Pattern Recognition 173 (2026) 112805

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Structure and sensitivity in 3D human pose similarity quantification and estimation

Kyoungoh Lee [a,1], Jungwoo Huh [b,1], Jiwoo Kang [c], Sanghoon Lee [b,*]

[a] Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea
[b] Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
[c] Department of IT Engineering, Sookmyung Women's University, Seoul, South Korea

ARTICLE INFO

Keywords:
3D Human pose estimation
Dual graph convolutional networks
Perceptual pose similarity metric

ABSTRACT

Recent advancements in deep learning have improved quantitative accuracy in 3D human pose estimation, but the estimated poses occasionally suffer from visual defects such as joint tremors and protrusions. While existing 3D pose similarity metrics and estimation models managed to reduce visual defects by addressing the structure of human poses, they still struggle in scenarios where visually sensitive joints are prevalent, particularly in cases of self-occlusion. In this paper, we identify these visually sensitive joints and demonstrate the significance of explicitly considering structure and sensitivity in the problem of 3D human pose estimation. Building upon the successful consideration of human pose structure, we first propose a new enhanced pose similarity metric PSIM$^+$, which models sensitivity similarity to further capture human perception and focus on visual defects. Furthermore, we introduce a new 3D pose estimation model Dual Graph-based Convolutional Neural Networks (DG-CNN), which reconstructs 3D poses by focusing on the spatio-temporal correlation of the skeletal structure and actively controlling visually sensitive joints. By incorporating a novel similarity loss function, our model can implicitly model the structure and sensitivity of human poses through its architecture and explicitly through direct supervision. Our model not only improves the accuracy of the estimated pose but also increases the perceptual quality as evaluated by PSIM$^+$, verifying the significance of structure and sensitivity awareness. Through rigorous benchmarking, we demonstrate that our metric and estimation model achieve the highest correlation with user scores and perform best in situations where visually sensitive joints are prevalent.

1. Introduction

Recent research in three-dimensional (3D) human pose estimation (HPE) has trended towards a two-step approach, where an estimated 2D pose is taken from an RGB image as a feature and then reconstructed into a 3D pose [1,2]. This has led to noticeable improvements in quantitative performance compared to direct approaches. Nevertheless, significant challenges persist, particularly in mitigating visual defects such as joint tremors, frame-to-frame jittering of joint positions, and joint protrusions, anatomically implausible dislocations of joints that appear to stick out unnaturally. These artifacts, as illustrated in Fig. 1(a), degrade the perceptual quality of the estimated 3D poses. Identifying and addressing such visual defects explicitly is thus essential to advancing 3D HPE methods.

Conventional Euclidean distance-based metrics, such as the mean per joint position error (MPJPE) [3], fail to capture structural similarity in 3D human poses, as they evaluate joints independently. To complement these traditional metrics, we introduced the pose similarity (PSIM) metric [4], which explicitly considers the structure of the human pose while measuring similarity and demonstrated a strong correlation between visual defects perceived by humans. From the estimation perspective, recent methods have focused on enhancing model estimation power by incorporating the spatio-temporal structure of the human pose [4,5]. These approaches have managed to reduce visual defects stemming from the insufficient consideration of human pose structure. Both research on visual defect identification and 3D HPE emphasize structural awareness, highlighting the importance of understanding and leveraging the human pose structure in 3D HPE.

Despite these advancements, structural considerations alone remain insufficient in challenging cases, specifically, in the presence of occlusion (e.g., Figs. 1(b) and (c)). Existing metrics for the two predicted 3D poses in the example deem them similar while most people perceive that the result in Fig. 1(c) is more similar and better estimated. The main reason why existing metrics exhibit such problems is that

* Corresponding author.
E-mail addresses: longweek7@etri.re.kr (K. Lee), gjwjddn9@yonsei.ac.kr (J. Huh), jwkang@sookmyung.ac.kr (J. Kang), slee@yonsei.ac.kr (S. Lee).
[1] These authors contributed equally.

https://doi.org/10.1016/j.patcog.2025.112805
Received 6 July 2024; Received in revised form 20 November 2025; Accepted 28 November 2025
Available online 2 December 2025
0031-3203/© 2025 Published by Elsevier Ltd.

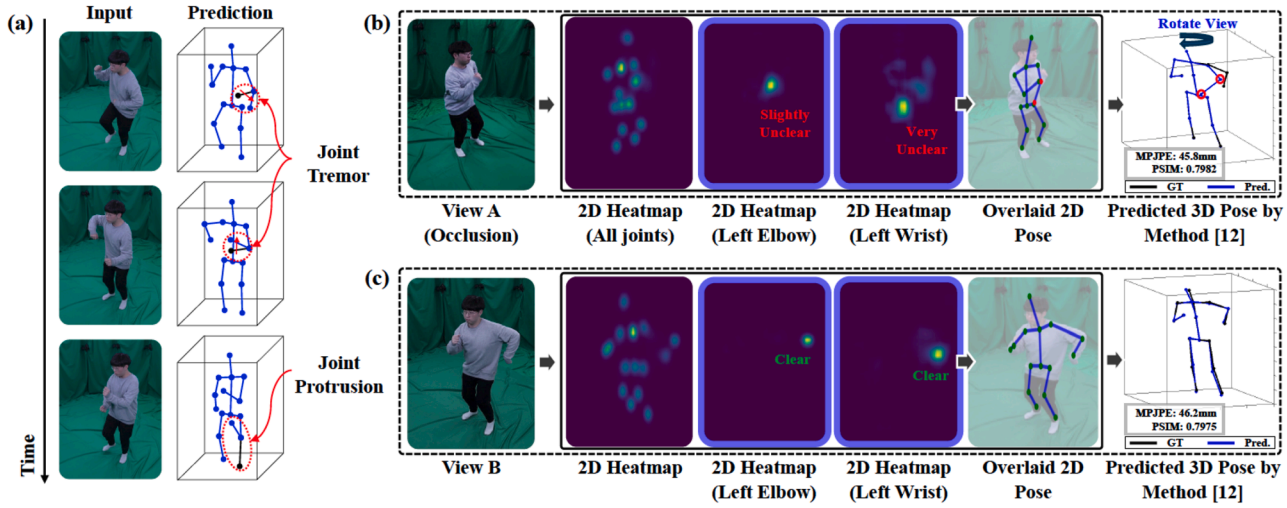

Fig. 1. (a) Typical visual defects in 3D pose estimation: joint tremor and joint protrusion. (b) For visually sensitive joints (left elbow/wrist, red circles), these defects become prominent. (c) Without sensitive joints, the joints are estimated more accurately. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

humans tend to focus on visually sensitive joints, in this case, joints with large errors such as the limb joints shown in Fig. 1(b). As a result, even when torso joints are well estimated, the higher uncertainty of limb joints results in a perceptual dissimilarity between the estimated and ground-truth poses, which is a common phenomenon in the human visual system and well-studied in image quality assessment research [6,7]. The case is analogous for 3D HPE methods. This example clearly demonstrates a critical limitation of existing metrics and 3D HPE methods, which do not explicitly account for visually sensitive joints.

