data. In its early stages, this task [31,34–36] heavily relied on handcrafted features and geometric constraints as the means to predict 3D human pose. Nevertheless, with the rapid evolution of deep learning, deep neural networks have emerged as a predominant approach for 3D human skeleton prediction. These approaches can be broadly categorized into two primary types: single-stage regression and multi-stage methods.

Single-stage regression stands as a direct approach to the end-to-end estimation of 3D human pose from images, characterized by its simplicity and effectiveness. For instance, Pavlakos et al. [13] employed voxel likelihood to represent the confidence of joint positions in 3D space and inferred joint details via 3D heat maps. Nevertheless, their model exhibits sensitivity to less pertinent factors, such as background and lighting conditions, and its generalization is impeded by a restricted volume of indoor data. In a different vein, Zhou et al. [15] adopted a weakly supervised training model that combined annotated 2D outdoor images with annotated 3D indoor images as inputs, effectively mitigating the prediction ambiguity from 2D to 3D. Additionally, Yang et al. [14] employed an adversarial network featuring a multi-source discriminator to compel the generator to produce plausible 3D human poses.

In contrast, multi-stage methods first employ CNN networks to detect 2D joint coordinates, followed by the utilization of this information in a 3D information detector to infer 3D human pose. These approaches heavily rely on a mature 2D pose detector [23,24] and primarily focus on the transition from 2D to 3D, a strategy that often yields heightened data accuracy. For instance, Martinez et al. [37] introduced a straightforward yet highly effective 3D pose detector that achieves precise predictions through a simplistic network structure. In another vein, Wandt et al. [38] proposed a generative adversarial training approach to address the shortage of labeled 3D datasets. They employed two parallel networks to simultaneously predict 3D human pose and camera parameters, harnessing a discriminator to enhance pose generation accuracy. Additionally, they re-projected generated 3D poses back to 2D poses, thereby enriching the training data.

The characteristics of 2D joint data differ from image data. While it avoids interference from factors such as background lighting in images, the compressed and concise feature information in 2D joint data necessitates an appropriate prior constraint for further discerning useful information between joints. Moreover, dealing with the continuity of a sequence of movements requires addressing the connections between joint data frames, emphasizing the significance of enlarging the receptive field. Disregarding these issues may compromise the accuracy of 3D human pose estimation. Therefore, effectively leveraging 2D joint data in multi-stage regression tasks becomes crucial. Based on this, our approach adopts a multi-stage regression method that integrates both graph convolution and Transformer techniques. In the subsequent sections, we will introduce recent research endeavors pertaining to human pose prediction through the utilization of graph convolution and Transformer.

2.2. Graph-Convolution-Based Methods

In recent years, significant strides have been made in the field of 3D human pose estimation through graph convolution, resulting in cutting-edge outcomes [26–28,39]. Zhao et al. [27] introduced a semantic graph convolutional network featuring a stacked structure as a non-local module. This architectural choice facilitated the learning of weights between neighboring nodes, consequently enhancing the connections among 2D joints. Furthermore, Xu and Takano [26] proposed an innovative graph hourglass network model. This model employed a distinctive combination of pooling and anti-pooling layers to acquire intermediate features, which were subsequently fused with the SE block [40]. In another notable development, Zhao et al. [28] amalgamated the strengths of both Transformer and graph convolutions. Initially, they extracted global features using the conventional Transformer [41], excluding the MLP [42]. Subsequently, they harnessed the implicit higher-order connections between joints by employing Chebyshev graph convolution [43]. This integrated approach yielded state-of-the-art results on the Human3.6M dataset, surpassing other graph-convolution-based methods. However, it is essential to acknowledge that this approach, despite its advantages in terms of parameter count, utilizes a learnable adjacency matrix following the global feature extraction stage. This allows the network to learn without constraints, introducing an element of uncertainty. The learned adjacency matrix may potentially disrupt the kinematic structure of the human skeleton and influence the accuracy of the data refined by subsequent network layers.

2.3. Transformer-Based Methods

The Transformer [41], initially prominent in the domain of natural language processing, has recently gained traction in the realm of computer vision tasks. Veličković et al. [44] introduced graph-attention networks that employ self-attention mechanisms to learn node weights within a graph, highlighting the challenging issue of limited receptive field size for graph convolution. In a pioneering effort, Zheng et al. [45] developed the first 3D human pose estimation model based on Transformer. They globally modeled relationships among human body joints within frames and temporal correlations between frames, ultimately predicting 3D human poses based on intermediate structures derived from frame averaging. On a related note, Lin et al. [16] proposed an end-to-end network rooted in the Transformer framework for the reconstruction of 3D human body poses and mesh vertices from a single image. Although both of these approaches [16,45] employ a Transformer to capture global features, thus addressing the issue of limited receptive field size, they neglect the potential utility of joint graphs. Neglecting human geometric constraints can lead to an incapacity to infer plausible joint information, ultimately resulting in a reduction in prediction accuracy.

Our approach leverages self-attention to extend the receptive field, simultaneously enhancing the interconnection among human joints through the integration of the graph convolution layer, thereby contributing to an overall enhancement in model performance. A detailed exposition of this approach will be provided in the third section.

3. Main Work

The model presented in this paper, SCGFormer, represents a fusion of Transformer and Graph Convolution methods. Figure 1 provides an overview of the comprehensive framework, wherein it takes 2D joint coordinates as input and yields predicted 3D poses as output. In this section, we will commence by introducing the Preliminaries in Section 3.1, with further elaboration provided in subsequent subsections under Section 3.2.

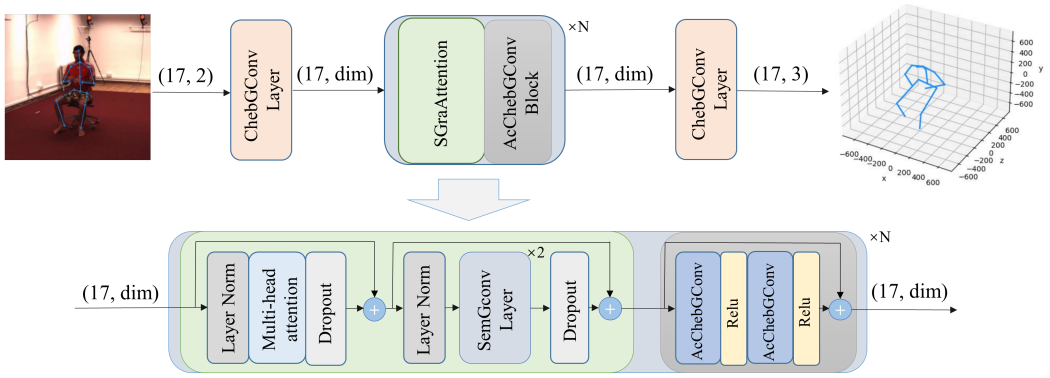

Figure 1. The general framework and core module structure of SCGFormer. By leveraging the provided 2D joint information, SCGFormer establishes a vital linkage between 2D and 3D skeletal information, facilitated by the incorporation of SGraAttention and the AcChebGConv block. This orchestrated process culminates in the accurate prediction of intricate 3D skeletal details.

3.1. Preliminaries

A human skeleton can be represented as a graph and predicted using graph convolution. Let J denote the number of joints, and D_l represent the dimension of the input data. $X^l \in R^{J \times D_l}$ is the input of the L-th layer. The two-dimensional joint coordinates are initialized as $X^0 \in R^{J \times 2}$, which is the input of SCGFormer. Then, the output X^{l+1} of the L-th layer is:

$$X^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\bar{A}\tilde{D}^{-\frac{1}{2}}X^l W), \quad (1)$$

where σ is the activation function RELU, $\bar{A} = A + I$, $A \in R^{J \times J}$ denotes the adjacency matrix of the joints, I is the unit matrix, $\tilde{D}$ means the diagonal matrix whose elements are the degrees of the nodes, and $W \in R^{D_l \times D_{l+1}}$ indicates a learnable parameter matrix

of the layer. The adjacency matrix A cannot aggregate the feature information of its joint, so the matrix $\bar{A}$ is obtained from A by adding the self-loop, which can learn the feature information of the joint itself. In the following two graph convolutions, we use $\bar{A}$, which helps to understand the correlation between joints.

