Signal, Image and Video Processing (2024) 18:3627–3641
https://doi.org/10.1007/s11760-024-03028-0

ORIGINAL PAPER

PVA-GCN: point-voxel absorbing graph convolutional network for 3D human pose estimation from monocular video

Minghao Liu[1] · **Wenshan Wang**[1] · **Wei Zhao**[1]

Received: 17 November 2023 / Revised: 26 December 2023 / Accepted: 15 January 2024 / Published online: 16 February 2024

Abstract
The evolution of 3D human pose estimation techniques has seen substantial progress over the past few decades, with notable advancements in accuracy and applications. While recent research primarily aims at enhancing estimated pose performance, it is important to acknowledge the challenges encountered when evaluating these estimations against ground truth pose data. Our findings emphasize the pivotal role of refining 2D pose data or integrating advanced 2D pose detectors in elevating the quality of estimated pose data. For instance, refining the accuracy of 2D pose data positively correlates with the precision of the final estimated 3D pose. To streamline computational complexity, techniques like OctreeGrid filtering and VoxelGraph construction are employed. OctreeGrid filtering involves organizing data in a hierarchical octree structure, facilitating the extraction of essential joint points and voxel representations. VoxelGraphs focus on capturing spatiotemporal relationships within point clouds and voxels, enhancing the model's understanding of 3D spatial configurations. Our model, PVA-GCN, underwent extensive evaluation on benchmark datasets including Human3.6M, HumanEva-I, and MPI-INF-3DHP, surpassing existing state-of-the-art methods. These validations indicate the model's robustness across diverse datasets and scenarios, contributing significantly to advancing 3D human pose estimation. This research significantly contributes to the advancement of 3D human pose estimation by leveraging ground truth data to enhance pose estimation quality, thereby laying a foundation for future developments in the field.

Keywords 3D human pose estimation · Voxel grid · Point cloud · Absorbing graph convolutional network

1 Introduction

The prediction of human body pose joints in three-dimensional space, known as 3D human pose estimation (HPE) in video content, serves various applications such as video surveillance, human–robot interaction, and physiotherapy [1]. Utilizing advanced motion sensors like motion capture systems, depth sensors, or stereoscopic cameras [2, 3] enables the direct extraction of 3D human poses. This task can be undertaken in either multi-view setups, involving multiple cameras, or monocular settings, where a single camera is used. Despite the generally superior performance, based on specific criteria, of state-of-the-art multi-view methods [4–7] compared to monocular ones, the cost-effectiveness and wide application of ordinary RGB monocular cameras in real-world surveillance scenarios make 3D HPE from monocular videos an essential and challenging task, attracting increasing research interest, particularly in areas such as feature extraction or real-time processing. Recent works in the monocular view domain can be categorized into model-based and model-free methods [8]. Model-based methods [9, 10] incorporate parametric body models such as kinematic [11], planar [12], and volumetric models [13] for 3D HPE. On the other hand, model-free methods can be further divided into single-stage and 2D to 3D lifting methods. Single-stage methods directly estimate the 3D pose from images in an end-to-end manner [14–19]. 2D to 3D lifting methods introduce an intermediate 2D pose estimation layer [20–23]. Notably, 2D to 3D lifting methods, particularly when implemented with ground truth 2D poses, demonstrate improved performance in terms of accuracy or robustness.

Advancements in the accuracy and efficiency of 2D human pose detection, achieved through detectors like Mask R-CNN (MRCNN) [24], cascaded pyramid network (CPN)

Minghao Liu and Wenshan Wang have contributed equally to this work.

✉ Wei Zhao
zhaowei1982122@126.com

[1] Faculty of Science, Dalian Minzu University, No. 31, Jinshi Road, Jinshitan, Dalian 116000, Liaoning Province, China

[25], stacked hourglass (SH) detector [26], and HR-Net [27], are notable. The intermediate 2D pose estimation stage, facilitated by these detectors, plays a pivotal role in significantly reducing the data volume and complexity associated with the 3D HPE task.

Regarding temporal information, mainstream methods [21–23, 28–30] have witnessed substantial improvements in accuracy and efficiency by processing extended sequences of 2D pose frames, contributing to the advancement of 2D to 3D lifting methods. Among them, [30] stands out for achieving state-of-the-art performance using ground truth 2D poses. Recent approaches [30, 31] have streamlined the process by fine-tuning these 2D pose detectors on target datasets, resulting in notable enhancements in the accuracy and efficiency of estimated 2D pose data. Despite these advancements, the performance still falls short in certain aspects compared to results obtained using ground truth 2D pose. This observation prompts a focused effort on enhancing specific aspects of 3D HPE, such as accuracy or robustness, through the utilization of ground truth 2D pose data, anticipating potential improvements with future, higher-quality estimated 2D pose data.

Motivated by the promising performance and advantages of 2D to 3D lifting methods, our work adds to the existing literature in this domain. Recent advanced models in various 2D to 3D lifting approaches, as categorized by [20]'s introduction of the fully connected network (FCN), fall into three main groups: temporal convolutional network (TCN)-based [21, 22], graph convolutional network (GCN)-based [23, 28, 32], and transformer-based models [29, 30, 33]. It is noteworthy that existing TCN- and transformer-based methods exhibit the capability to handle large receptive fields, allowing for the representation of extended 2D pose sequences, through the utilization of strided convolutions, providing enhanced context or spatial information. However, a significant challenge arises in designing intuitive methods to backtrace local joint features based on the pose structure, especially given that the 2D pose sequence is flattened and fed to the model, necessitating innovative solutions for effective feature extraction. Additionally, these methods rely on the same fully connected layer for estimating different pose joints, potentially neglecting the unique and independent characteristics of distinct pose joints, which could impact the accuracy of joint predictions. Conversely, GCN-based models explicitly preserve the structure of 2D and 3D human poses during convolutional propagation. Yet, the potential advantages of GCN in this context remain underexplored, and further exploration could unveil novel insights or improvements in 2D to 3D lifting methods. Existing GCN-based methods [23, 32] also employ a fully connected layer for estimating different 3D pose joints, potentially overlooking the structural features of GCN representations.

In pursuit of the stated objective, we introduce an innovative and effective framework for 3D human pose estimation, utilizing a dual absorbing graph representation strategy. The initial step involves downsampling the input dense event stream into a sparse event stream and dividing it into non-overlapping voxel grids. Subsequently, distinct dual absorbing graph models are crafted for the point and voxel streams, each encompassing all sparse point/voxel nodes and a dedicated absorbing node. The subsequent phase introduces a novel absorbing graph convolutional network (AGCN) designed for absorbing graph representation and learning in the context of 3D human pose estimation. The AGCN model provides three distinct advantages. Firstly, it adeptly captures the importance of event nodes in learning the graph-level representation through the introduced absorbing node. Secondly, the AGCN's absorbing node dynamically aggregates information from all event nodes, improving the summarization of node representations compared to conventional pooling layers. Thirdly, in AGCN, each node aggregates messages from both its neighbors and the absorbing node, concurrently preserving local and global structures to enhance the learning of graph representation. Finally, the outputs of the dual AGCN branches are concatenated to extract complementary information from both streams. This combined information is then fed into a linear layer for accurate 3D human pose estimation.

In summary, the primary innovations presented in this work are as follows:

- Employing OctreeGrid filtering and voxel construction streamlines computational complexity, extracting vital joint points and voxel representations through downsampling for effective 3D pose estimation.
- Introducing CPointGraph and VoxelGraph, our absorbing graphs focus on specific spatiotemporal relationships in point clouds and voxels. The innovative absorbing graph convolutional network (AGCN) utilizes graph convolutional networks (GCNs) to learn feature descriptors crucial for accurate 3D pose estimation.
- AGCN's multilayer design with a residual connection facilitates seamless information flow, enhancing the model's understanding of intricate spatiotemporal structures. Absorbing nodes play a pivotal role in consolidating information for improved graph-level representations.