Motivated by this observation, we propose a novel 3D human pose similarity quantification and estimation method that explicitly incorporates sensitivity awareness to effectively account for visually sensitive joints. We first propose an enhanced perceptual pose similarity metric, PSIM$^+$, explicitly designed to incorporate sensitivity awareness in addition to structural considerations. PSIM$^+$ measures the overall similarity between two poses by combining three types of similarity measures: positional, structural, and sensitivity similarities. The positional similarity measures the similarity based on joint distance and the structural similarity considers human structure by incorporating body part information and joint hop distance into the similarity measure. While the former two metrics capture the structural components, the sensitivity similarity selectively evaluates the errors of visually sensitive joints after identifying them. Compared to existing similarity measures, PSIM$^+$ simplifies structural evaluation while explicitly emphasizing visually sensitive joints, thereby significantly enhancing perceptual relevance.

For 3D HPE, we propose a novel Dual Graph-based Convolutional Neural Network (DG-CNN) explicitly designed to incorporate sensitivity awareness into human pose estimation. Unlike the traditional two-step estimation paradigm, our method first predicts occlusion information in addition to 2D joint positions to accurately identify visually sensitive joints. Subsequently, these features are lifted from 2D to 3D using a dual-graph architecture, where one graph models structural relationships among joints, and the other explicitly emphasizes visually sensitive joints. By integrating these complementary graphs, DG-CNN significantly improves the stability and robustness of pose estimation, particularly under occlusion scenarios, compared to traditional methods that rely solely on structural awareness. Finally, we apply a perceptual similarity loss function based on our proposed PSIM$^+$ metric, which reinforces structural and sensitivity considerations both implicitly through the model architecture and explicitly via direct supervision.

Our main contributions are summarized as follows::

- We propose PSIM$^+$ that explicitly incorporates sensitivity awareness in addition to structural considerations, providing a more perceptually relevant evaluation of 3D human poses.
- We design DG-CNN that models both structural and sensitivity relationships among joints, enabling stable and robust 3D pose estimation even under severe occlusion.
- We demonstrate through extensive experiments on multiple benchmarks that our method achieves consistent improvements in both quantitative accuracy and perceptual quality especially in challenging scenarios, surpassing existing state-of-the-art approaches.

2. Related work

2.1. Similarity measurement in pose

Most pose estimation studies rely on a Euclidean distance-based metric (*e.g.*, MPJPE) as a performance indicator [3]. Some authors have introduced the Procrustes MPJPE (P-MPJPE) which is a modified MPJPE metric that uses Procrustes alignment [8] to remove subtle transformations and focus on measuring overall pose similarity. Moreover, the authors of [9] provided the 3D percentage of correct key points (PCK), which is suitable for outdoor environments where errors are likely to be large. While these metrics have been widely used in this field, they fail to capture the discrepancy between quantitative and perceptual results. They are designed only for a single frame, making it difficult to quantify visual problems such as tremors. To solve these problems, we presented PSIM [4], which takes into account structural similarity and uses a temporal edge-based pooling scheme. Although PSIM captured many structural errors and had the highest correlation with the user score, it could not easily measure the perceptual quality of estimated poses under various visual defects, especially self-occlusion.

2.2. Structure awareness in 3D pose estimation

Estimating 3D human pose from a single image is highly challenging, and many approaches have attempted to address this by incorporating structural awareness. Early works adopted Graph Convolutional Network (GCN)-based formulations to encode human skeletal topology and capture spatial-temporal relationships. Bin et al. [10] introduced a GCN that explicitly models the articulated skeletal

structure, while Zhao et al. [11] incorporated semantic relationships between local and global nodes. Cai et al. [12] further extended this idea to capture spatial-temporal dependencies, and more recently, Yu et al. [13] enhanced this paradigm by adaptively modeling global-local joint relationships. With the rise of Transformers, several works explored attention-based structural modeling [14]. Zhang et al. [15] combined spatial and temporal Transformers to exploit their complementary information. Peng et al. [16] modeled joint kinematics and motion trajectories with separate Transformer modules to emphasize their structural validity, while Lang et al. [17] further incorporated event camera metadata to strengthen temporal reasoning. Hybrid architectures combining GCN and Transformers have also been explored. Cheng et al. [18] integrates GCN modules into a Transformer backbone to reinforce joint-level relational cues, whereas Li et al. [19] replaces the Transformer attention mechanism with a lightweight MLP-GCN hybrid, improving computational efficiency. Despite these advances, structure-aware methods are still sensitive to severe self-occlusion or missing joint observations, as their structural priors dominate ambiguous visual cues.

2.3. Sensitivity awareness in 3D pose estimation

For stable 3D pose estimation, many works have explored sensitivity awareness, particularly in handling occlusions and ambiguous observations. Cheng et al. [20] filtered out occluded joints in 2D by combining joint heatmaps with optical-flow consistency across video frames. Han et al. [21] implicitly modeled occlusion through heatmap uncertainty, using the variance of joint distributions to refine 3D pose predictions. Li et al. [22] further addressed ambiguity by allowing multiple pose hypotheses for uncertain joints. Multi-view approaches offer another direction: Bragagnolo et al. [23] fused multi-view pose features for more robust estimation, while Zhang et al. [24] employed a self-supervised multi-view consistency loss to improve reconstruction quality. However, these methods either model occlusion implicitly or rely on multiple viewpoints to circumvent sensitivity issues. None of them explicitly incorporate such sensitivity cues into the core 3D pose estimation process.

3. Perceptual pose similarity (PSIM$^+$)

3.1. Human perception in pose estimation

Most human pose estimation methods use various metrics [3,4,9] to measure the error between two postures. However, these metrics have limitations in capturing perceptual differences in 3D poses.

Fig. 2 depicts the correlation of the PSIM and sensitivity similarity on the self-occlusion-centered similarity validation set. Certain joints with large errors in this set are frequently observed due to occluded joints. Thus, these joints are visually sensitive, and the results of the sensitivity similarity significantly correlate with the user MOS. From this tendency, humans feel the two poses are more similar if the errors on the sensitive joints are lower. Therefore, we propose a new pose metric by exploiting human perception. The perception-based metric distinguishes structural differences and sensitive points, simultaneously.