Inspired by the work [27], we add a semantic graph convolution (SemGconv) to the SGraAttention module, which better uses the adjacent node relations and graph structure. The general output X^{l+1} of the L-th layer of the semantic graph convolution is as follows:

$$X^{l+1} = \sigma(WX^l\rho_i(M \odot \bar{A})), \tag{2}$$

$$M \odot \bar{A} = \begin{cases} -\infty & \text{if } a_{ij} \neq 1 \\ m_{ij} & \text{if } a_{ij} = 1 \end{cases}, \tag{3}$$

In Equation (2), ρ_i is the activation function to normalize the numerical weights; $M \in R^{J \times J}$ is a learnable weight transformation matrix that performs elemental operations $\odot$ on the adjacency matrix $\bar{A} = [a_{ij}]$, the specific formula of which is as shown in Equation (3). If $a_{ij} = 1$, then the value of the corresponding position m_{ij} is returned; otherwise, negative infinity is returned. In the actual network layer, a smaller value (-9×10^{15}) will be taken instead of negative infinity. W is the parameter matrix of the layer and is divided into two parts, W_0 and W_1. W_0 relates to the feature-transformed representation of the node itself, while W_1 learns the feature representations of all nodes except its own. Then, the detailed output X^{l+1} of the L-th layer in the semantic graph convolution is as follows:

$$X^{l+1} = \sigma(I \otimes W_0X^l\rho_i(M \odot \bar{A}) + (1 - I) \otimes W_1X^l\rho_i(M \odot \bar{A})), \tag{4}$$

where $\otimes$ represents the element-by-element multiplication in the matrix, all nodes of the previous graph convolution share a parameter matrix [29], and semantic graph convolution [27] expresses the correlation between individual nodes better.

The Transformer model, which is now the mainstream for natural language, has achieved good results in computer vision [16,44–46] due to its self-attention mechanism [41]. The self-attention mechanism processes the feature tensor X^l through three multilayer perceptions (MLPs) [42] to output three matrices: Q^l, K^l, and V^l. Then, the output feature is obtained by matrix multiplication. The output of the L-th layer of the Transformer is formulated as follows:

$$X^{l+1} = \sigma(\frac{Q^l{K^l}^T}{\sqrt{d_k}})V^l, \tag{5}$$

where σ represents softmax [47,48]. Its elements are scaled individually with $\sqrt{d_k}$ to avoid a sharp distribution of matrix elements.

We chose the network layers introduced above to compose our network architecture. Initially, we employ the Transformer to globally extract joint features, addressing the issue of a relatively small receptive field. This aids in establishing connections between joint frames. Subsequently, for the globally extracted information, which does not inherently contain knowledge of human kinematic structure, we utilize semantic graph convolution with a first-order adjacency matrix incorporating information about human joints to refine the local information of joints. Following this, Chebyshev graph convolution with K-order Chebyshev functions is applied to extend more comprehensive local feature information. This approach is adopted to construct a more effective human skeleton.

3.2. SCGFormer

The architecture of the novel network model proposed in this paper is visually represented in Figure 1. Specifically, the Chebyshev graph convolution serves a dual role, encompassing the preprocessing of the initial feature data as well as the regression of the final decoded features. In the ensuing sections, we will provide an in-depth exploration of the two central modules integral to SCGFormer, namely the SGraAttention module and the AcChebConv module.

3.2.1. SGraAttention Module

The SGraAttention module is equipped with two residual connections [49], primarily comprising a multi-head self-attention mechanism and two layers of semantic graph convolution, along with a LayerNorm (LN) layer [50] and a dropout layer [51]. As illustrated in Figure 2, the 17 joints undergo pre-encoding in the Chebyshev graph convolution layer before being fed into the SGraAttention module. The input feature vectors are initially normalized by the LayerNorm (LN) layer [50] and subsequently passed through the multi-head self-attention layer. The multi-head self-attention module employed here corresponds to the traditional Transformer [41] architecture. Initially, global feature information is extracted through the self-attention mechanism, as depicted in Figure 3a, aiding the model in comprehending contextual relationships among human joint points.

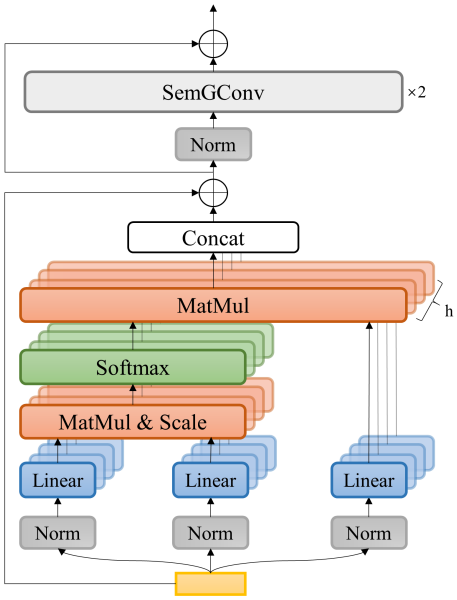

Figure 2. Structure of SGraAttention module.

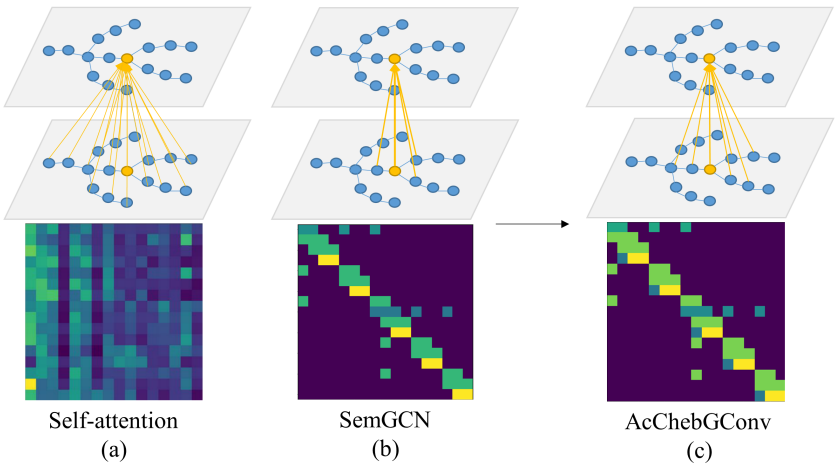

Figure 3. Strategies for human skeleton feature extraction in the model structure. (**a**) the weight matrix V in self-attention is capable of capturing global information; (**b**) the adjacency matrix in semantic graph convolution, with constraints based on human body structure, is utilized to enhance local information; (**c**) the adjacency matrix in Chebyshev graph convolution, combining adjacency matrix information, is employed to broaden the receptive field, further strengthening local information. The thickness of the yellow arrows represents the strength of the correlation and the black arrow indicate the order of feature extraction.

Following the self-attention mechanism, we eliminate the MLP [42] and introduce semantic graph convolution layers. While the MLP layer possesses robust data fitting capabilities, its excessive parameter count results in some neurons not contributing to the transmission of network information, leading to a wastage of spatial resources. Hence, we opt for semantic graph convolution layers as a replacement for the MLP. Semantic graph convolution layers are more suitable for handling graph-structured data like human skeletal data. Each layer enhances the local feature information of human joint points by learning correlated weights between joints, as illustrated in Figure 3b.