2 Related work

2.1 2D to 3D lifting

Early attempts to infer 3D positions from 2D projections, like [34–36], often relied on manually chosen parameters based

on assumptions about joint mobility. While methods such as [10, 37] have made impressive strides in estimating 3D pose from fewer frames, they sometimes neglect the consideration of temporal information evolution over the sequence. Recent advancements in 2D human pose estimation, exemplified by [25, 26, 38], have paved the way for 2D to 3D lifting approaches, which have outperformed other methods. Building upon the principles of [20], more sophisticated learning architectures have emerged, with a particular emphasis on utilizing temporal information. These approaches, collectively known as 2D to 3D lifting, can be categorized into three directions: TCN-based, GCN-based, and transformer-based architectures [21–23, 28–30, 32].

TCN-based methods, exemplified by [21, 22], have significantly advanced the field of 2D to 3D pose lifting, particularly through their effective handling of temporal sequences and dimensionality reduction. This design enables the features to undergo a reduction in dimensionality, effectively transforming a 2D pose sequence into a feature embedding for 3D pose estimation through a final fully connected layer. The fully connected layer typically has 1024 channels and is used to predict the 3D positions of all pose joints. Research has extensively explored various numbers of input 2D pose frames, revealing that a reasonable number of frames benefits the 3D pose reconstruction. The strided design efficiently reduces the feature size by decreasing the number of temporal frames during the propagation of several TCN blocks, contributing to enhanced computational efficiency and improved real-time processing. Building upon this strided structure, transformer-based methods, particularly [30], have exhibited promising performance. [30] capitalizes on weighted and temporal loss functions, surpassing GCN-based methods optimized with an additional motion loss [23, 32]. Notably, the effectiveness of the motion loss was found to be limited in [30]. These observations prompt the exploration of effective models in the realm of GCN-based models. The aim is to incorporate the inspiring designs from the TCN-based methods without relying on novel loss functions. This research direction seeks to strike a balance between innovation and simplicity in architectural design, aiming for a model that is both advanced and easily interpretable.

2.2 Graph convolutional network

A widely used approach for representing pose data is the Spatial Temporal GCN (ST-GCN), initially designed to model large receptive fields for improved skeleton-based action recognition. Building on this, more sophisticated GCN models like [23, 32, 39, 40] have emerged to further advance 3D human pose estimation (3D HPE). These models aim to enhance the understanding and accuracy of 3D pose estimation, leveraging the principles established by ST-GCN.

In the realm of graph convolutional network (GCN)-based models dedicated to 3D human pose estimation (3D HPE), several innovative architectures have emerged in recent years. Ci et al. [39] introduced the locally connected network (LCN), which combines ideas from both fully connected networks and GCNs. Specifically, LCN performs graph convolutions over a neighbor set defined by distance, similar to the design in ST-GCN [41]. Zhao et al. [32] presented SemGCN, a novel model that stacks GCN layers followed by a flattening fully connected layer. By optimizing with both joint positions and bone vectors, SemGCN achieves strong performance on 3D HPE. Choi et al. [40] offered a new perspective by utilizing GCN to lift 2D poses to 3D, demonstrating its effectiveness in recovering 3D human poses and meshes. Liu et al. [42] investigated different weight sharing schemes in GCNs for the pose lifting task and identified the superiority of the pre-aggregation scheme in terms of performance. The architecture proposed in [42] shares similarities with SemGCN. The aforementioned GCN-based approaches have exhibited compelling results given single-frame 2D poses as input. However, they did not fully take advantage of the temporal information available in 2D pose sequences. This reveals opportunities for future investigation into modeling the temporal dynamics to further enhance the performance of GCN-based 3D human pose estimation.

The U-shaped graph convolution networks (UGCN), as seen in [23, 28], represent a significant advancement in GCN-based methods for 3D human pose estimation (3D HPE). UGCN excels by considering the temporal characteristics of pose motion, specifically addressing the reconstruction of a single 3D pose frame from multiple 2D pose frames. UGCN leverages the spatial–temporal GCN [41] to predict a 3D pose sequence from a 2D pose sequence, regulating the temporal trajectory of pose joints with a motion loss term based on the prediction and the corresponding ground truth 3D pose sequence. While prior works like SemGCN and UGCN have made improvements by introducing novel loss terms, our contribution to the literature of 2D-3D lifting focuses on the application of a consistent loss term, inspired by the proven effectiveness of [21, 22]. In our model design, we propose to incorporate strided convolutions into a GCN-based model to represent the global information of a 2D pose sequence. Leveraging the structure of GCN representation, we explicitly employ the structured features of different pose joints to locally predict their corresponding 3D pose locations. This approach builds upon the existing literature while enhancing the accuracy and robustness of 3D HPE.

3 Method

In this section, we first give an overview of our proposed 3D human pose estimation model and initial event repre-

sentation. Then, we dive into details of the absorbing graph representation learning method, focusing on graph construction and absorbing graph convolutional networks (AGCN).

3.1 Overview

For an input video stream containing hundreds of thousands of events, we initially employ OctreeGrid filtering [43] and Voxel construction. This process aims to derive point cloud and voxel representations, respectively. Next, we construct two absorbing graphs, namely CPointGraph and VoxelGraph, to capture the spatiotemporal relationships between the point clouds and voxels. Subsequently, we introduce a novel absorbing graph convolutional network (AGCN) designed to learn feature descriptors from the information captured by the two graphs. In the final step, we integrate the information from the two graphical representations to perform 3D human pose estimation. Figure 1 outlines the overall framework, with details of each module provided in the following sections.

3.2 Initial event representation

Within the realm of 3D pose estimation, where the challenges of dealing with substantial data volumes and computational complexity are evident, the use of downsampling techniques is of utmost importance to effectively curtail the number of events. In this paper, we employ two distinct sampling methods to obtain concise event representations, seamlessly integrating them with the 3D pose estimation process. The first crucial step involves the extraction of representative joint points, which are explicitly designated as center points.

For a more detailed elaboration, let us focus on the original event stream, denoted as $\mathcal{E}$, which encompasses N events. Our initial phase involves the application of the OctreeGrid filtering algorithm [43], a pivotal step in efficiently reducing the data. This results in the extraction of a set of representative events, specifically designated as center points, which we refer to as $\mathcal{C} = \{c_1, c_2 \dots c_M\}$. Each of these representative events, denoted as ci, is encapsulated within a 4D tuple, presented as follows:

$$c_i = (x_i, y_i, t_i, p_i). \tag{1}$$

The variables x_i and y_i are employed to represent spatial coordinates, while t_i corresponds to the event's timestamp. Furthermore, the variable p_i denotes the event's attribute or polarity. In the context of our 3D pose estimation research, our primary emphasis is placed on (x_i, y_i, t_i), which collectively characterizes the spatiotemporal coordinates or positions of an event. It is noteworthy that the sampled set $\mathcal{C}$, in contrast to the original events in set $\mathcal{E}$, contains a substantially reduced number of events, yet it effectively preserves the fundamental spatiotemporal structure.

Our approach not only involves identifying center points as $\mathcal{C}$ but also incorporates voxelization to obtain voxel representations. More specifically, when considering the original event stream $\mathcal{E}$ within a spatial–temporal 3D space defined by dimensions H, W, and T, we partition this space into voxels, each having dimensions h', w', and t'. Consequently, each voxel typically encompasses multiple events, and the resulting event voxels within this spatial–temporal space are characterized by dimensions H/h', W/w', and T/t'. In practice, the aforementioned voxelization process typically still results in the generation of tens of thousands of voxels. To further reduce the number of voxels and mitigate the impact of noisy voxels in the context of human pose estimation, we also implement a voxel selection procedure. This procedure identifies the top K voxels based on the number of events contained within each voxel. Let $\mathcal{O} = \{o_1, o_2 \cdots o_K\}$ represent the collection of the final selected voxels. Each event voxel, denoted as o_i, is associated with a feature descriptor $\mathbf{a}_i \in \mathbb{R}^C$ that incorporates attributes, including polarity, from the events it encompasses. Consequently, each $o_i \in \mathcal{O}$ is represented as:

$$o_i = (x_i, y_i, t_i, \mathbf{a}_i), \tag{2}$$

where x_i,y_i,t_i represent the 3D coordinates of each voxel.