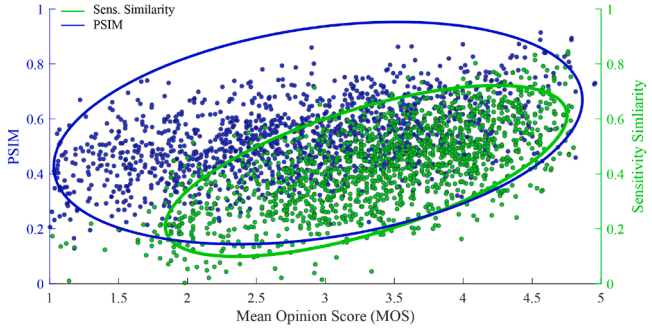

Fig. 2. Comparison of sensitivity similarity and PSIM in the self-occl.-centered similarity validation dataset. Each ellipse displays an estimated distribution of the scatter.

3.2. Positional similarity

Let the 3D pose vector for the reference pose be $\bar{J}_{\mathrm{R}} = [j_1, j_2, \cdots, j_k], \forall k \in K, j^k \in \mathbb{R}^{1\times 3}$ where $\bar{J}_{\mathrm{R}}$, j_k, and K denote the joint position vector of the reference pose, the 3D position of the k^{th} joint, and the set of joints, respectively. To obtain the positional similarity between two 3D poses, we exploit the similarity function from [25]. Let $\bar{J}_{\mathrm{T}}$ be the joint position vector of the target pose (*i.e.*, prediction). Then, the positional similarity is measured as follows:

$$\mathrm{P}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) = \frac{2(\bar{J}_{\mathrm{R}} \cdot \bar{J}_{\mathrm{T}})}{(\bar{J}_{\mathrm{R}})^2 + (\bar{J}_{\mathrm{T}})^2}. \tag{1}$$

This similarity role is the same as Euclidean pooling (*e.g.*, MPJPE). However this approach does not take into account the structural characteristics of the 3D posture as a natural signal. Thus, it is insufficient to capture structural information for the posture and ignores perceptual accuracy.

3.3. Structural similarity

Human posture is represented by structural elements (joint position, limb length, body parts, etc.), and considering these elements have boosted performance in human posture-related tasks. In our previous work [4], we demonstrated that a high correlation between structural and perceptual similarities exists in human posture. To quantify this structural information, we first measure the hop vector (*i.e.*, vector between two joints) at the body part level. As depicted on the left side of Fig. 3, the joints are classified into five groups: spine, 2 arms, and 2 legs. We compute the structural similarity of the body part, which is the smallest structural unit. Let $\bar{p}_1$ be the structure vector of body part 1 for the left arm, represented on the right side of Fig. 3. Within this body part, we measure three hop vectors from the root joint, which is the starting point of the body part, as $\bar{l}_h^{p_1} = j_h^{p_1} - o^{p_1}, \forall h \in \{1,2,3\}$, where h, $\bar{l}_h^{p_1}$, and o^{p_1} are the joint index in the body part, hop vector from the origin point, and origin point, respectively.

The difference in the scale should be minimized to compare the similarity of the two poses from a structural perspective. Inspired by mean-subtracted contrast normalization [26], we normalize the three hop vectors for each body part as structural elements to minimize the scale difference. The structural elements in body part 1 are expressed as $\bar{p}_1 = [\hat{l}_1^{p_1}, \hat{l}_2^{p_1}, \hat{l}_3^{p_1}], \hat{l}_h^{p_1} = \bar{l}_h^{p_1} / u^{p_1}, \forall h \in \{1,2,3\}$, where $\bar{p}_1$ and u^{p_1} are the structural vector and unit limb vector ($= ||\bar{l}_1^{p_1}||$) in body part 1, respectively. Finally, let $\bar{P}_{\mathrm{R}}$ be the structural vector of reference pose, consisting of several body part's structural vectors for several body parts $\bar{P}_{\mathrm{R}} = [\bar{p}_1, \bar{p}_2, \bar{p}_3, \bar{p}_4, \bar{p}_5]$. Then, the structural similarity is calculated as follows:

$$\mathrm{S}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) = \frac{2(\bar{P}_{\mathrm{R}} \cdot \bar{P}_{\mathrm{T}})}{(\bar{P}_{\mathrm{R}})^2 + (\bar{P}_{\mathrm{T}})^2}. \tag{2}$$

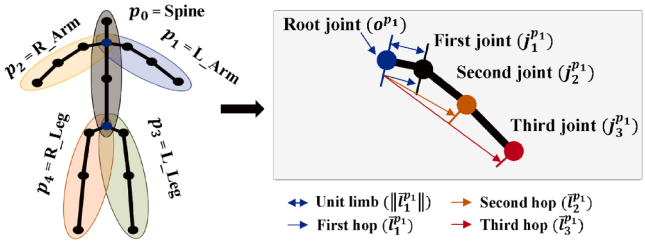

Fig. 3. Body part category and structural similarity components in a body part 1.

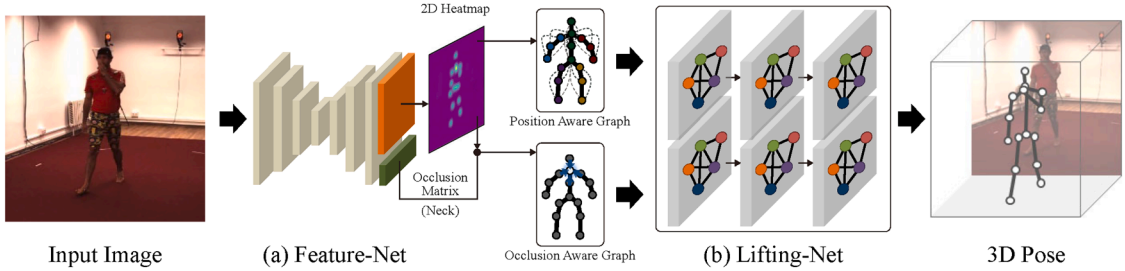

Fig. 4. The overall framework of our 3D HPE method.

The proposed hop vector-based structural similarity is similar to PSIM's structural similarity. The hop vectors include the direction and magnitude between two joints in a body part, which can be obtained very simply.

3.4. Sensitivity similarity

The human visual system tends to focus more on objects that move slowly or in different directions with respect to the background [7]. In a structural signal, visual attraction is prominent in areas with large errors [27], such as the limb joints in the case of posture. Thus, we compare the similarity by selecting joints sensitive to human vision. Let δ be the threshold to determine the joint sensitivity. This value is obtained using the MPJPE between two 3D poses. From this threshold, we obtain the sensitive joint index set $\mathcal{A} = \{k|\varepsilon_k \geq \delta\}, \forall k \in K$, where ε_k is the k^{th} joint distance error between the reference and target poses. Then, the sensitive joint vector of the reference pose is created by the selected joint index as $\bar{V}_{\mathrm{R}} = [j_1, j_2, \cdots, j_v], \forall v \in \mathcal{A}$, where $\bar{V}_R$ is the sensitive joint vector consisting of the sensitivity joint index set $\mathcal{A}$. The similarity of sensitive joint vectors between the target and reference poses is calculated in the following equation:

$$\mathrm{V}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) = \frac{2(\bar{V}_{\mathrm{R}} \cdot \bar{V}_{\mathrm{T}})}{(\bar{V}_{\mathrm{R}})^2 + (\bar{V}_{\mathrm{T}})^2}. \tag{3}$$

This measure is particularly effective for limb joints, which generally exhibit higher sensitivity due to their frequent occlusion compared to torso joints. As a result, to enhance PSIM, we propose a new pose similarity (PSIM$^+$):

$$\mathbf{PSIM}^+(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) \quad = w_{\mathbf{p}}\mathrm{P}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) + w_{\mathbf{s}}\mathrm{S}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}) + w_{\mathbf{v}}\mathrm{V}(\bar{X}_{\mathrm{R}}, \bar{X}_{\mathrm{T}}), \tag{4}$$

where the weights $w_{\mathbf{p}}, w_{\mathbf{s}}$, and $w_{\mathbf{v}}$ are the scale factors that control the magnitude and importance of each similarity.