The structure of the SGraAttention module is depicted in Figure 2, where 'h' represents the number of heads. The module initially utilizes a multi-head self-attention mechanism to capture the global information of joint points, followed by semantic graph convolution to refine the preliminary features of joint points.

3.2.2. AcChebGConv Module

The second crucial module is the Chebyshev graph convolution [43]. We introduce second-order neighborhood information into the prior constraints of Chebyshev graph convolution. Specifically, the adjacency matrix that combines both first-order and second-order information is employed as the primary constraint. This enhanced module is referred to as the AcChebGconv module, denoting Chebyshev graph convolution with a fusion of adjacency matrix information. The AcChebGConv module comprises two instances of AcChebGconv and two rectified linear unit (ReLU) layers, as depicted in Figure 4. Building upon the globally refined preliminary features extracted in the preceding module, this module further refines local feature information by expanding the receptive field, as illustrated in Figure 3c.

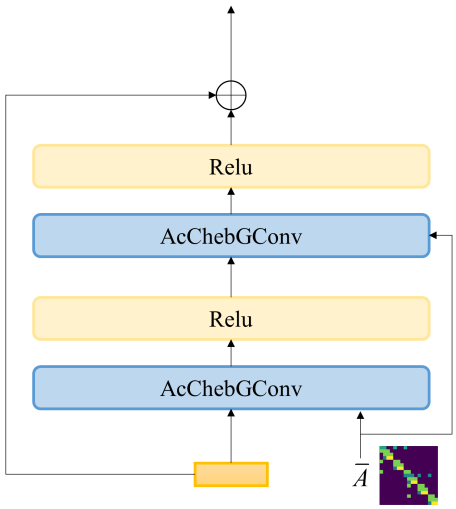

Figure 4. Structure of AcChebGconv module.

The output X^{l+1} at the L-th layer of the AcChebGConv module is given by:

$$X^{l+1} = \sum_{k=0}^{K-1} T_k(\tilde{L}) X^l \theta_k, \tag{6}$$

where T_k is the k-th order chebyshev correlation function (i.e., $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$), $k \geq 2$, $T_0 = 1$, and $T_1 = x$. $\theta_k \in R^{D_l \times D_{l+1}}$ is the matrix of trainable parameters in the graph convolution of this layer. $\tilde{L} = 2L/\lambda_{max} - I$ denotes the Laplacian operator based on the maximum eigenvalue rescaling, and λ_{max} is the maximum eigenvalue in the normalized graph Laplacian L. The normalized graph Laplacian L is formulated as follows:

$$L = I - \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \tag{7}$$

where the adjacency matrix $\tilde{A} = A + A^2$, it incorporates more extensive adjacency information by summing the first-order adjacency matrix with the second-order adjacency matrix.

In alignment with the existing literature [28], the updated Chebyshev graph convolution layer inherits the K-order polynomial function from its foundational Chebyshev graph convolution, thereby affording the advantage of an expanded receptive field. This extension allows for the aggregation of information from the K nearest neighbors. Furthermore, the innovative constraint introduced here, which integrates the second-order adjacency matrix, enhances the local structural constraints of the human body. This augmentation, in conjunction with the semantic graph convolution, leads to a noteworthy enhancement in the model's performance, as will be substantiated in the forthcoming experimental sections.

Our model incorporates two variants of graph convolution: the semantic graph convolution and the Chebyshev graph convolution enriched with fused adjacency matrix information. In comparison to prior graph convolution methods [28], although the computational load is notably increased, the required dataset remains relatively uncomplicated, involving only 17 nodes. As a result, the computation entails the calculation of a 17×17 matrix.

3.2.3. Loss Function

Given the human joint set $S = \left\{ J_i^{2d}, J_i^{3d} \right\}_{i=1}^{N}$, where $J_i^{2d} \in R^{j \times 2}$ is the 2D ground truth of the human joint, $J_i^{3d} \in R^{j \times 3}$ is the corresponding 3D ground truth, and $N = 17$ is the number of joints that make up the human skeleton. We use the 3D ground truth in the dataset to correct the predicted 3D value, and the 2D ground truth is used to evaluate the model's performance. We train the SCGFormer with the mean square error L (i.e., Equation (8)) to minimize the error between the predicted and ground truth.

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(\left\| \tilde{J}_i^{3d} - J_i^{3d} \right\|^2 \right), \tag{8}$$

where $\tilde{J}_i^{3d} \in R^{j \times 3}$ is the 3D coordinate value predicted by the network. The error measure is in millimeters.

The algorithm is shown in Algorithm 1.

Algorithm 1 Training

Input: 2D joint data of CPN network $\tilde{J}^{2d}$, first order adjacency matrix $\bar{A}$, adjacency matrix $\tilde{A}$ that combines first- and second-order adjacency information, the ground truth of 3D human joint data J^{3d}.
Output: 3D human skeleton $\tilde{J}^{3d}$ predicted by the network

repeat
Input $\tilde{J}^{2d}$ into the preprocessing layer and project it onto high-dimensional features $\tilde{J}^{H}$.
Use Transformer to extract features from $\tilde{J}^{H}$ based on Equation (5)
Apply SemGConv and combine with prior constraint $\bar{A}$ to extract features according to Equation (3).
Apply AcChebGConv and combine with prior constraint $\tilde{A}$ to extract features according to Equation (6).
Map the extracted high-dimensional feature values back to the 3D human skeleton $\tilde{J}^{3d}$ through a decoder.
Take gradient descent step on $\nabla_\theta \left\| \tilde{J}^{3d} - J^{3d} \right\|^2$
until converged

4. Experiments

In this section, we will commence by providing an elaborate account of the experimental procedures, encompassing the configuration of various training parameters employed throughout the experiments. Subsequently, we will conduct an in-depth analysis of the

experimental outcomes achieved by SCGFormer, comparing them with those of state-of-the-art methods. Finally, we will substantiate the effectiveness of SCGFormer through a series of ablation experiments.

4.1. Experimental Details

The dataset, evaluation metrics, and training details used in this work are described below.

Dataset: We use two popular human pose datasets, Human3.6M [30,31] and MPI-INF-3DHP [12], to evaluate the SGraFormer.

Human3.6M [30,31] is the most widely used dataset for 3D human pose estimation tasks [28,30,52]. It encompasses a diverse set of 15 actions, ranging from common activities such as greeting and smoking to walking a dog, all enacted by a total of 11 different actors, captured across four cameras. Within the dataset, subjects denoted as S1, S5, S6, S7, and S8, sourced from seven actors, serve as the training and validation sets. Conversely, subjects S9 and S11, corresponding to the remaining four actors, are exclusively designated for testing purposes. In aggregate, the dataset comprises 1,559,752 frames for training and 543,344 frames for testing.

MPI-INF-3DHP [12] presents images captured in three distinct scenarios: a studio environment with a green screen (GS), a studio environment without a green screen (noGS), and an outdoor setting (Outdoor). We have employed this dataset to assess the generalization capabilities of the proposed architecture.

Evaluation metrics: In the case of Human3.6M [30,31], two distinct evaluation protocols, **Protocol 1** and **Protocol 2**, have been established, delineating different approaches based on subject selection within the dataset [26,27,37]. **Protocol 1** involves the utilization of subjects S1, S5, S6, S7, and S8 as the training set, with subjects S9 and S11 designated for testing. Conversely, **Protocol 2** employs subjects S1, S5, S6, S7, S8, and S9 for training, reserving S11 as the sole component of the testing set. In accordance with recent research practices [26–28,37,52,53], we have adhered to the first evaluation protocol. Regarding the MPI-INF-3DHP dataset [12], we have adhered to established conventions [26,28] by employing 3D-PCK (percentage of correct keypoints) and AUC (area under the curve) as the chosen evaluation metrics.