3.3 Absorbing graph representation learning

After obtaining the initial event representations from $\mathcal{C}$ and $\mathcal{O}$, we introduce an effective method to learn distinctive representations tailored for 3D human pose estimation tasks. These initial representations in $\mathcal{C}$ and $\mathcal{O}$ encapsulate crucial spatiotemporal relationships among event units, whether they are points or voxels. Recognizing the significance of these relationships, we leverage graph models and a learning approach to represent the pre-processed event streams associated with 3D human pose estimation.

In the upcoming sections, we delve into the specifics of our graph construction techniques for the data from $\mathcal{C}$ and $\mathcal{O}$, focusing on their relevance to 3D human pose estimation. Subsequently, we unveil a novel absorbing graph convolutional network (AGCN) designed to adeptly learn and generate effective representations for the event data originating from $\mathcal{C}$ and $\mathcal{O}$. This integrated approach is pivotal in improving the performance of our 3D human pose estimation tasks by capturing the inherent spatiotemporal relationships within the event data.

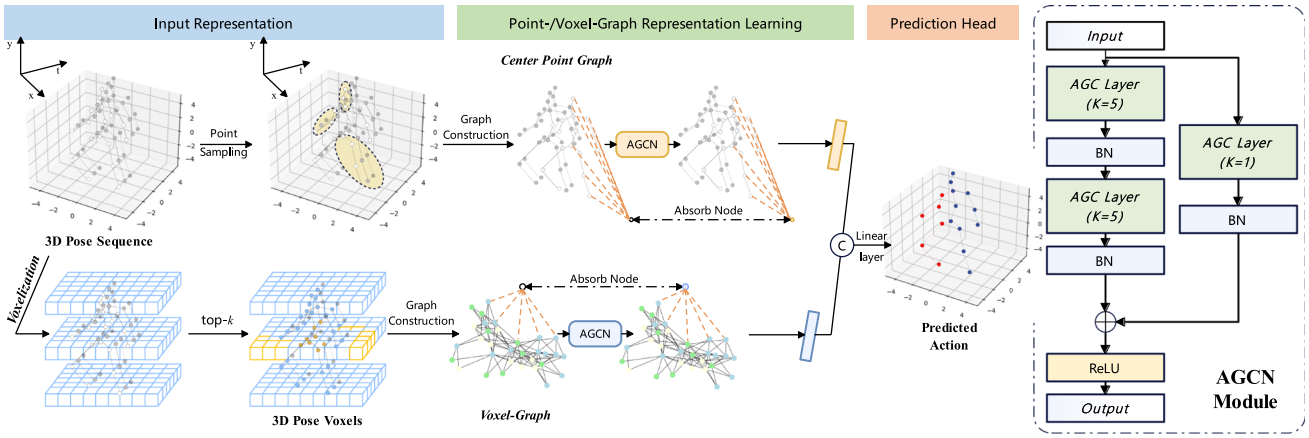

Fig. 1 A comprehensive overview of our proposed absorbing graph representation learning framework tailored for human pose estimation. In this context, we initially convert the input representing human poses into dual forms, namely the sparse pose cloud and voxelized representations. Subsequently, we establish dual graphs based on these two inputs, with each point or voxel-grid serving as a graph node. It is noteworthy that we incorporate absorbing nodes into the graph structure to capture global information crucial for accurate pose estimation. The absorbing graph convolutional network (AGCN) is meticulously crafted to specialize in structured feature learning and simultaneous global feature aggregation, aligning with the unique demands of human pose estimation. In the final stage, the predictions from the dual branches are concatenated and fed through a linear layer to yield the ultimate human pose estimation

3.3.1 Graph construction

The core innovation is the introduction of graph convolutional networks (GCNs) to model relationships between points and voxels in event streams.

Center points graph. For each center point event datum c_i in $\mathcal{C}$ with attributes (x_i, y_i, t_i, p_i), we add a node v_i to G^c. Nodes v_i and v_j are adjacent if the spatial distance between c_i and c_j is less than R. This geometric graph G^c with node set V^c and edge set E^c captures the relationships between nearby events.

$$d(c_i, c_j) < R \tag{3}$$

where R is a preset parameter. In our experiments, $d(c_i, c_j)$ calculates the distance between events c_i and c_j.

$$d(c_i, c_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (t_i - t_j)^2} \tag{4}$$

An absorbing node $\bar{v}$ is introduced and linked to every event node v_i in V^c by adding edges. This augmented center point graph is shown in Fig. 1.

Voxel graph. When addressing the topic of 3D pose estimation, we apply a similar approach to handle voxel event data, denoted as $\mathcal{O}$. Here, we establish a geometric neighborhood graph labeled as $G^{\text{pose}}(V^{\text{pose}}, E^{\text{pose}})$. In this context, each node v_i within V^{pose} represents a pose instance $p_i = (x_i, y_i, z_i, \hat{\mathbf{p}}_i)$ from the dataset $\mathcal{P}$, described by a feature vector $\mathbf{p}_i \in \mathbb{R}^D$. We define an edge $e_{ij} \in E^{\text{pose}}$ connecting v_i and v_j if the Euclidean distance between their 3D coordinates is less than a predefined threshold D_{lim}, as defined in Eq. (4). Moreover, we introduce an absorbing node v^*, which establishes connections with all pose nodes within V^{pose}. Its role is to aggregate and integrate information from all pose instances, facilitating the extraction of a comprehensive, global-level representation for the entire pose graph. For a visual representation of this pose estimation process, please refer to Fig. 1.

3.3.2 Absorbing GCN

Introducing the absorbing graph convolutional network (AGCN)—an inventive model crafted for the autonomous derivation of meaningful representations in the realm of 3D pose estimation. The inspiration for this innovative approach is rooted in the central point mentioned earlier and voxel graphs incorporating absorbing nodes. AGCN is structured with multiple learning layers, including a residual connection between the initial and final layers, as exemplified in Fig. 1 (right). Each layer plays a pivotal role in facilitating the seamless process of message passing across the graph. To delve further into this concept, within every AGCN layer, each event node v_i adeptly aggregates features from its neighboring nodes as

$$f'_d(v_i) \leftarrow \sigma\Big(\sum_{v \in \{\mathcal{N}(v_i) \bigcup \bar{v}\}} \omega_d(v_i, v) f(v)\Big), d = 1, 2 \ldots D \tag{5}$$

The absorbing node, denoted as $\tilde{v}$, collects and consolidates messages from all the remaining nodes in the following manner:

$$f'_d(\bar{v}) \leftarrow \sigma\Big(\sum_{v \in V} \omega_d(\bar{v}, v) f(v)\Big), d = 1, 2 \ldots D \tag{6}$$

The activation function, typically ReLU, is applied. We use $\omega_d(v_i, v)$ and $\omega_d(\bar{v}, v)$ to represent the flexible convolution kernel weights. Following prior research [44], [45], we define these weights as a Gaussian mixture model (GMM) function [45] based on the pseudocoordinate. In a general context, for any node pair (u, v), we calculate their $pseudocoordinate$[3], referred to as $z_{u,v}$, and then proceed to train the weight kernel $\omega_d(v_i, v)$.

$$\omega_d(u, v) = \sum_{k=1}^{K} \alpha_k \exp(-\frac{1}{2}(z_{uv} - \mu_k)^T \Sigma_k{}^{-\Psi}(z_{uv} - \mu_k)) \tag{7}$$

the parameters μ_k and Σ_k are adaptable and undergo a learning process, while α_k characterizes the importance assigned to the k-th Gaussian kernel. The total number of Gaussian kernels used is represented by K.

By implementing the layer-wise message passing method explained earlier, we can construct a multilayer AGCN architecture that integrates a residual connection, bridging the initial and final layers. This architectural design is illustrated in Fig. 1 (right).