3.5. Temporal pooling

Following the temporal edge in [4], we set a sliding window Z that moves according to the frame index t. The pose variation at the t^{th} frame is computed with respect to the mean pose of the window. The mean-subtracted 3D pose at the sliding window index z is expressed as $\mathrm{D}(z) = X(z) - \mu_{\mathbf{X}}$, where $X(z)$ and $\mu_{\mathbf{X}}$ are the z^{th} 3D pose and the mean 3D pose in window. Finally, let $\bar{X}_{\mathrm{R}}(t)$ and $\bar{X}_{\mathrm{T}}(t)$ be the reference and target temporal poses at the t^{th} frame. Then, the temporal PSIM$^+$ in terms of the entire frame is:

$$\mathbf{PSIM}^+(\mathbf{X}_{\mathrm{R}}, \mathbf{X}_{\mathrm{T}}) = \frac{1}{|T|}\sum^{T} w_{\mathbf{t}}(t) \cdot \mathbf{PSIM}^+(\bar{X}_{\mathrm{R}}(t), \bar{X}_{\mathrm{T}}(t)), \tag{5}$$

where $\mathbf{X}_{\mathrm{R}}$ and $\mathbf{X}_{\mathrm{T}}$ are a series of sequences for the reference and target poses, respectively. In each frame, the temporal poses at the unit window level measure the similarity, and the change is calculated as a weight $w_{\mathbf{t}}(t) = 1 - (\frac{1}{|Z|}\sum^{Z} ||\mathrm{D}_{\mathrm{R}}(z) - \mathrm{D}_{\mathrm{T}}(z)||_2)$. When visual defects, such as joint tremors or protrusions, occur within a window, the difference in temporal variation acts as a penalty, further reducing the degree of similarity.

4. Dual graph-based pose estimation (DG-CNN)

An overview of DG-CNN is presented in Fig. 4. The design is inspired by the PSIM$^+$ metric, particularly its modeling of perceptual sensitivity to visually important joints. DG-CNN first predicts joint sensitivity from features extracted by FeatureNet, including 2D heatmaps and occlusion information, as illustrated in Fig. 4(a). This sensitivity information is then integrated into the 2D-to-3D lifting process in LiftingNet, which predicts the 3D pose relative to the root joint, as shown in Fig. 4(b). Furthermore, PSIM$^+$ is also used as part of the similarity loss to provide perceptual supervision, encouraging the model to produce poses that are more visually plausible.

4.1. Feature-Net

2D Pose Estimation. We define the pose features from an image as a combination of the 2D pose and joint occlusion vector. We exploit a convolutional neural network (CNN) model [28] as a feature extractor to estimate the joint positions on the image domain. The joint position is represented as a 2D Gaussian heatmap with a single peak on the joint position. Through integral regression[29], the loss function of the feature extractor is defined as follows:

$$\mathcal{L}_{2\mathrm{D}} = \sum_{i=1}^{K} \left\| j_i - j_i^{\mathrm{GT}} \right\|, \tag{6}$$

where K, i, j, and j^{GT} are the number of joints, joint index, predicted joint location, and GT joint location, respectively.

Occluded Joint Detection. Joints are frequently occluded in monocular images; thus, estimating the exact 2D location of an obscured joint is challenging. This produces incomplete 2D poses accompanied by visually sensitive joints, which is a critical problem for two-step approach models. We stress that including occlusion information for each joint can greatly improve the 3D HPE performance by explicitly incorporating sensitivity awareness. To detect occluded joints from an image, we use the properties of 2D HPE models to determine the joint location. Thus, we present a novel joint occlusion detection module based on multitask learning. The joint occlusion information is a 1D binary vector, and inferring this vector is formulated as a multi-label binary classification. However, the dataset for occluded joint detection is unbalanced because each joint has different degrees of freedom (DOFs) [30]. Therefore, we introduce the focal loss to deal with the occlusion label imbalance problem[31]. The classification loss to detect the occluded joint is expressed as follows:

$$\mathcal{L}_{\mathrm{OC}} = -\alpha \cdot (1 - e^{-C_E})^{\gamma} \cdot C_E, \tag{7}$$

where α, γ, and C_E are the control parameters and the binary cross entropy function, respectively. The binary cross entropy is defined as $C_E = c^{\mathrm{GT}} \log(c) + (1 - c^{\mathrm{GT}}) \log(1 - c)$ where $c \in \{0, 1\}^{1\times K}$ and $c^{\mathrm{GT}} \in \{0, 1\}^{1\times K}$ are the predicted and GT occlusion labels, respectively. We merge the two losses in Eqs. (6) and (7) to define the final loss function for FeatureNet:

$$\mathcal{L}_{Feat} = \omega_1 \mathcal{L}_{2\mathrm{D}} + \omega_2 \mathcal{L}_{\mathrm{OC}}, \tag{8}$$

where $\omega_{2\mathrm{D}}$ and ω_{OC} are the loss weights for each task.

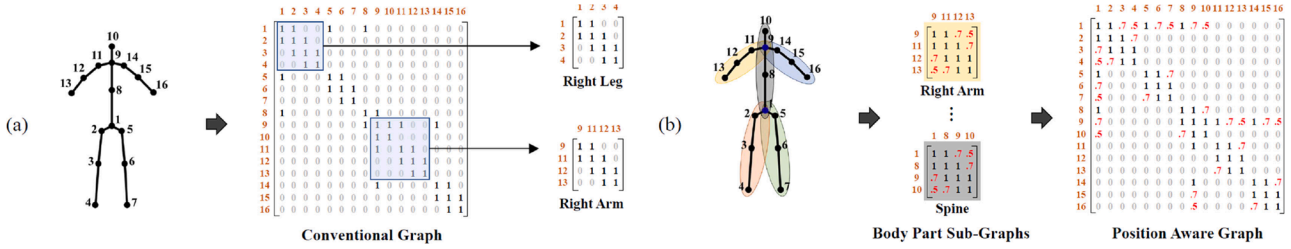

Fig. 5. Difference between the conventional graph and position-aware graph of the skeletal structure. (a) Simple adjacency matrix, (b) our position-aware matrix.