Protocol 1 uses the mean error per joint position (MPJPE) $E_{MPJPE}(f, S)$ as the evaluation metric, as shown in Equation (9). Given skeleton S with N_S joints, it first aligns the root joint (the pelvic joint), then calculates the mean error of the joint position. The smaller the value, the more accurate the prediction.

$$E_{MPJPE}(f, S) = \frac{1}{N_S} \sum_{i=1}^{N_S} \left\| m_{pe,S}^{(f)}(i) - m_{gt,S}^{(f)}(i) \right\|_2, \tag{9}$$

where $m_{pe,S}^{(f)}(i)$ is a function that returns the coordinates of the i-th joint of skeleton S, at frame f, from the pose estimator pe, and $m_{gt,S}^{(f)}(i)$ is the i-th joint of the ground truth frame f.

Protocol 2 employs P-MPJPE (Procrustes mean per joint position error) as the evaluation metric, wherein the predicted values are aligned with the ground truth through rigid transformations, encompassing translation and rotation. Subsequently, MPJPE is computed to quantify the error, with millimeters serving as the unit of measurement. It is noteworthy that the performance outcomes under both protocols typically exhibit consistency. Therefore, in alignment with established practices in the literature [27,28], we have chosen to utilize **Protocol 1**, specifically the mean error per joint position (MPJPE), as the designated evaluation metric for our experiments.

Training parameter settings: During the training phase, we configured the SCGFormer model with a total of five SGraAttention modules (as denoted by N in Figure 1), and each multi-headed self-attention utilized four heads (H). The intermediate feature dimension for data processing within the model was set to 96, accompanied by a dropout rate of 0.25. For optimization during training, we employed the classical Adam [54] optimizer, commencing with an initial learning rate of 0.001. The training data was organized into batches of size 64, and a learning rate decay rate of 0.9 was applied. Specifically, the current learning rate

was decreased by a factor of 0.9 every 50,000 steps throughout the training process, which spanned a duration of 100 epochs in total to train the SCGFormer model.

4.2. Comparison with State-of-the-Art Methods

In previous research, 3D human pose estimation methods have been categorized into two distinct groups based on their network inputs. The first category involves the direct extraction of essential feature information from images to predict joint points [12–15,17,18,55,56]. The second category utilizes a well-established 2D pose detector to forecast crucial point information, after which the network deduces the 3D spatial positions of the joints [25–28,37,53,57,58]. The results comparing the model presented in this paper with previous works [12–15,18,25–28,37,53,55–58] are summarized in Table 1.

Among the aforementioned studies, some [12–15,18] have adopted an end-to-end approach that takes images as input. However, these approaches have yielded unsatisfactory estimation results primarily due to the limited feature information available in the images. When confronted with occlusion issues, these models, which rely solely on a single view, struggle to accurately infer hidden features. Furthermore, these models exhibit heightened sensitivity to various factors, such as background and lighting conditions, which contributes to reduced model robustness and generalization. In contrast, leveraging the two-dimensional skeleton provides access to a wealth of spatial kinematic information [28,37], enabling the exclusion of extraneous image features during model training. As depicted in Table 1, all of these models utilize the output from the cascaded pyramid network (CPN) [21] as their input. Notably, among the 15 actions assessed in the Human3.6M [30,31] dataset, our approach achieves the lowest overall prediction error for nine actions and excels in scenarios involving occluded frames, such as greeting, phoning, and sitting. The average error across all actions experiences a substantial reduction of nearly one and a half points compared to the most recent work [28]. The bar chart is shown in Figure 5. These results demonstrate the superior performance of our method in comparison to state-of-the-art techniques, with the exception of action poses. Furthermore, when compared to other methods [26,27,53] that employ graph convolution, our approach consistently delivers superior results for almost all actions.

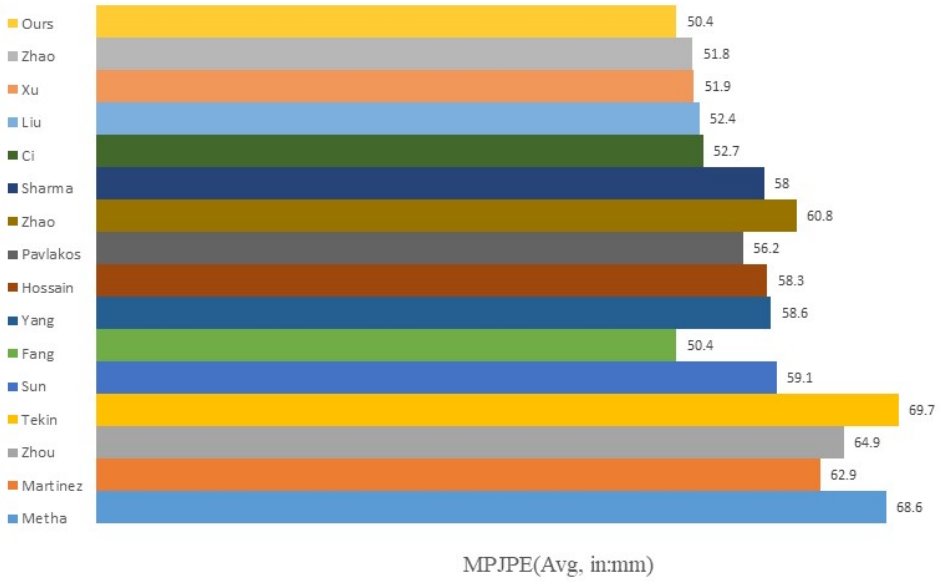

Figure 5. Bar chart of the overall average MPJPE (in mm) corresponding to the methods in Table 1.

In Table 2, we conducted experiments using the ground truth (GT) values of the 2D key points. Our method outperforms others by achieving the best results for nine actions and yielding the lowest overall mean error. When compared to recent works utilizing graph convolution [26–28,53], our model excels in handling accurate 2D information, demonstrating our significant advancements in processing 2D information with noise. The bar chart of average error is shown in Figure 6.

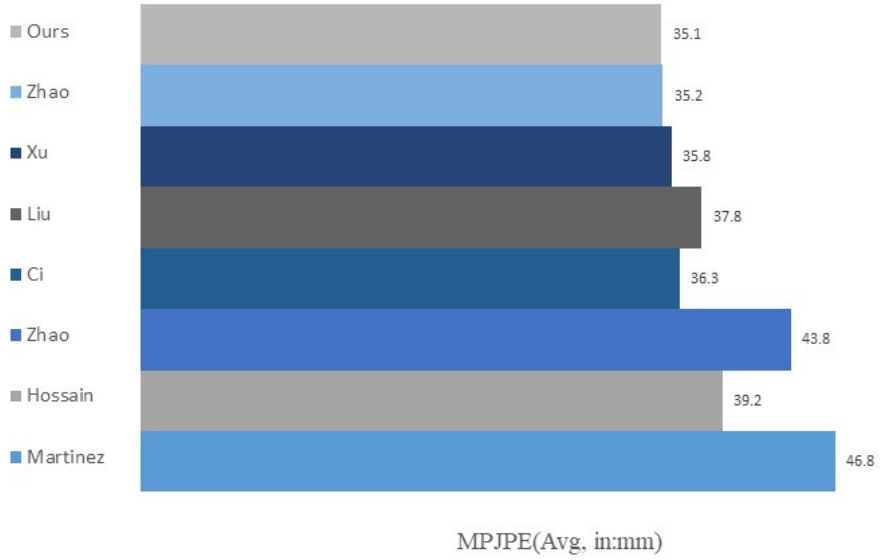

Figure 6. Bar chart of the overall average MPJPE (in mm) corresponding to the methods in Table 2.