To represent the results obtained from the two branches following the application of the AGCN module, we employ Y^c to denote the output of the first branch, with dimensions of Md, and Y^o for the second branch, which is of size Ld.

$$Y^c = \text{AGCN}(G^c, \Omega^c), Y^o = \text{AGCN}(G^o, \Omega^o) \tag{8}$$

In the realm of 3D pose estimation, we refer to Ω^c and Ω^o as encompassing all the parameters associated with the two branches.

3.4 Classification head and network training

We employ $Y^c_{\tilde{v}}$ and $Y^o_{\tilde{v}}$ to denote the representations associated with the absorbing nodes in both the center point and voxel graphs. As detailed in 3.3.2, the absorbing node's remarkable ability to consolidate information from all event nodes positions it as an excellent representation of the overall graph-level information. Consequently, we combine $Y^c_{\tilde{v}}$ and $Y^o_{\tilde{v}}$ and employ a MLP for the final classification, ultimately predicting the class label for 3D pose estimation.

$$Y = \text{MLP}(Y^c_{\tilde{v}} \| Y^o_{\tilde{v}}) \tag{9}$$

The || symbol signifies the concatenation operation in our approach. We also introduce dropout and batch-normalization layers between the layers of the MLP to mitigate possible challenges related to overfitting and gradient issues. The entire network is trained in a holistic end-to-end fashion. For the optimization of the complete network, we employ the negative log likelihood loss [46] as our chosen loss function in the context of 3D pose estimation.

4 Experiments

4.1 Datasets and evaluation

Our experiments are conducted on three widely used datasets in the field of 3D human pose estimation: Human3.6M, HumanEva-I, and MPI-INF-3DHP.

For Human3.6M, we utilize data from human subjects labeled as S1, S5, S6, S7, and S8 for training, aligning with established practices in the field [21–23, 32]. Data from subjects S9 and S11 are reserved for testing.

In the case of HumanEva-I, following the approach taken in [20] and [22], we use data for the "walk" and "jog" actions from subjects S1, S2, and S3 for both training and testing.

Regarding MPI-INF-3DHP, we adhere to the experimental settings outlined in the recent state-of-the-art work [54] to ensure a fair and rigorous comparison.

We employ standard evaluation protocols for our experiments, including Mean Per Joint Position Error (MPJPE) and pose-aligned MPJPE (P-MPJPE). MPJPE is calculated as the mean Euclidean distance between the predicted 3D pose joints and the ground truth 3D pose joints, aligned to the root joint. P-MPJPE incorporates additional post-processing steps, such as scale, rotation, and translation, to align the predicted 3D pose more rigidly with the ground truth. These evaluation metrics are consistent with previous studies [58–60]. This ensures that our results are comparable and meaningful within the field of 3D human pose estimation.

4.2 Implementation details

We provide a comprehensive overview of the implementation details for our PVA-GCN model, covering three primary aspects: 2D pose detections, model configuration, and training hyperparameters. For a fair and consistent comparison with prior works such as [21, 22], we adopt the 2D pose detections from Human3.6M and HumanEva-I. CPN's 2D pose detection involves 17 joints, and MRCNN's detection involves 15 joints, influencing the granularity of the pose information used in our experiments.

Our model's tunability provides flexibility for optimization, allowing us to explore and adjust various parameters for improved performance. Ablation studies on Human3.6M were conducted to systematically vary channels and pose frames (C_{out}, T) and assess their impact on the model's per-

Table 1 Protocol #1 evaluates the reconstruction error using Mean Per Joint Position Error (MPJPE) in millimeters on the Human3.6M dataset

Method	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez et al. [20] (ICCV'17)	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
Fang et al. [47] (AAAI'18)	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	60.4
Pavlakos et al. [48] (CVPR'18)	48.5	54.4	54.4	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
Lee et al. [49] (ECCV'18) †	40.2	49.2	47.8	52.6	50.1	75.0	50.2	43.0	55.8	73.9	54.1	55.6	58.2	43.3	43.3	52.8
Zhao et al. [32] (CVPR'19)	47.3	60.7	51.4	60.5	61.1	49.9	47.3	68.1	86.2	**55.0**	67.8	61.0	42.1	60.6	45.3	57.6
Ci et al. [39] (ICCV'19)	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
Pavllo et al. [21] (CVPR'19) †	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
Cai et al. [50] (ICCV'19) †	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
Xu et al. [11] (CVPR'20) †	**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 et al. [22] (CVPR'20) †	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
Zeng et al. [51] (ECCV'20) †	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
Xu and Takano [52] (CVPR'21)	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
Zhou et al. [53] (PAMI'21) †	38.5	45.8	40.3	54.9	**39.5**	**45.9**	**39.2**	43.1	**49.2**	71.1	**41.0**	53.6	44.5	33.2	34.1	45.1
Li et al. [33] (CVPR'22) †	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	**41.1**	43.8	29.8	30.6	43.0
Shan et al. [54] (ECCV'22) †	38.9	**42.7**	40.4	41.1	45.6	49.7	40.9	**39.9**	55.5	59.4	44.9	42.2	**42.7**	29.4	29.4	42.8
Our PCA-GCN (T=243, CPN) †	40.1	43.0	**39.5**	**40.4**	44.7	53.0	41.0	40.4	56.5	61.6	43.9	41.7	44.7	**28.2**	**28.7**	**41.7**
Martinez et al. [20] (ICCV'17)	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Lee et al. [49] (ECCV'18) †	32.1	36.6	34.3	37.8	44.5	49.9	40.9	36.2	44.1	45.6	35.3	35.9	30.3	37.6	35.5	38.4
Zhao et al. [32] (CVPR'19)	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 et al. [39] (ICCV'19)	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 et al. [22] (CVPR'20) †	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Xu and Takano [52] (CVPR'21)	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
Zheng et al. [53] (ICCV'21) †	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Li et al. [33] (CVPR'22) †	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
Shan et al. [54] (ECCV'22) †	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
Our PCA-GCN (T=243, GT) †	**25.3**	**26.0**	**28.0**	**24.2**	**27.0**	**30.5**	**28.3**	**25.7**	**36.5**	**38.5**	**28.7**	**25.8**	**26.1**	**19.3**	**19.6**	**27.3**
Wang et al. [23] (ECCV'20) †*	23.0	25.7	22.8	22.6	24.1	30.6	24.9	24.5	31.1	35.0	25.6	24.3	25.1	19.8	18.4	25.6
Li et al. [29] (TMM'22) †*	27.1	29.4	26.5	27.1	28.6	33.0	30.7	26.8	38.2	34.7	29.1	29.8	26.8	19.1	19.8	28.5
Hu et al. [28] (MM'22) †*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22.7
Zhang et al. [30] (CVPR'22) †*	21.6	22.0	20.4	21.0	**20.8**	**24.3**	24.7	21.9	26.9	**24.9**	21.2	21.5	20.8	14.7	15.7	21.6
Our PCV-GCN (T=243, GT) †*	**19.8**	**20.9**	**19.7**	**19.2**	21.1	26.2	**22.9**	**19.5**	**26.5**	28.9	**20.4**	**19.8**	**18.9**	**12.5**	**13.5**	**20.6**

The top table displays results where input 2D pose sequences are detected by the cascaded pyramid network (CPN). The bottom table showcases input 2D pose sequences with ground truth (GT). The optimal performance is denoted in bold, the second-best is underlined, and lower values are preferable. The symbol † signifies the utilization of temporal information, while * indicates the reconstruction of an intermediate 3D pose sequence