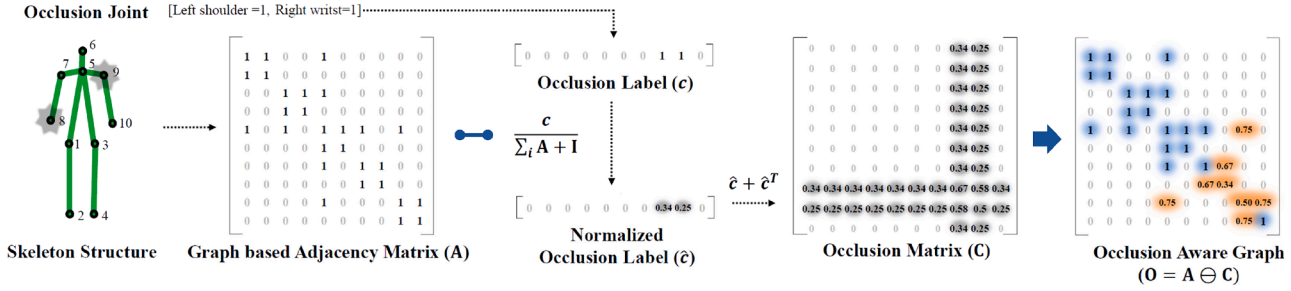

Fig. 6. Procedure for constructing an occlusion matrix from the occlusion label.

4.2. Lifting-Net

For lifting the dimension of the pose, we introduce a GCN, which has recently been widely applied in learning topological structures. Furthermore, in this paper, we propose a novel GCN based on dual graphs to model both structural and sensitivity awareness. The dual graphs are composed of a position-aware graph and an occlusion-aware graph in parallel. The position-aware graph is designed in the form of multiple sub-graphs that reflect the elements of the structural similarity metric. The occlusion-aware graph is structured to directly handle occlusions, which are visually sensitive joints.

Graph Convolution. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ express a graph, where $\mathcal{V}$ is a set of nodes (K joints corresponding to T frames), and $\mathcal{E}$ is a set of edges (limbs and temporal connections). Based on this graph, the l^{th} convolutional operation in vanilla GCN is written as follows: $\mathrm{X}^{(l)} = \sigma(\mathrm{W}\mathrm{X}^{(l-1)}\tilde{\mathrm{A}})$, where $\mathrm{X}^{(l)} \in \mathbb{R}^{N\times V}$, $\tilde{\mathrm{A}} \in \mathbb{R}^{V\times V}$, and $\mathrm{W} \in \mathbb{R}^{M\times N}$ are all node features, the normalized adjacency matrix, and the learnable parameter matrix, respectively. The dimensions V, M, and N are the numbers of vertices of the adjacency matrix with $V = KT$, l^{th} nodes, and $(l-1)^{th}$ nodes, respectively.

Position-aware Graph.

The spatial correlation should be strengthened to estimate accurate 3D poses from the 2D pose features. Fundamentally, the human body is represented by a fixed graph in which each joint is connected. Thus, all joints influence each other in estimating the pose [1]. In structural similarity, the influence is inversely proportional to the distance (hop) between the two joints at the body part level. Based on this concept and the five body part configuration in Fig. 5(b), we present the position-aware GCN. Unlike conventional CGN depicted in Fig. 5(a), position-aware GCN divides one graph into five sub-graphs corresponding to each body part. The graph reflects the distance between each joint at the body part level. To aggregate features of nodes with high spatial correlation, the position-aware GCN collects features according to the edge connection. However, 3D HPE has a differential spatial correlation depending on the distance between all connected joints. The distance between two nodes is calculated to collect the spatial correlation weighted by the hop for all connections in the body part level as $w_d(v,u) = 1/(1 + d(v,u))$, where $w_d(v,u)$ and $d(v,u)$ are the edge distance weight and distance counter, respectively. The distance counter calculates the hop between the target node v and reference node u ($d(v,u) \in \{0,1,2,3\}$).

After calculating the distance between all node pairs in the graph, the feature of the reference node u with respect to target node v is calculated by multiplying the edge distance weights by the concatenated features:

$$f_{v,u} = w_d(v,u) \cdot \mathrm{CONCAT}(h_v^{l-1}, h_u^{l-1}), \quad \forall\, v,u \in \mathcal{G}, \tag{9}$$

where $f_{v,u}$, h_u^{l-1}, h_v^{l-1}, and CONCAT are the position embedded feature and reference and target node features at the $(l-1)^{th}$ layer, and concatenation function, respectively. Now let $\mathcal{S}$ be the set of body parts. The aggregation procedure for features of each body part is $m_{S_i,u} = w_s \mathrm{AGG}_s(\{f_{v,u}, \quad \forall\, v \in S_i\})$, where $m_{S_i,u}$, s_i, w_s, $\mathcal{S}$, and AGG_s are the aggregated u^{th} node message from the i^{th} subgraph, aggregated features based on subgraph $\mathcal{S}$, body part weight, number of subgraphs, and aggregator, respectively. The messages aggregated at the subgraph level are aggregated again based on the entire graph: $\tilde{s}_u = \mathrm{AGG}_l(\{m_{S_i,u}, \quad \forall\, S_i \in \mathcal{S}\})$, where $\tilde{s}_u$ and AGG_l denote the aggregated features for node u and aggregator of the l^{th} layer, respectively. Finally, the aggregated features are transformed using the weight matrix and non-linear function: $h_u^{(l)} \leftarrow \sigma(W^l \cdot \tilde{s}_u)$, where $h_u^{(l)}$ represents the l^{th} layer u node value. We use the mean aggregation for all aggregator functions in this position-aware graph network. In summary, the position-aware graph uses and expresses the fine connectivity of the skeletal structure in more detail.

Occlusion-aware Graph. Occlusion is generally the main cause of sensitive joints such as the limb joints, which makes the sensitivity similarity difference even larger. Therefore, in 3D HPE, these joints should be directly controlled to improve accuracy in terms of perceptual similarity. To deal with occlusion, we propose an approach that exploits an occlusion matrix that regulates the dependence on incomplete joints. The matrix is generated by the occlusion vector c detected using Feature-Net. The occlusion vector is a 1D binary vector, with 1 set for the obscured joint and 0 otherwise. Because each joint has a different dependency on adjacent joints, an occluded joint is normalized based on its connectivity. For instance, joints with many adjacent joints are more robust to occlusion due to sufficient reference information, and thus receive lower scores compared to joints with fewer neighbors.