We visualized the predictive performance of a majority of actions, as depicted in Figure 7, encompassing various movements such as posing (Pose), waiting (Wait), sitting down (SittingD), walking (Walk), and walking together (WalkT). The corresponding errors for these actions are tabulated in Table 1, measured in millimeters, as 48.3 for pose, 48.0 for wait, 66.4 for sitting, 38.9 for walking, and 42.1 for walking together. Among these actions, Wait, Walk, and WalkT demonstrate superior outcomes, with Pose achieving the second-best performance. For additional actions, including directions, walking dog (WalkD), phone usage (Phone), greeting (Greet), and sitting, the corresponding errors, reported in millimeters in Table 1, are 44.6, 52.7, 52.7, 49.4, and 58.2, respectively. In comparison to previous methodologies, Directions, Phone, Greet, and Sitting exhibit the best predictive performance, denoted by the lowest prediction errors within their respective action categories. A visual representation of these outcomes is presented in Figure 8. Remarkably, the predictions made by SCGFormer closely align with the ground truth labels in these scenarios. These findings underscore the model's good learning capability and predictive performance.

To assess the generalization prowess of our model, we trained it exclusively on the Human3.6M dataset [30,31] and subsequently evaluated it on the test set of MPI-INF-3DHP [12]. The outcomes are presented in Table 3. Impressively, our method outperforms the majority of other approaches, achieving an average PCK (percentage of correct keypoints) of 79.2 and an AUC (area under curve) of 43.9. These results underscore the robust generalization capabilities of our architecture, even when applied to previously unseen datasets.

Table 1. Results (in mm) of the quantitative evaluation using MPJPE on Human3.6M [30,31] according to protocol I. The notation (*) in the table indicates that the model uses the picture as input, while the rest use the 2D skeleton key points detected by the CPN network [21] as input, and the notation (+) indicates that the model was trained using additional data from MPII [59]. The best-performing data have been bolded.

Protocol#1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Pavlakos [13] (*)	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
Metha [12] (*)	52.6	64.1	55.2	62.2	71.6	79.5	52.8	68.6	91.8	118.4	65.7	63.5	49.4	76.4	53.5	68.6
Martinez [37]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Zhou [15] (*)	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	51.4	63.2	55.3	64.9
Tekin [17]	54.2	61.4	60.2	61.2	79.4	78.3	63.1	81.6	70.1	107.3	69.3	70.3	74.3	51.8	63.2	69.7
Sun [18] (+) (*)	52.8	54.8	54.2	54.3	61.8	**53.1**	53.6	71.7	86.7	61.5	67.2	53.4	47.1	61.6	53.4	59.1
Fang [57]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	50.4
Yang [14] (+) (*)	51.5	58.9	50.4	57.0	62.1	65.4	49.8	52.7	69.2	85.2	57.4	58.4	**43.6**	60.1	47.7	58.6
Hossain [58]	48.4	50.7	57.2	55.2	63.1	72.6	53.0	51.7	66.1	80.9	59.0	57.3	62.4	46.6	49.6	58.3
Pavlakos [55] (+)	48.5	54.4	54.5	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Zhao [27]	48.2	60.8	51.8	64.0	64.6	53.6	51.1	67.4	88.7	**57.7**	73.2	65.6	48.9	64.8	51.9	60.8
Sharma [56]	48.6	54.5	54.2	55.7	62.2	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
Ci [25] (+)	46.8	52.3	**44.7**	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
Liu [53]	46.3	52.2	47.3	50.7	55.5	67.1	49.2	**46.0**	60.4	71.1	51.5	50.1	54.5	40.3	43.7	52.4
Xu [26]	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Zhao [28]	45.2	50.8	48.0	50.0	54.9	65.0	**48.2**	47.1	60.2	70.0	51.6	48.7	54.1	39.7	43.1	51.8
Ours	**44.6**	**49.7**	46.2	**49.4**	**52.7**	61.1	48.3	46.5	**58.2**	66.4	**50.7**	**48.0**	52.7	**38.9**	**42.1**	**50.4**

Table 2. Results (in mm) of quantitative evaluation using MPJPE on Human3.6M [30,31] according to protocol I. The models in the table all use 2D key point ground truth as input, and the notation (+) indicates that the model was trained using additional data from MPII [59]. The best-performing data have been bolded.

Protocol#1	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SittingD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez [37]	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
Hossain [58]	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2
Zhao [27]	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	**42.2**	53.5	44.3	40.5	47.3	39.0	43.8
Ci [25] (+)	36.3	38.8	**29.7**	37.8	34.6	42.5	39.8	32.5	36.2	39.5	34.4	38.4	38.2	31.3	34.2	36.3
Liu [53]	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
Xu [26]	35.8	38.1	31.0	35.3	35.8	43.2	37.3	31.7	38.4	45.5	35.4	36.7	36.8	27.9	30.7	35.8
Zhao [28]	**32.0**	38.0	30.4	34.4	**34.7**	43.3	**35.2**	31.4	38.0	46.2	**34.2**	35.7	36.1	27.4	30.6	35.2
Ours	33.3	**36.9**	30.8	**33.5**	36.6	**41.2**	35.4	**31.2**	**37.5**	48.3	35.1	**35.6**	**34.5**	**26.9**	**30.2**	**35.1**

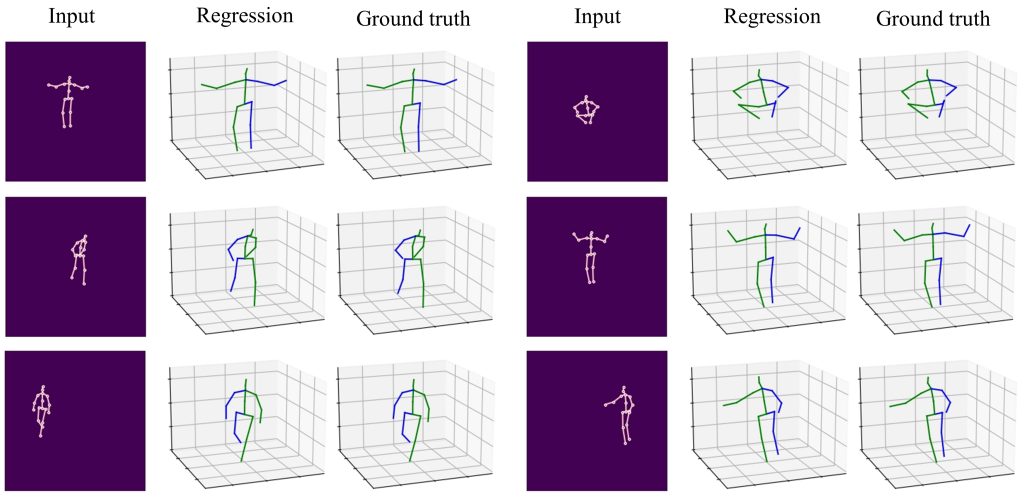

Figure 7. Visualization results of actions such as Pose, Wait, SittingD, Walk, and WalkT.

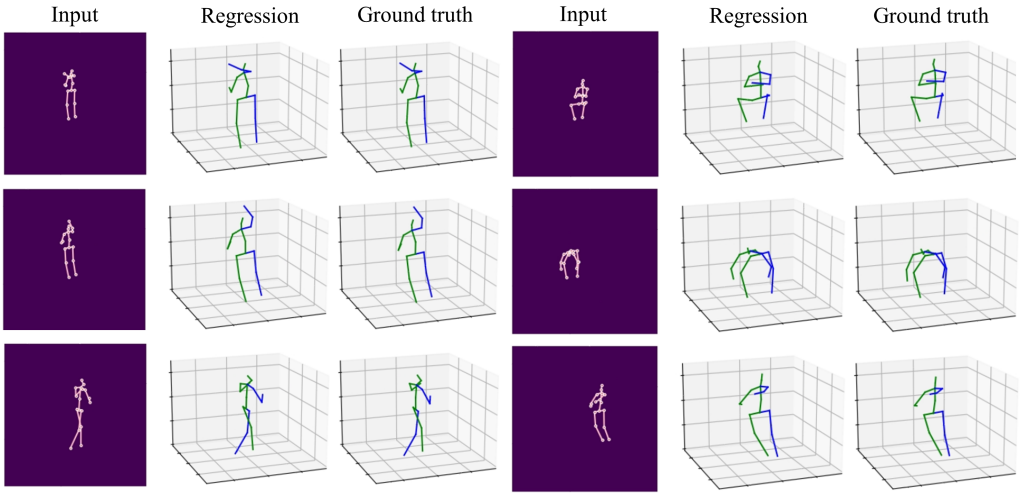

Figure 8. Visualization results of actions such as Directions, WalkD, Phone, Greet, and Sitting.