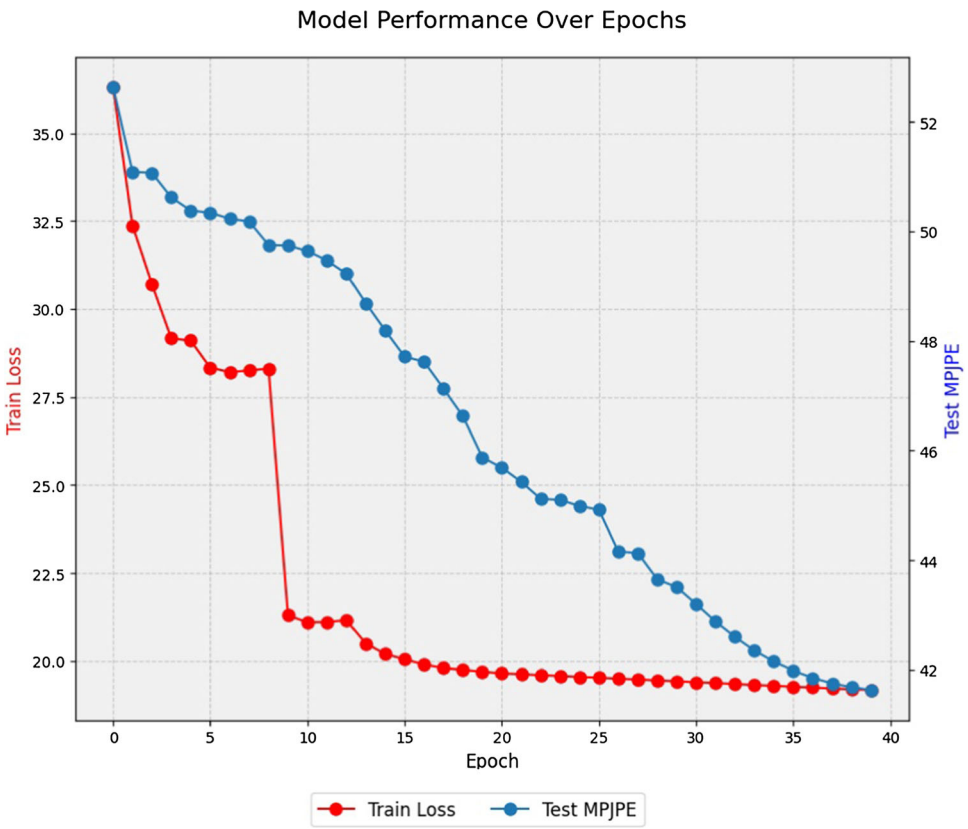

Fig. 2 Loss on the training set and MPJPE on the test set

formance. Experiments, conducted on four GeForce GTX 4060 GPUs, utilized batch sizes of 512, 256, and 256 for Human3.6M, HumanEva-I, and MPI-INF-3DHP, respectively. Leveraging sparse 3D supervision, our approach achieved state-of-the-art performance in lifting 2D to 3D poses, surpassing previous methods while requiring significantly fewer 3D labels.

4.3 Comparison with state-of-the-art

Table 1 and Fig. 3 present a detailed comparison between our PVA-GCN approach and state-of-the-art methods. These tables showcase results obtained on the Human3.6M and HumanEva-I datasets using *Protocol* #1 and *Protocol* #2, respectively. The optimization of our implementation, whether based on ground truth (GT) 2D pose with or without the loss for reconstructing the intermediate 3D pose sequence, significantly impacts the obtained results. This optimization strategy is crucial for understanding the trade-offs between model accuracy and computational efficiency, offering valuable insights into the robustness and generalizability of our approach across diverse datasets and scenarios.

Figure 2 provides insights into the training process of our PVA-GCN on the Human3.6M dataset, illustrating convergence dynamics and visually representing the evolution of the model's loss function over training epochs.

Results in Table 2 on the HumanEva-I dataset under *Protocol*#2 confirm the superiority of our method over state-of-the-art alternatives, particularly in reducing the MPJPE. Notably, this improvement is achieved solely through utilizing the MPJPE loss, highlighting the efficacy of our model in enhancing the accuracy of 3D human pose estimation.

To further evaluate our approach, we qualitatively compare it with a state-of-the-art method lacking a 3D pose sequence reconstruction module [22]. This comparison aims to highlight the nuanced improvements achieved by our model in capturing the intricacies of 3D human pose. Our visual analysis, focusing on specific instances like the "S11 WalkT." action, reveals that our method produces more accurate and coherent representations of joint movements compared to the alternative method. The absence of a 3D pose sequence reconstruction module in the compared approach becomes apparent in scenarios where capturing temporal dynamics is crucial for accurate pose estimation. This qualitative evaluation further substantiates the advantage of our model in effectively leveraging temporal information for superior 3D human pose estimation.

Previous research predominantly focuses on performance evaluation using estimated 2D pose data, such as CPN or HR-Net pose data. This emphasis has implications for benchmarking, where evaluation criteria may favor methods adept at handling lower-quality 2D pose data, potentially influenc-

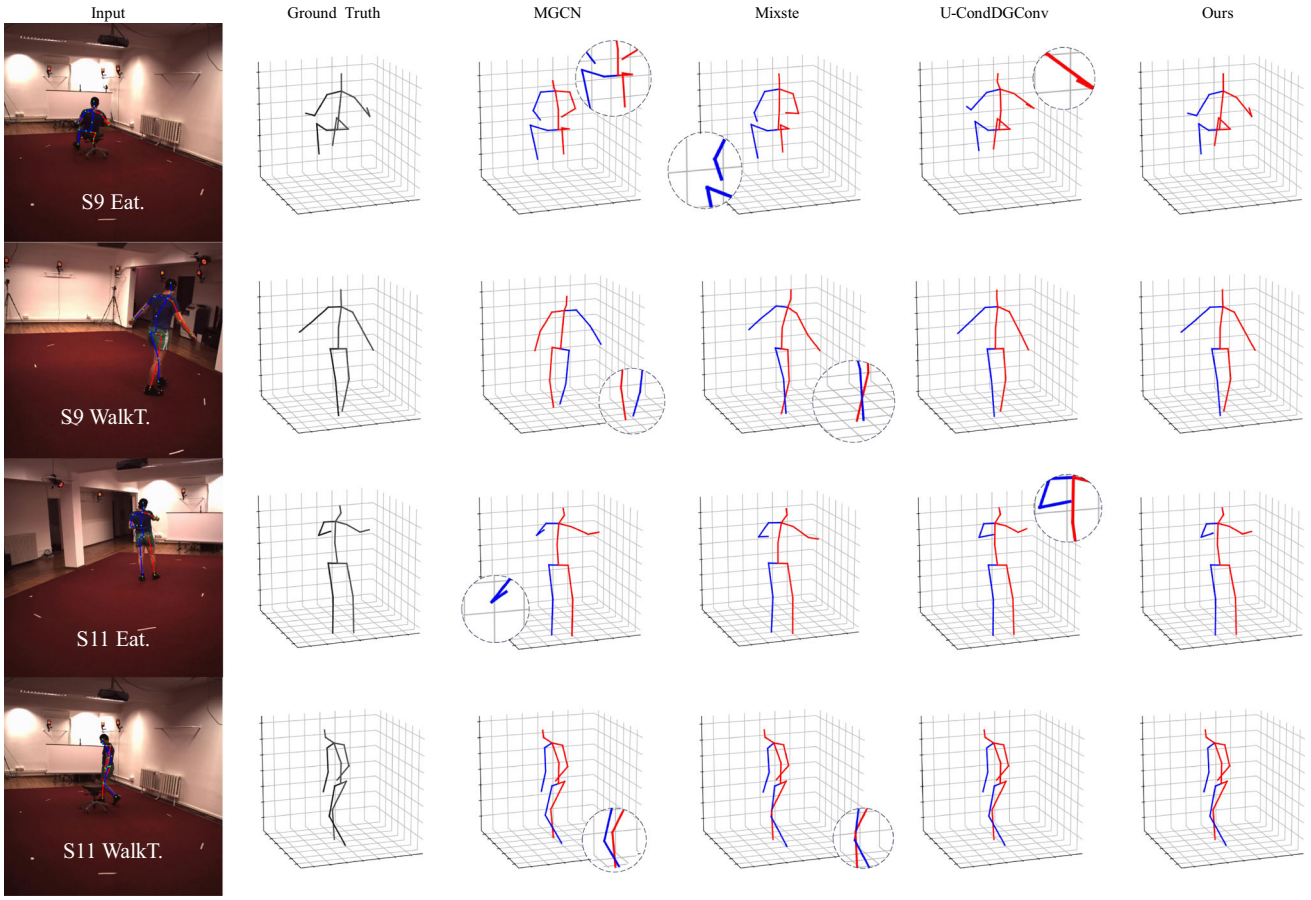

Fig. 3 A qualitative comparison is conducted with MGCN, Mixste and U-CondDGConv for subjects S9 and S11 on two actions within the Human3.6M dataset. Noticeable improvements are emphasized and magnified

ing the perceived efficacy of different approaches (Tables 3, 4).