To adjust the node dependency according to the number of edges, we create a normalized occlusion vector $\hat{c}$ calculated as $\hat{c} = \frac{c}{\sum_i \mathrm{A}+\mathrm{I}}$. The normalized occlusion vector $\hat{c}$ is converted to the occlusion matrix C, which is a symmetric matrix. By subtracting the occlusion matrix C from the adjacency matrix A, the occlusion weighted adjacency matrix

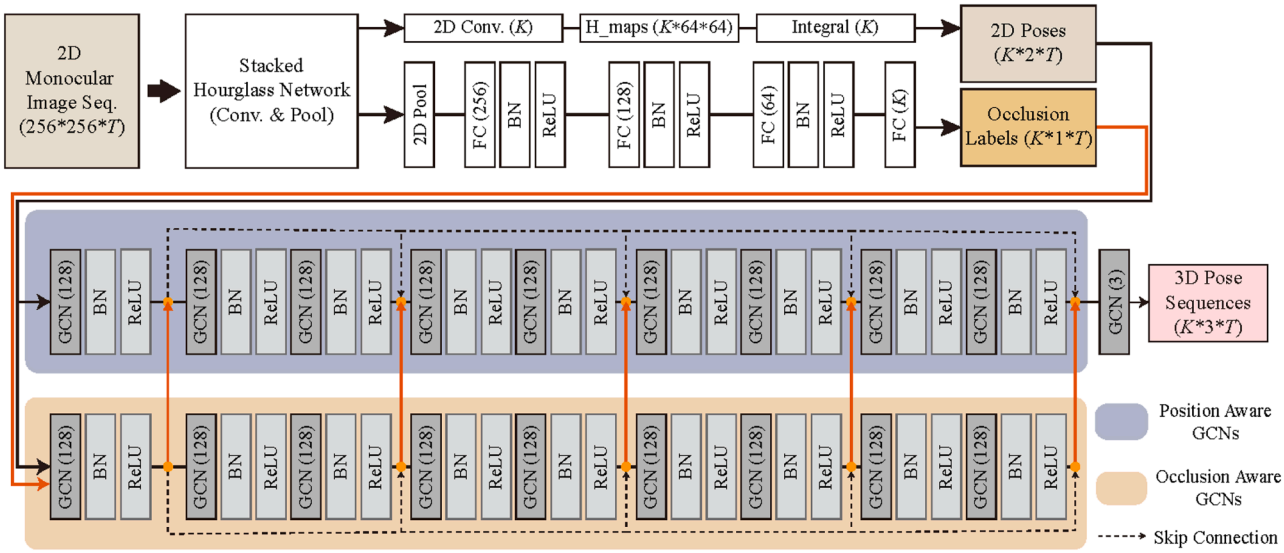

Fig. 7. Detailed dual graph-based 3D pose estimation model. The orange arrow represents a summation step for collecting the intermediate outputs between position and occlusion-aware GCNs.

(*i.e.*, occlusion-aware graph O) is generated, representing the occlusion status for each joint. Thus, the first given skeletal structure-based adjacency matrix is replaced by the occlusion-aware graph. Fig. 6 conceptually presents the transformation of the adjacency matrix into the occlusion-aware graph using occlusion information. The left shoulder and right wrist are occluded in a detected 2D pose. The occlusion label c given from Feature-Net is converted to the occlusion matrix C after normalization. Then, the occlusion-weighted adjacency matrix created by element-wise subtraction reduces the self-value and dependency-value of the occluded joint. This is demonstrated in the dependency-weight (7, 8) and self-weight (8, 8): the reliability of the seventh joint is weighted to estimate the obscured eighth joint. In summary, the occlusion-aware graph enhances occluded joint reconstruction by reducing self-dependency and leveraging emphasizing attention to reliable neighbors, particularly benefiting limb joints.

4.3. Network details

Fig. 7 illustrates the detailed structure of each network of the proposed method. Feature-Net conducts two tasks, 2D HPE, and occluded joint detection, using the stacked Hourglass network as the backbone. The number of joints representing the human pose is K. For 2D HPE, a 2D heatmap with 64×64 resolution with K channels is extracted via the final 2D convolution of the upper branch of Feature-Net. The heatmaps are transformed into a 2D pose through an integral function layer. For occlusion detection, the occlusion label is extracted through a series of fully connected layers. The occlusion information is obtained using the Sigmoid activation of final features. The 2D pose and occlusion information are expressed as 1D binary vectors of T frames stacked into a single dimension. This approach has long been a convention and used in previous 2D-to-3D lifting-based 3D HPE studies. However, to further enforce temporal consistency, we add temporal positional encoding to the stacked feature vectors in a manner similar to Transformers [14]. Let the input 2D pose feature be $f \in R^{T \times KD}$ where T, K, D are the number of frames, joints, and feature dimensions, respectively. We added the temporal position encoding to the input feature, resulting in the updated feature $\tilde{f}$ as follows:

$$\tilde{f} = f + PE, \quad PE(t,i) = \begin{cases} \sin\left(\frac{t}{10000^{2i/KD}}\right), & \text{if } i \text{ is even} \\ \cos\left(\frac{t}{10000^{2i/KD}}\right), & \text{if } i \text{ is odd} \end{cases} \tag{10}$$

where t and i are the indices of the frame and spatial feature dimension.

In addition, Lifting-Net comprises 19 GCNs ($9 \times 2 + 1$) in series. The top side GCN estimates the 3D pose based on the position-aware graph which provides the skeletal structure, including the joint distance. The bottom side estimates the 3D pose based on the occlusion-aware graph, which controls the defects of the input features according to the obscured joints. The output of the two branches of GCNs is combined and processed in a final GCN layer, obtaining the 3D pose from the final GCN output. The 3D pose loss is as follows:

$$\mathcal{L}_{3D} = \sum_{i=1}^{E} \left\| B_i - B_i^{GT} \right\|^2 + \sum_{i=1}^{K} \left\| Y_i - Y_i^{GT} \right\|^2, \tag{11}$$

where B, E, and Y are the 3D limb length, number of limbs, and 3D joint location, respectively. Finally, while the architecture of DG-CNN is designed to consider both structure and sensitivity, each branch may not be learned as expected without direct guidance to the model. Thus, we add a novel similarity loss by measuring the similarity between the output and ground-truth 3D pose using our PSIM$^+$ metric. This provides direct supervision and enables the model to estimate poses similar to the ground-truth with both structure and sensitivity awareness. The final loss function for Lifting-Net is:

$$\mathcal{L}_{Lift} = \omega_{3D}\mathcal{L}_{3D} + \omega_{sim}\mathcal{L}_{sim}, \tag{12}$$

where $\mathcal{L}_{sim}$, ω_{3D}, ω_{sim} are the similarity loss and weights for each loss term.