Table 3. Results on MPI-INF-3DHP [12] test set, and the best-performing data have been bolded.

Methods	Training Data	PCK				AUC
		GS	noGS	Outdoor	Avg	All
Martinez [37]	H36M	49.8	42.5	31.2	42.5	17.0
Mehta [12]	H36M	70.8	62.3	58.8	64.7	31.7
Yang [14]	H36M + MPII	-	-	-	69.0	32.0
Zhou [15]	H36M + MPII	71.1	64.7	72.7	69.2	32.5
Luo [60]	H36M	71.3	59.4	65.7	65.6	33.2
Ci [25]	H36M	74.8	70.8	77.3	74.0	36.7
Zhou [61]	H36M + MPII	75.6	71.3	**80.3**	75.3	38.0
Xu [26]	H36M	**81.5**	**81.7**	75.2	**80.1**	**45.8**
Zhao [28]	H36M	80.1	77.9	74.1	79.0	43.8
ours	H36M	80.3	77.6	74.0	79.2	43.9

4.3. Ablation Experiments

In this section, we conduct ablation experiments to investigate the impact of different graph convolution layers and determine the most optimal structure for the model.

Impact of Different Graph Convolutions: Zhao et al. [28] initially employed a self-attention layer to capture global features, followed by the utilization of an LAM (learnable adjacency matrix) graph convolution layer to acquire information about adjacent

joints. They subsequently applied a Chebyshev graph convolution layer to learn implicit higher-order connectivity relationships between non-adjacent joints. While the LAM graph convolution enables the network to learn valuable information between neighboring nodes using the Laplacian operator, it may introduce instability during the learning process, potentially disrupting human kinematic structural constraints and weakening local connections between neighboring nodes. In this study, we replace the LAM graph convolution with semantic graph convolution (SemGconv) after extracting global features with self-attention. The semantic graph convolution layer provides stability to the results by using a fixed adjacency matrix. The visualization of the adjacency matrices used for semantic graph convolution and LAM graph convolution is presented in Figures 9 and 10, respectively. Brighter matrices indicate stronger associations between two nodes. In LAM, the network learns the adjacency matrix, as shown in Figure 9, capturing more distant neighborhood relationships. However, this fragile relationship comes at the cost of compromising the fundamental human kinematic structure. Semantic graph convolution employs a fixed adjacency matrix to learn node features after global feature extraction, reinforcing local adjacency relationships based on fixed human structural constraints. While this increases the number of parameters, it provides a more reasonable basis for the subsequent Chebyshev graph convolution to learn distant neighborhood relationships between nodes.

We introduced modifications to the adjacency matrix, which serves as a prior constraint for Chebyshev graph convolution, by incorporating second-order adjacency information. Ablation experiments revealed that replacing the original Chebyshev graph convolution with this new version alone did not lead to improvements in test results or affect the performance of the original model. However, when combined with the improved Chebyshev graph convolution and semantic graph convolution, the prediction error decreased once again. This observation suggests that the LAM (learnable adjacency matrix) graph convolution may compromise body structure information and impact subsequent feature extraction. The experimental results, as presented in Table 4, utilize the 2D node information provided by the CPN (cascaded pyramid network) [21]. In cases where the authors did not provide a pre-trained model, the data in the table was reproduced using the configuration and parameters outlined in the original literature. Without modifying the prior constraint of ChebGConv, the result of the model with semantic graph convolution (SGraAttention) decreased from 52.0 mm to 51.3 mm, a reduction of almost one point compared to the original LAM graph convolution model. When the ChebGConv prior constraint was modified (AcChebGConv), the error values further decreased with the use of semantic graph convolution, decreasing from 51.3 mm to 51.1 mm. This indicates that the new Chebyshev blocks are effective. The ablation experiments presented in Table 4 demonstrate that strengthening stable local inter-joint adjacency relationships leads to the extraction of more valuable and practical features.

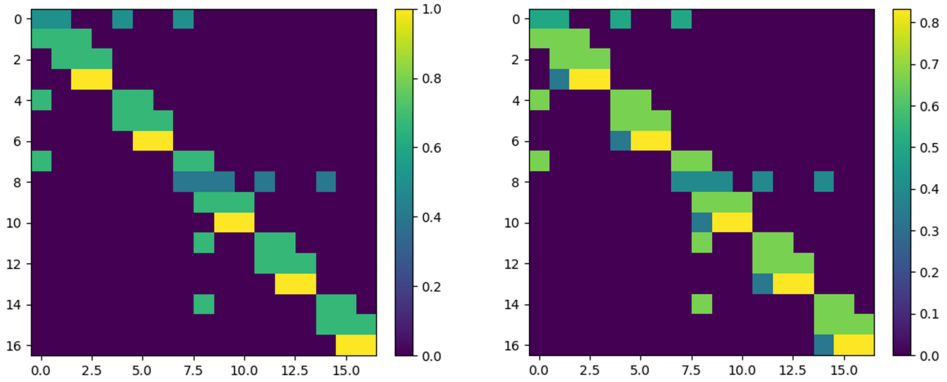

Figure 9. Visualization of the adjacency matrix used by the SCGormer, with the semantic graph convolution using a first-order adjacency matrix (first from the left) and the Chebyshev graph convolution in the AcChebGConv block using an adjacency matrix that combines first- and second-order adjacency information (first from the right).

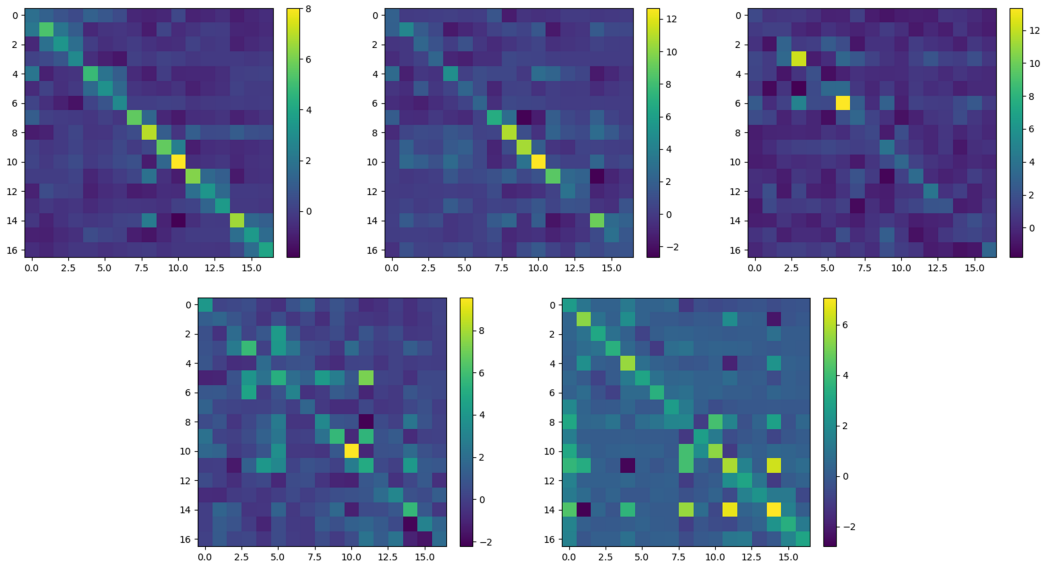

Figure 10. Visualization of the learnable adjacency matrix used in the five-layer LAM graph convolution, with the lowercase letters indicating the results for visualization of the corresponding layer.