Acknowledging a limitation, our method exhibits subpar performance compared to recent approaches [33], [54], [30] in scenarios with relatively low-quality estimated 2D pose data. This highlights a specific challenge and prompts a deeper exploration of methodologies to enhance the robustness of 3D human pose estimation models under varied 2D pose data quality conditions. Addressing this challenge is crucial for advancing the applicability of such models in real-world scenarios with diverse data sources (Table 5).

Discussion: the Effect of 2D Pose Quality. Going back to the inception of 3D pose lifting research, Martinez et al. [20] employed the SH 2D pose detector, fine-tuned on the Human3.6M dataset, to enhance 3D human pose estimation (HPE). This refinement led to a significant reduction in the average Mean Per Joint Position Error (MPJPE), from 67.5mm to 62.9mm, underscoring the pivotal role of high-quality 2D pose data in 3D HPE. Recent works, including [23, 30, 33], leveraged the advanced 2D pose detector HR-Net, achieving even better performance, with an average MPJPE of 39.8mm. Furthermore, Zhu et al. [31] achieved notable progress by fine-tuning the SH network [26] on the Human3.6M dataset, resulting in an average MPJPE of 37.5mm. However, it is important to note that these advancements still fall short of the results achieved when using ground truth (GT) 2D pose data.

The same pattern holds true when considering the HumanEva-I and MPI-INF-3DHP datasets. As depicted in Table 2, our method demonstrates a substantial 41% decrease in P-MPJPE on the HumanEva-I dataset. Notably, with ground truth (GT) 2D pose data, the P-MPJPE reduces from 15.3mm to 9.3mm when compared to the best-performing state-of-the-art algorithm. Meanwhile, on the MPI-INF-3DHP dataset, the MPJPE decreases from 32.2mm to 26.76mm.

As a result, the performance improvement of estimated poses predominantly hinges on the quality of 2D pose data. This quality can be achieved either by employing advanced 2D pose detectors capable of generating pose data closely resembling ground truth (GT) 2D pose or by fine-tuning existing pose detectors as necessary. In contrast, the utility of reconstructed 3D pose data generated by advanced pose detectors remains uncertain in certain scenarios. One

Table 2 Results from Protocol #2 for HumanEva-I are presented

Method	Walk			Jog			Avg
	S1	S2	S3	S1	S2	S3	
Martinez et al. [20] (ICCV'17)	19.7	17.4	46.8	26.9	18.2	18.6	24.6
Fang et al. [47] (AAAI'18)	19.4	16.8	37.4	30.4	17.6	16.3	23.0
Pavlakos et al. [48] (CVPR'18)	18.8	12.7	29.2	23.5	15.4	14.5	19.0
Lee et al. [49] (ECCV'18)†	18.6	19.9	30.5	25.7	16.8	17.7	21.5
Pavllo et al. [21] (CVPR'19)†	13.9	10.2	46.6	20.9	13.1	13.8	19.8
Liu et al. [22] (CVPR'20)†	13.1	9.8	26.8	**16.9**	12.8	13.3	15.5
Zheng et al. [53] (ICCV'21)†	14.4	10.2	46.6	22.7	13.4	13.4	20.1
Li et al. [29] (TMM'22)†*	14.0	10.0	32.8	19.5	13.6	14.2	17.4
Zhang et al. [30] (CVPR'22)†*	12.7	10.9	**17.6**	22.6	15.8	17.0	16.1
Ours (T=27, MRCNN)†	**12.3**	**9.4**	26.5	18.2	**12.3**	**12.5**	**15.3**
Li et al. [29] (TMM'22)†*	9.7	7.6	15.8	12.3	9.4	11.2	11.1
Ours (T=27, GT)†	**8.9**	**6.5**	**11.3**	**9.9**	**8.5**	**10.2**	**9.3**

The symbol† denotes the utilization of temporal information. Optimal performance is highlighted in bold, the second-best is underlined, and * signifies the reconstruction of an intermediate 3D pose sequence

Table 3 Outcomes from Protocol #1 for MPI-INF-3DHP are provided

Method	PCK↑	AUC↑	MPJPE↓
Mehta et al. [55] (3DV'17, T=1)	75.7	39.3	117.6
Pavllo et al. [21] (CVPR'19, T=81) †	86.0	51.9	84.0
Lin et al. [56] (BMVC'19, T=25) †	83.6	51.4	79.8
Wang et al. [23] (ECCV'20, T=96) †*	86.9	62.1	68.1
Zheng et al. [53] (ICCV'21, T=9) †	88.6	56.4	77.1
Chen et al. [57] (TCSVT'21, T=81) †	87.9	54.0	78.8
Hu et al. [28] (MM'22, T=96) †*	97.9	69.5	42.5
Shan et al. [54] (ECCV'22, T=81) †	97.9	75.8	32.2
Our PVA-GCN (T=27) †	**97.19**	**77.53**	**31.54**
Our PVA-GCN (T=81) †	**96.53**	**77.12**	**26.76**

The symbol † indicates the incorporation of temporal information. The optimal result is presented in bold, the second-best is underlined, and * denotes the reconstruction of an intermediate 3D pose sequence

such scenario is 3D human pose estimation in real-world conditions, typically evaluated through qualitative visualization [29]. Nevertheless, the question of whether 3D pose reconstructed from estimated 2D pose data can effectively contribute to pose-based tasks remains an area that has not been thoroughly explored. Given the straightforward nature of improving the performance of estimated 2D pose and the absence of clearly defined practical use cases, we argue that comparisons based on GT 2D pose data offer a more accurate representation of a model's 3D human pose estimation (HPE) capability than comparisons based on estimated 2D pose data.

4.4 Ablation studies

In our analysis, we eliminate gradients from our model design, which comprises Voxel, Points, and Proxy-Node layers. The appropriateness of AGCN layers is assessed by comparing our model with a version implemented using ST-GCN [41] blocks, resulting in the ablation of AGCN. The results for Protocol #2 across datasets Human3.6M and HumanEva-I consistently indicate superior performance with the use of AGCN blocks, as shown in Table 6. To ablate the strided design, we apply average pooling to the second dimension (i.e., temporal) of the feature map, as an alternative approach. The absence of the strided design not only results in a larger feature map representation, increasing from $F(C_{out}, 1, N)$ to $F(C_{out}, T, N)$, but also adversely affects the accuracy of 3D human pose estimation (3D HPE).