5. Experiments

5.1. Datasets and evaluation metrics

The following datasets were used for model training and evaluation.
Human3.6M [3]. The Human3.6M dataset consists of 3.6 million images with paired 3D poses and has been widely used as the main benchmark for 3D pose estimation. The dataset consists of 11 subjects with 15 actions captured from four camera views. We followed a major protocol for performance comparison which uses cross subject-based validation with five subjects (S1, S5, S6, S7, and S8) for training and two subjects (S9 and S11) for testing.
MPI-INF-3DHP [9]. Unlike Human3.6M, this dataset consists of images with indoor and outdoor backgrounds paired with 3D poses captured by motion capture. In addition, it contains sequences with diverse actions performed by eight subjects (four males and four females). The dataset was used for both training and testing, following the split based on the

Table 1
Comparison between various metrics on Pose Similarity Validation Dataset.

Type of Dataset	SET A [4]				SET B			
Metrics	PLCC [↑]		SROCC [↑]		PLCC [↑]		SROCC [↑]	
	Spatial	Temporal	Spatial	Temporal	Spatial	Temporal	Spatial	Temporal
MPJPE	0.7928	0.6319	0.8154	0.6239	0.7521	0.6119	0.7965	0.6004
P-MPJPE	0.8299	0.6732	0.8253	0.6430	0.7635	0.6954	0.8033	0.6267
3D PCK	0.7530	0.5523	0.7328	0.5839	0.7513	0.5499	0.7340	0.5895
PSIM [4]	0.8518	0.7628	0.8735	0.7911	0.7988	0.7501	0.8246	0.7540
PSIM$^+$ (P)	0.7866	0.6232	0.7916	0.6119	0.7611	0.6098	0.7548	0.5966
PSIM$^+$ (P+S)	0.8430	0.6852	0.8765	0.7316	0.7835	0.6710	0.8168	0.7150
PSIM$^+$ (avg. pool)	**0.8695**	0.7039	**0.8781**	0.8010	**0.8661**	0.7054	**0.8843**	0.8108
PSIM$^+$ (ours)	**0.8695**	**0.7548**	**0.8781**	**0.8155**	**0.8661**	**0.7608**	**0.8843**	**0.8274**

experimental setup of the recent state-of-the-art method [13] for a fair comparison.

Occlusion Analysis Dataset. We constructed a new 3D pose dataset designed to analyze the effects and limitations of 3D HPE under diverse occlusion scenarios. Eight subjects with varying body shapes performed free-form actions in a motion capture studio, resulting in a wide range of occlusions due to diverse body shapes, motions, and camera viewpoints. Occlusions were categorized into three levels of severity: joint occlusion (1–4 joints occluded), body part occlusion (up to two body parts covered), and half-body occlusion (three or more body parts occluded, with upper or lower body folded). Each category contains 35,000 RGB images paired with 3D pose annotations.

Pose Quality Validation Dataset. As no existing dataset supports similarity verification under self-occlusion, we constructed the Pose Quality Validation Dataset, following the validation protocol used in image quality assessment [25,32,33]. The dataset contains pairs of ground-truth and predicted poses, each annotated with MOSs. Predicted poses were generated using five methods: Zhao *et al.*[11], Pavllo *et al.*[2], Cai *et al.*[12], Lee *et al.*[4], and our DG-CNN. Two subsets were used for validation: a Human3.6M-centered set based Human 3.6M (2,000 image-pose pairs), and a self-occlusion-centered set built from the Occlusion Analysis Dataset (300 image-pose pairs). Each sample consists of 120 consecutive frames, and subjective pose quality measurement (PQM) was conducted to obtain MOS annotations.

The following metrics are used to report the performance of our method. **PLCC, SROCC.** Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SROCC) are used to measure the correlation between the subjective MOS and similarity metric values. A correlation value closer to 1 for both indicators indicates higher correlation.

MPJPE, P-MPJPE. These metrics are conventionally used to measure the accuracy of 3D pose estimation models. They both measure the Euclidean distance between the GT and predicted poses while P-MPJPE measures the distance after conducting Procrustes alignment [8].

PCK, AUC. These metrics are used to further measure 3D pose accuracy on databases such as MPI-INF-3DHP. PCK measures the correctness of 3D joint predictions within a certain threshold (e.g. 30mm) while AUC measures the area under the curve formed by varying thresholds of PCK.

PSIM, PSIM$^+$. These are recent metrics used to measure the 3D pose accuracy based on perceptual similarity [4]. As these metrics are designed to reflect human-perceived visual quality, it is used as a measure of the perceptual quality of predicted 3D poses.

5.2. Implementation details

To train the proposed DG-CNN, Feature-Net was first trained independently, followed by training Lifting-Net with the weights of Feature-Net kept fixed. Feature-Net was trained on the MPII and Human3.6M datasets, using both 2D pose annotations and occlusion labels. For datasets that provide occlusion annotations (e.g., MPII), we used the provided labels directly. For datasets without such annotations, we adopted the labeling method from [20]. A simplified 3D cylinder model was fit between adjacent joints, projected onto the 2D image plane, and a joint was labeled as occluded if it intersected with any projected cylinder, indicating potential self-occlusion or overlap. Lifting-Net was trained separately on Human3.6M and MPI-INF-3DHP for evaluation, following the standard protocols described in Section 5.1. For cross-dataset evaluation on MPI-INF-3DHP and the Occlusion Analysis Dataset, we used the model trained on Human3.6M without additional fine-tuning.

The following settings were used for PSIM$^+$ and DG-CNN. In Eq. (4), the scale weights for the pose similarity metric were set to $w_{\mathbf{p}} = 0.6$, $w_{\mathbf{s}} = 0.15$, and $w_{\mathbf{v}} = 0.25$. In the occlusion loss of Eq. (7), α and γ were set to 1 and 2, respectively. Since the classification loss from Feature-Net was relatively large, we applied scaling to balance it with other objectives: $\omega_{2D} = 0.999$, $\omega_{OC} = 0.001$ in Eq. (8). We exploited 40 consecutive frames with a dilation value of 2 as one input set for Lifting-Net. As the 3D pose loss measures similar objective to positional similarity in PSIM$^+$, we have reduced it scale: $\omega_{3D} = 0.1$, $\omega_{sim} = 0.9$ in Eq. (12). For optimization, we used RMSProp[34] with a learning rate $lr = 1e^{-3}$, $\alpha = 0.9$, $\epsilon = 1e^{-8}$, and $\omega_{decay} = 1e^{-4}$. Training was stabilized using a 5-epoch warmup followed by a ReduceLROnPlateau scheduler based on validation loss.