Table 4. MPJPE results (in mm) for different graph convolution structures in the model on the Human3.6M [30,31] dataset.

Method	MPJPE
LAM-GConv × 2 ChebGConv Block(ChebGConv × 2)	52.0
SemGConv × 2 ChebGConv Block(ChebGConv × 2)	51.3
LAM-GConv × 2 AcChebGConv Block(AcChebGConv × 2)	52.0
SemGConv × 2 AcChebGConv Block(AcChebGConv × 2)	51.1

In the case of modified prior constraint information, we conducted further discussions regarding the number of SemGconv (semantic graph convolution) and AcChebGconv (Chebyshev graph convolution with fused adjacency matrix information) modules to determine the optimal configuration for achieving the best results. We set up 12 groups of modules with varying numbers but identical structures to perform ablation experiments. As shown in Table 5, when the number of one type of graph convolution module is fixed while the other is changed, it is evident that having two semantic graph convolution modules and four new Chebyshev graph convolution modules produces the best results. In comparison to the state-of-the-art method [28], our approach achieves a reduction in error by one and a half points and outperforms the competition in multiple actions.

Table 5. The test results (in mm) of MPJPE on Human3.6M [30,31] for different structures in the model. SemG represents SemGConv, A-C represents AcChebGConv.

MPJPE	A-C × 2	A-C × 3	A-C × 4	A-C × 5
SemG × 2	51.2	51.1	50.4	51.1
SemG × 3	51.3	50.9	50.5	51.8
SemG × 4	51.4	50.7	51.0	-
SemG × 5	-	51.3	-	-

For actions involving occlusion, leveraging the full utilization of 2D joint relations can significantly enhance the accuracy of 3D joint predictions. In Figure 11, we selected actions

like Phone, Photo, and WalkD for visualization. The human skeleton predicted by GraFomer, particularly the arms, deviates significantly from the ground truth. In contrast, our results are almost in perfect alignment with the ground truth. These visualizations provide a more intuitive demonstration of how employing suitable constraints to extract useful 2D features can accurately predict 3D information and mitigate issues caused by occlusion.

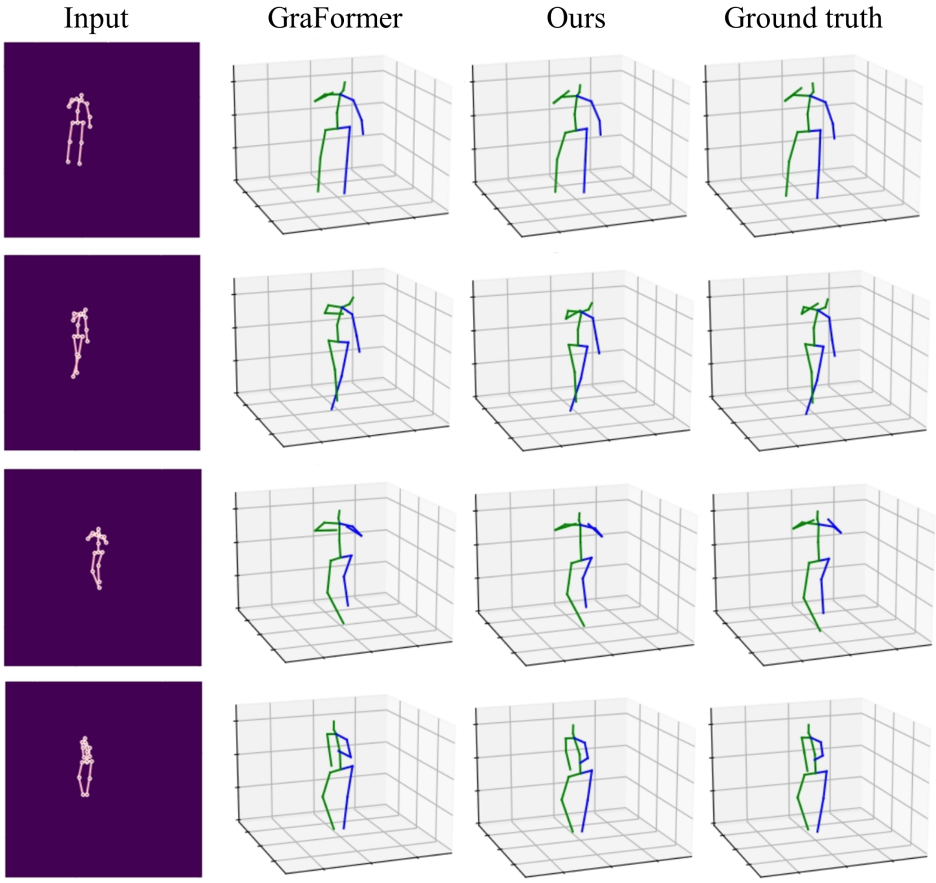

Figure 11. Visual comparison of some occluded frames between GraFormer [28] and our model in the test set of Human3.6M [30,31].

4.4. Analysis of Computational Complexity

A comparison of the complexity between our model and GraFormer is shown in Table 6. Compared to GraFormer, SCGraFormer exhibits an increase in both parameter count and computational complexity. Specifically, SCGraFormer has a parameter count of 1.24 million, which is 0.28 million more than GraFormer. In terms of computational complexity, SCGraFormer achieves a floating-point operations per second (FLOPs) rate of 1.53 billion, surpassing GraFormer by 0.41 billion. Despite these heightened metrics, our approach ultimately achieves an error value nearly 1.5 points lower than GraFormer.

Table 6. SGraFormer complexity comparison.

Methods	MPJPE	Params	FLOPs
GraFormer	51.8	0.96 M	1.12 G
SCGraFormer	50.4	1.24 M	1.53 G

5. Discussion

Through experimental validation, SCGraFormer has demonstrated promising results. In comparison to Graformer [28], we excluded LAM-GConv and employed SemGConv as a subsequent network layer for Transformer. As indicated by the ablation experiments in

Table 4, the use of a learnable adjacency matrix in LAM-GConv (as shown in Figure 10) leads to larger errors compared to the fixed adjacency matrix used in SemGConv (as shown on the left side of Figure 9). Subsequently, we refined the adjacency matrix of ChebGConv and named this network layer AcChebGConv. The error was further reduced, with a decrease of approximately one and a half points compared to GraFormer.

However, our work has limitations: the model's performance suffers when capturing rapid and significant movements, as illustrated in Figure 12. For instance, when a person rapidly raises their arm from bottom to top (top side of the figure), our model's prediction places the arm too high compared to the ground truth. In another scenario, when a person transitions from running to a sudden stop (bottom side of the figure), the model exhibits errors at the ends of the arms compared to the ground truth. These actions involve rapid and wide-ranging motions occurring in a short duration. The red circles highlight these noticeable errors. In future research, exploring more refined human body structural constraints to capture critical information about extremities (e.g., hand and foot joints) could be a promising avenue for improvement.