In order to validate the effectiveness of our Proxy-Node layer design, we compare it with a fully connected layer that uses the expanded feature map as its input. The results, as presented in Table 6, demonstrate a significant enhancement in performance achieved by our individual connected layer in effectively leveraging the structured representation of GCN. Visualizations of feature distinctions before the pre-

Table 4 Protocol #2 assesses the reconstruction error following rigid alignment, measured with P-MPJPE (millimeters), on the Human3.6M dataset

Method	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez et al. [20] (ICCV'17)	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Fang et al. [47] (AAAI'18)	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
Pavlakos et al. [48] (CVPR'18)	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
Lee et al. [49] (ECCV'18) †	34.9	35.2	43.2	42.6	46.2	55.0	37.6	38.8	50.9	67.3	48.9	35.2	31.0	50.7	34.6	43.4
Cai et al. [50] (ICCV'19) †	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 et al. [21] (CVPR'19) †	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 et al. [11] (CVPR'20) †	**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 et al. [22] (CVPR'20) †	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
Shan et al. [54] (MM'21) †	32.5	36.2	33.2	35.3	35.6	42.1	32.6	31.9	42.6	47.9	36.6	32.1	34.8	24.2	25.8	35.0
Our PVA-GCN (T=243, CPN) †	30.9	33.1	**30.3**	**31.7**	**33.1**	39.1	**31.1**	**30.5**	**42.5**	**44.5**	**34.0**	**30.8**	**32.6**	**22.1**	**22.9**	**32.5**
Martinez et al. [20] (ICCV'17)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.1
Ci et al. [39] (ICCV'19)	24.6	28.6	24.0	27.9	27.1	31.0	28.0	25.0	31.2	35.1	27.6	28.0	29.1	24.3	26.9	27.9
Our PVA-GCN (T=243, GT) †	**20.4**	**22.0**	**21.3**	**20.0**	**21.7**	**24.4**	**22.8**	**20.7**	**28.4**	**33.0**	**22.4**	**20.3**	**20.1**	**16.2**	**16.4**	**22.7**
Our PVA-GCN (T=243, GT) †*	**16.4**	**18.3**	**16.1**	**16.8**	**16.5**	**22.1**	**19.4**	**17.3**	**22.6**	**26.2**	**17.8**	**16.2**	**16.1**	**11.0**	**11.4**	**18.4**

In the top table, input 2D joints are obtained through detection using the cascaded pyramid network (CPN). The bottom table features input 2D joints with ground truth (GT). The symbol † denotes the utilization of temporal information, while * indicates the reconstruction of an intermediate 3D pose sequence. Optimal performance is highlighted in bold, and the second-best result is underlined

Table 5 Comparison with state-of-the-art methods on Human3.6M is conducted, implementing various receptive fields for ground truth 2D pose in the evaluation of Protocol #1

Method	Frames T	Parameters	P1(mm)
Pavllo et al. [21] (CVPR'19) †	27	8.56M	40.6
Liu et al. [22] (CVPR'20) †	27	5.69M	38.9
Li et al. [29] (TMM'22) †*	27	18.92M	**34.3**
Our PVA-GCN †	27	2.84M	34.5
Pavllo et al. [21] (CVPR'19) †	81	12.75M	38.7
Liu et al. [22] (CVPR'20) †	81	8.46M	36.2
Li et al. [29] (TMM'22) †*	81	≥18.92M	32.7
Our PVA-GCN †	81	3.61M	**31.8**
Pavllo et al. [21] (CVPR'19) †	243	16.95M	37.8
Liu et al. [22] (CVPR'20) †	243	11.25M	34.7
Our PVA-GCN †	243	4.56M	**27.3**
Wang et al. [23] (ECCV'20, T=96) †*	96	**1.69**M	25.6
Hu et al. [28] (MM'22, T=96) †*	96	3.42M	22.7
Our PVA-GCN($C_{out=96}$) †	243	4.73M	23.3
Li et al. [29] (TMM'22) †*	351	≥18.92M	30.5
Zhang et al. [30] (CVPR'22) †*	243	33.70M	21.6
Our PVA-GCN($C_{out=512}$) †	243	43.61M	**20.8**

The reported results include the reconstruction of an intermediate 3D pose sequence, marked by *
† indicates the introduction of time information

diction layers (i.e., individually and fully connected layers) are depicted in the upper and lower rows of Fig. 4. These visualizations in Fig. 4 really drive home the point about the power of our individual connected layer in making predictions. It is like having a detailed map of interpretable features, which a regular fully connected layer just cannot match. The independence of arm and leg joints in actions like "eating" and "walking" speaks volumes about the effectiveness of our approach in maintaining predictive accuracy. It is like having a sharper lens to capture the nuances of each movement.

Discussion: Limitation on Model. Similar to state-of-the-art methodologies, our approach confronts the challenge of heightened computational overhead. Notably, the data presented in the lower section of Table 5 underscores that our model surpasses the performance of state-of-the-art methods while requiring slightly more model parameters. This accentuates the dual limitation of increased computational demands and a marginal rise in model complexity. Addressing these challenges constitutes a central focus for our future work, where advanced techniques like model pruning will be explored to optimize efficiency without compromising performance.

Furthermore, echoing the constraints of existing methodologies [28, 30, 33], our approach exhibits a reliance on extensive training data. Despite achieving superior performance compared to state-of-the-art methods [23, 30], our model shows a dependency on a larger volume of training data. Subsequent endeavors will be dedicated to refining the model's generalization capabilities and diminishing its dependence on extensive datasets, thereby enhancing overall efficiency and applicability.

Table 6 An ablation study is performed to analyze the key designs of our PVA-GCN

No.	Method	Human3.6M		HumanEva-I	
		CPN	GT	MRCNN	GT
1	Point	38.9	27.9	18.1	11.6
2	Voxel	41.1	30.5	22.5	12.6
3	Point-AGCN	38.9	27.6	17.5	12.3
4	Voxel-AGCN	38.2	28.0	16.3	12.2
5	PVA-GCN(T=27) †	**37.9**	**25.9**	**15.3**	**9.3**

The results are derived from the average values obtained in Protocol #2, implemented with 27 receptive fields, considering various 2D pose detections across the Human3.6M and HumanEva-I datasets
† indicates the introduction of time information

5 Conclusion

In this paper, we propose a novel point-voxel absorbing graph convolutional network (PVA-GCN) method for addressing the problem of 3D human pose estimation. Our approach involves transforming the event stream into a sparse event cloud and voxel grids, creating a joint representation that strikes a balance between performance and efficiency. The dual representations facilitate improved performance by addressing the challenges of fragmented node feature learning and global classification feature aggregation encountered

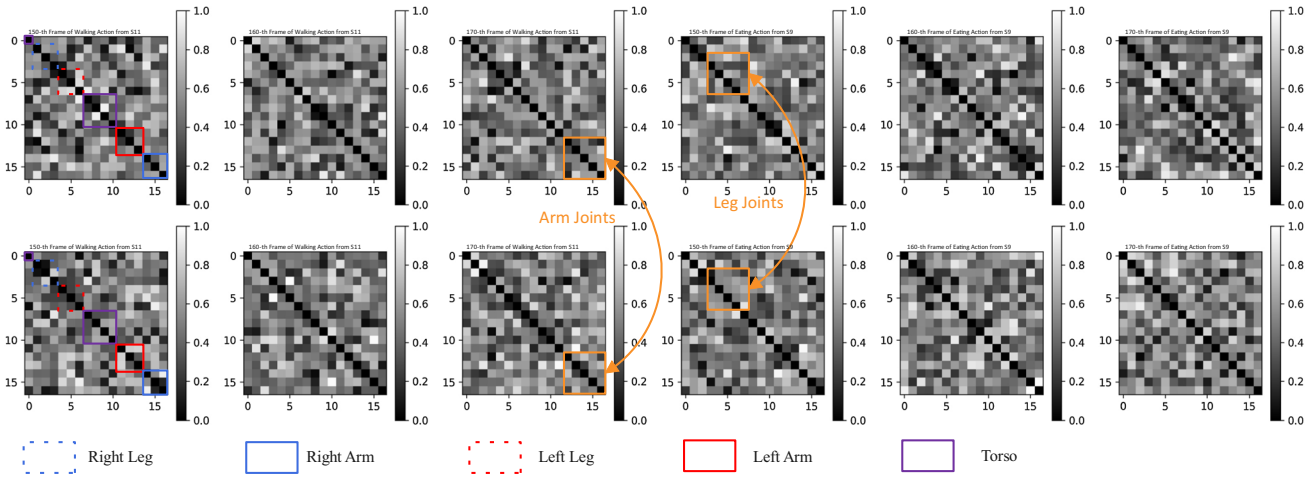

Fig. 4 Visualizations of inter-joint feature cosine similarity are presented for actions "Walking" (first three columns) and "Eating" (last columns) in the Human3.6M dataset

in previous event-based classification models. To achieve this, we introduce absorbing nodes into the dual graphs for global information aggregation, and employ absorbing graph convolution networks (AGCN) for structured feature learning and global feature aggregation simultaneously.

Our PVA-GCN framework's efficacy has been thoroughly validated through extensive experiments on multiple benchmark datasets for event-based classification. The results of these experiments showcase the superior performance of PVA-GCN when compared to state-of-the-art methods, utilizing ground truth (GT) 2D poses across datasets like Human3.6M, HumanEva-I, and MPI-INF-3DHP. We have substantiated the appropriateness of our model design through comprehensive ablation studies and visualizations. Additionally, studies such as [61, 62] offer valuable insights into leveraging graph-based methodologies for point cloud registration and innovating image quality assessment. These insights contribute significantly to discussions on 3D human pose estimation. In our future work, we plan to address the challenge of parameter efficiency by incorporating tuning techniques [63]. Furthermore, we intend to explore the impact of our model in diverse application scenarios, such as human behavior understanding. Additionally, we will delve into the examination of other loss terms, including those based on bone features [57] and motion trajectory [23], to further refine our approach.

Data availability The data used in this research work is obtained from publicly available datasets and will be made accessible for research purposes. The researchers acknowledge the importance of data sharing to promote reproducibility and further advancements in the field of 3D human pose estimation.

Declarations

Conflict of interest The authors declare no conflict of interest that could influence the interpretation of the results or the objective presentation of the research findings. This research is purely academic and does not involve any financial or personal relationships that could lead to bias.

References

1. 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. **152**, 1–20 (2016)
2. Presti, L.L., La Cascia, M.: 3D skeleton-based human action classification: a survey. Pattern Recognit. **53**, 130–147 (2016)
3. Yu, B., Liu, Y., Chan, K.: A survey of sensor modalities for human activity recognition. In: Proceedings of the 12th International Joint Conference on Knowledge Discovery, Budapest, Hungary, pp. 2–4 (2020)
4. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7718–7727 (2019)
5. Reddy, N.D., Guigues, L., Pishchulin, L., Eledath, J., Narasimhan, S.G.: Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15190–15200 (2021)
6. Zhang, Z., Wang, C., Qiu, W., Qin, W., Zeng, W.: Adafuse: adaptive multiview fusion for accurate human pose estimation in the wild. Int. J. Comput. Vis. **129**, 703–718 (2021)
7. He, Y., Yan, R., Fragkiadaki, K., Yu, S.-I.: Epipolar transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7779–7788 (2020)
8. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Comput. Vis. Image Underst. **192**, 102897 (2020)
9. Cheng, Y., Yang, B., Wang, B., Tan, R.T.: 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10631–10638 (2020)
10. Gong, K., Zhang, J., Feng, J.: Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8575–8584 (2021)
11. Xu, J., Yu, Z., Ni, B., Yang, J., Yang, X., Zhang, W.: Deep kinematics analysis for monocular 3d human pose estimation. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 899–908 (2020)
12. Urtasun, R., Fua, P.: 3D human body tracking using deterministic temporal motion models. In: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004. Proceedings, Part III 8, pp. 92–106 (2004). Springer
13. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3D human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3395–3404 (2019)
14. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2848–2856 (2015)
15. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, pp. 752–768 (2020). Springer
16. Chen, X., Wei, P., Lin, L.: Deductive learning for weakly-supervised 3D human pose estimation via uncalibrated cameras. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1089–1096 (2021)
17. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3D human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11199–11208 (2021)
18. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3D human pose estimation: a unified perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6238–6247 (2021)
19. 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, pp. 2344–2353 (2019)
20. 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, pp. 2640–2649 (2017)
21. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2019)
22. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.-C., Asari, V.: Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5064–5073 (2020)
23. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3D pose estimation from videos. In: European Conference on Computer Vision, pp. 764–780 (2020). Springer
24. He, K., Gkioxari, G.: P. doll ar, and r. girshick,"mask r-cnn,". In: Proceedings of IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
25. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
26. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14, pp. 483–499 (2016). Springer
27. 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, pp. 5693–5703 (2019)
28. Hu, W., Zhang, C., Zhan, F., Zhang, L., Wong, T.-T.: Conditional directed graph convolution for 3d human pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 602–611 (2021)
29. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimedia **25**, 1282–1293 (2022)
30. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13232–13242 (2022)
31. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: Unified pretraining for human motion analysis. arXiv preprint arXiv:2210.06551 (2022)
32. 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, pp. 3425–3435 (2019)
33. 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, pp. 13147–13156 (2022)
34. Lee, H.-J., Chen, Z.: Determination of 3D human body postures from a single view. Comput. Vis. Graph. Image Process. **30**(2), 148–168 (1985)
35. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1014–1021 (2009). IEEE
36. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR 2011, pp. 1385–1392 (2011). IEEE
37. Zhan, Y., Li, F., Weng, R., Choi, W.: Ray3d: ray-based 3d human pose estimation for monocular absolute 3D localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13116–13125 (2022)
38. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
39. 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, pp. 2262–2271 (2019)
40. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2d human pose. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, pp. 769–787 (2020). Springer
41. 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, vol. 32 (2018)
42. 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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, pp. 318–334 (2020). Springer
43. Lee, K., Woo, H., Suk, T.: Point data reduction using 3d grids. Int. J. Adv. Manuf. Technol. **18**, 201–210 (2001)
44. Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Cui, L., Wen, H.: Event-stream representation for human gaits identification using deep neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3436–3449 (2021)
45. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model CNNs. In: Proceedings of the IEEE Confer-

ence on Computer Vision and Pattern Recognition, pp. 5115–5124 (2017)
46. Miranda, L.J.: Understanding softmax and the negative log-likelihood. ljvmiranda921. github. io (2017)
47. 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, vol. 32 (2018)
48. 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, pp. 7307–7316 (2018)
49. Lee, K., Lee, I., Lee, S.: Propagating LSTM: 3D pose estimation based on joint interdependency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 119–135 (2018)
50. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2272–2281 (2019)
51. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: SRNET: Improving generalization in 3D human pose estimation with a split-and-recombine approach. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pp. 507–523 (2020). Springer
52. 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, pp. 16105–16114 (2021)
53. Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J.: Hemlets posh: learning part-centric heatmap triplets for 3d human pose and shape estimation. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3000–3014 (2021)
54. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-STMO: pre-trained spatial temporal many-to-one model for 3D human pose estimation. In: European Conference on Computer Vision, pp. 461–478 (2022). Springer
55. 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: 2017 International Conference on 3D Vision (3DV), pp. 506–516 (2017). IEEE
56. Lin, J., Lee, G.H.: Trajectory space factorization for deep video-based 3D human pose estimation. arXiv preprint arXiv:1908.08289 (2019)
57. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3D human pose estimation with bone-based pose decomposition. IEEE Trans. Circuits Syst. Video Technol. **32**(1), 198–209 (2021)
58. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4966–4975 (2016)
59. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3D body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 991–1000 (2016)
60. 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, pp. 7025–7034 (2017)
61. Sun, L., Zhang, Z., Zhong, R., Chen, D., Zhang, L., Zhu, L., Wang, Q., Wang, G., Zou, J., Wang, Y.: A weakly supervised graph deep learning framework for point cloud registration. IEEE Trans. Geosci. Remote Sens. **60**, 1–12 (2022)
62. Sun, S., Yu, T., Xu, J., Zhou, W., Chen, Z.: Graphiqa: learning distortion graph representations for blind image quality assessment. IEEE Trans. Multimedia **25**, 2912–2925 (2023). https://doi.org/10.1109/TMM.2022.3152942
63. Yu, B.X., Chang, J., Liu, L., Tian, Q., Chen, C.W.: Towards a unified view on visual parameter-efficient transfer learning. arXiv preprint arXiv:2210.00788 (2022)