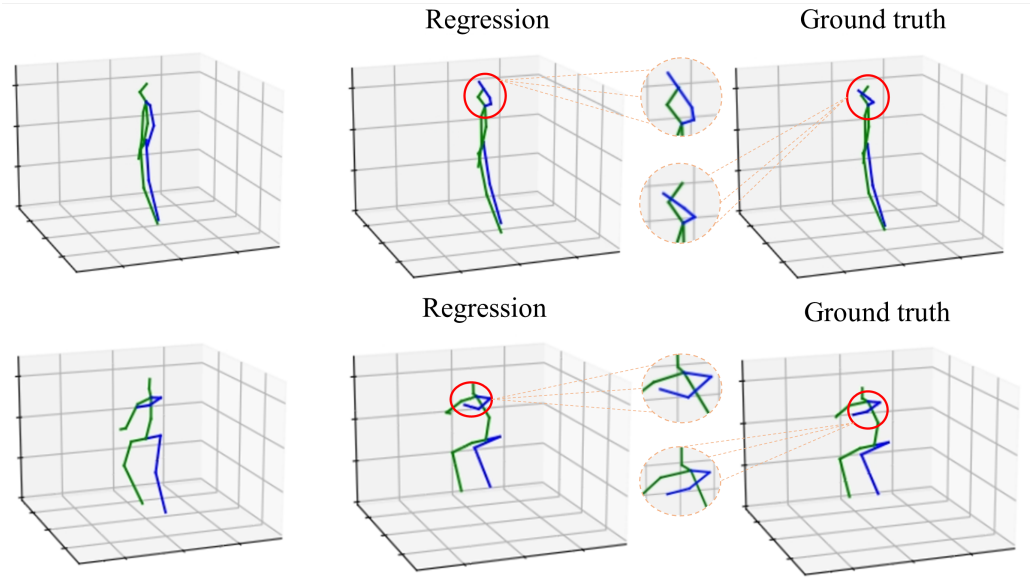

Figure 12. Limitations of our work.

In terms of application, 3D human pose estimation has broad prospects in the field of action recognition. It can provide detailed joint positions and motion information, providing a strong foundation for motion analysis and recognition. This is of great significance for understanding human actions in different scenarios, such as sports competitions, medical research, and interactive actions in virtual reality applications.

6. Conclusions

We propose a novel model architecture, SCGFormer, which significantly enhances the local feature extraction capabilities of graph convolution. The improved Chebyshev graph convolution enables the learning of connections between more distant joints, thereby maximizing the utilization of 2D human body pose information and enhancing the accuracy of 3D human pose estimation. Our results on widely used datasets outperform state-of-the-art methods based on graph convolution.

Nevertheless, our work still has limitations. When there is a short period of large-amplitude movement, leaf nodes such as hands will have significant differences from the ground truth. In the future, exploring more reasonable human skeleton constraints or obtaining more information from multiple perspectives will help to address the limitations of our work.

Author Contributions: Methodology, J.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and M.Y.; supervision, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Natural Science Foundation of China (No.61762007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Staudemeyer, R.C.; Morris, E.R. Understanding LSTM—A tutorial into long short-term memory recurrent neural networks. *arXiv* **2019**, arXiv:1909.09586
2. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556
3. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral networks and locally connected networks on graphs. *arXiv* **2013**, arXiv:1312.6203.
4. Zhang, Z.; Cui, P.; Zhu, W. Deep learning on graphs: A survey. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 249–270. [CrossRef]
5. Gan, J.; Wang, W. In-air handwritten English word recognition using attention recurrent translator. *Neural Comput. Appl.* **2019**, *31*, 3155–3172. [CrossRef]
6. Lu, D.; Luo, L. Fmkit: An in-air-handwriting analysis library and data repository. In Proceedings of the CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Virtual, 14–19 June 2020.
7. Weng, J.; Weng, C.; Yuan, J. Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4171–4180.
8. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
9. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
10. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 3413–3423.
11. Harada, T.; Sato, T.; Mori, T. Pressure distribution image based human motion tracking system using skeleton and surface integration model. In Proceedings of the 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164), Seoul, Republic of Korea, 21–26 May 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 4, pp. 3201–3207.
12. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3d human pose estimation in the wild using improved cnn supervision. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 506–516.
13. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7025–7034.
14. Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X. 3d human pose estimation in the wild by adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5255–5264.
15. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 398–407.
16. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 1954–1963.
17. Tekin, B.; Katircioglu, I.; Salzmann, M.; Lepetit, V.; Fua, P. Structured prediction of 3d human pose with deep neural networks. *arXiv* **2016**, arXiv:1605.05180.
18. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional human pose regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2602–2611.
19. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
20. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part VIII 14; Springer: New York, NY, USA, 2016; pp. 483–499.

21. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
22. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
23. Osokin, D. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. *arXiv* **2018**, arXiv:1811.12004.
24. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
25. Ci, H.; Wang, C.; Ma, X.; Wang, Y. Optimizing network structure for 3d human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2262–2271.
26. Xu, T.; Takano, W. Graph stacked hourglass networks for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 16105–16114.
27. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic graph convolutional networks for 3d human pose regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3425–3435.
28. Zhao, W.; Wang, W.; Tian, Y. GraFormer: Graph-oriented transformer for 3D pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 20438–20447.
29. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
30. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
31. Ionescu, C.; Li, F.; Sminchisescu, C. Latent structured models for human pose estimation. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 2220–2227.
32. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Comput. Surv.* **2022**, *55*, 1–41. [CrossRef]
33. Sarafianos, N.; Boteanu, B.; Ionescu, B.; Kakadiaris, I.A. 3d human pose estimation: A review of the literature and analysis of covariates. *Comput. Vis. Image Underst.* **2016**, *152*, 1–20. [CrossRef]
34. Agarwal, A.; Triggs, B. Recovering 3D human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *28*, 44–58. [CrossRef] [PubMed]
35. Ohashi, T.; Ikegami, Y.; Yamamoto, K.; Takano, W.; Nakamura, Y. Video motion capture from the part confidence maps of multi-camera images by spatiotemporal filtering using the human skeletal model. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4226–4231.
36. Takano, W.; Nakamura, Y. Action database for categorizing and inferring human poses from video sequences. *Robot. Auton. Syst.* **2015**, *70*, 116–125. [CrossRef]
37. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
38. Wandt, B.; Rosenhahn, B. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7782–7791.
39. Liu, J.; Rojas, J.; Li, Y.; Liang, Z.; Guan, Y.; Xi, N.; Zhu, H. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Madrid, Spain, 1–5 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3374–3380.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
42. Gardner, M.W.; Dorling, S. Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. *Atmos. Environ.* **1998**, *32*, 2627–2636. [CrossRef]
43. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 5–8 December 2016; Volume 29.
44. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903
45. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3d human pose estimation with spatial and temporal transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtually, 11–17 October 2021; pp. 11656–11665.
46. Li, W.; Liu, H.; Tang, H.; Wang, P.; Van Gool, L. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 13147–13156.

47. Goodman, J. Classes for fast maximum entropy training. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, pp. 561–564.
48. Mikolov, T.; Kombrink, S.; Burget, L.; Černockỳ, J.; Khudanpur, S. Extensions of recurrent neural network language model. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, UT, USA, 7–11 May 2001; IEEE: Piscataway, NJ, USA, 2011; pp. 5528–5531.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
51. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
52. Zhao, W.; Tian, Y.; Ye, Q.; Jiao, J.; Wang, W. Graformer: Graph convolution transformer for 3d pose estimation. *arXiv* **2021**, arXiv:2109.08364.
53. Liu, K.; Ding, R.; Zou, Z.; Wang, L.; Tang, W. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part X 16; Springer: New York, NY, USA, 2020; pp. 318–334.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal depth supervision for 3d human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7307–7316.
56. Sharma, S.; Varigonda, P.T.; Bindal, P.; Sharma, A.; Jain, A. Monocular 3d human pose estimation by generation and ordinal ranking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2325–2334.
57. Fang, H.S.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.C. Learning pose grammar to encode human body configuration for 3d pose estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
58. Hossain, M.R.I.; Little, J.J. Exploiting temporal information for 3d human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 68–84.
59. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
60. Luo, C.; Chu, X.; Yuille, A. Orinet: A fully convolutional network for 3d human pose estimation. *arXiv* **2018**, arXiv:1811.04989.
61. Zhou, K.; Han, X.; Jiang, N.; Jia, K.; Lu, J. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2344–2353.