5.3. Pose similarity metric

Fig. 8 illustrates the correlation as a scatter plot between subjective and objective scores in the two validation datasets. The two columns on the right of Fig. 8 are comparisons in the Human3.6M-centered dataset [4], and the other two columns are comparisons in the self-occlusion-centered dataset. We selected three objective metrics in this visualization: MPJPE, PSIM, and PSIM$^+$. All scores were normalized from 0 to 1 for a fair comparison. In addition, the logistic fitting curve (dotted red line) presents the linearity of the correlation, and the ellipse (solid line) expresses the estimated distribution. In this figure, both PSIM and PSIM$^+$ are more consistent than MPJPE. In particular, in the temporal domain, the correlation between the subjective score and MPJPE, which cannot catch joint tremors and protrusions, is significantly reduced. Moreover, both PSIM and PSIM$^+$ show a fairly consistent correlation with both domains. In the self-occlusion-centered dataset, PSIM$^+$ shows an exceptionally excellent correlation.

Table 1 details the quantitative correlations between various existing and proposed metrics. Parenthesized notations *P*, *S*, and *avg. pooling* in *PSIM$^+$* denote the position, structural similarities, and simple average pooling, respectively. In the temporal results, the reference metrics apply an average pooling scheme. Sets A and B indicate Human3.6M-centered [4] and self-occlusion-centered datasets, respectively. While both PSIM and PSIM$^+$ exhibit competitively good correlations in Set A, in Set B, the proposed metric displays remarkably excellent performance compared to the existing metrics. Finally, the PSIM$^+$ predicts posture quality better than other metrics, regardless of the domain and dataset.

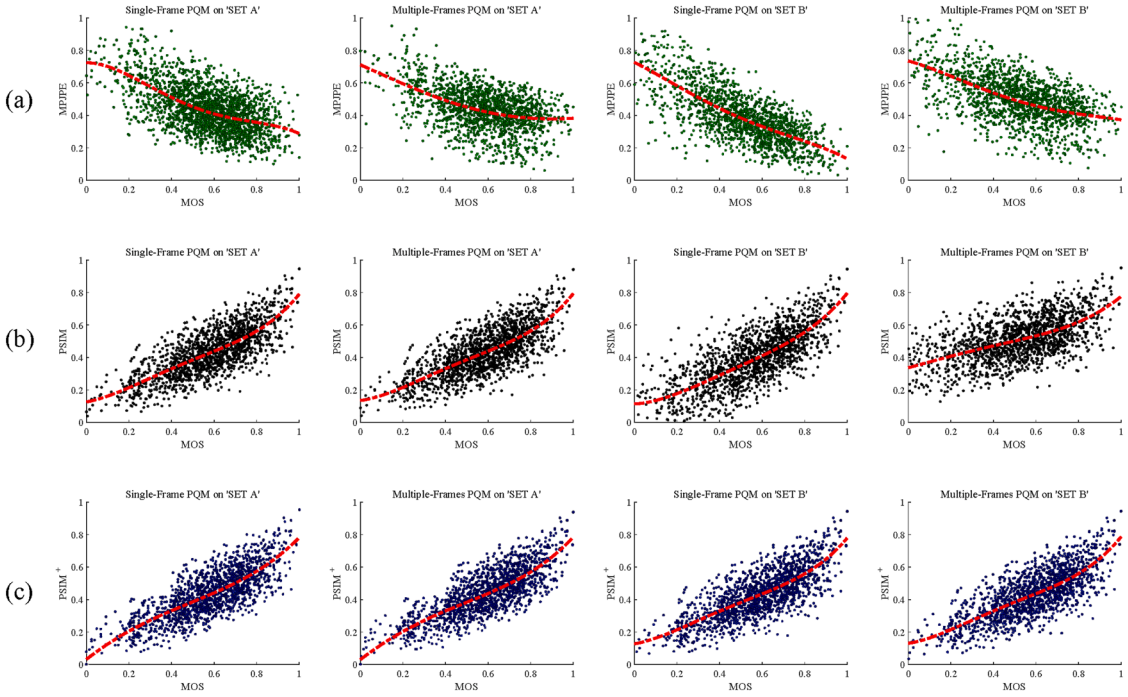

Fig. 8. Normalized correlation scatter plots between the similarity metrics and MOS on various similarity validation datasets: (a) MPJE, (b) PSIM, and (c) PSIM$^+$.

Table 2
Performance Comparison with Action Scenarios on Human3.6M.

Method		Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	MPJPE	P-MPJPE	PSIM$^+$
Zhao et al. [11]*	(SH)	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	47.7	0.667
Lee et al. [1]	(SH)	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	44.8	0.692
Han et al. [21]*	(Im)	50.2	54.4	50.1	52.0	53.6	57.1	46.7	50.1	61.5	65.1	52.2	49.2	54.5	41.2	46.6	52.8	–	–
Cai et al. [12]	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	48.7	39.0	0.757	
Pavllo et al. [2]	(CPN)	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	36.5	0.763
Li et al. [14]	(CPN)	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	43.8	30.5	–
Zhang et al. [15]	(CPN)	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8	30.6	–
Lee et al. [4]	(SH)	35.6	40.6	40.4	41.2	40.7	51.0	42.6	36.5	50.8	53.1	43.2	41.6	35.6	31.4	28.8	40.5	31.8	0.790
Zhu et al. [5]	(SH)	36.3	38.7	38.6	33.6	42.1	50.1	36.2	35.7	50.1	56.6	41.3	37.4	37.7	25.6	26.5	39.2	32.9	0.788
Yu et al. [13]	(CPN)	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4	34.8	0.771
Lang et al. [17]	(CPN)	36.1	38.2	37.9	32.1	41.6	45.1	30.5	38.2	47.2	54.7	40.1	36.3	35.6	26.7	27.2	37.8	28.3	0.812
Peng et al. [16]	(CPN)	30.1	32.1	29.1	30.6	35.4	39.3	32.8	30.9	43.1	45.5	34.7	33.2	32.7	22.1	23.0	33.0	26.2	0.869
DG-CNN†	(CPN)	35.5	37.4	39.6	38.0	39.3	47.0	37.0	36.1	45.8	46.1	39.4	36.4	33.4	32.8	32.9	**38.9**	**31.2**	**0.776**
DG-CNN	(SH)	32.4	34.3	36.6	34.9	36.1	43.8	34.0	33.1	42.6	42.9	36.3	33.3	30.2	29.7	29.8	**35.8**	**27.7**	**0.874**
DG-CNN	(GT)	28.2	30.5	27.1	31.3	33.8	34.2	34.0	27.4	27.3	32.8	28.1	30.3	29.8	26.6	25.1	**30.1**	**24.8**	**0.903**

5.4. 3D Pose estimation

We used the Human3.6M dataset for evaluating 3D HPE performance as shown in Table 2. To ensure a fair comparison, methods that use multi-view input images were excluded. The * symbol denotes methods that estimate 3D poses from a single image. For 2D input data, (SH) indicates poses predicted by the Stacked Hourglass network [28], (CPN) by the Cascaded Pyramid Network [35], and (Im) refers to direct image input without intermediate 2D pose estimation. Models with (GT) are trained with ground-truth 2D poses for comparison, offering an upper bound on performance by leveraging precise input data. The † symbol in DG-CNN indicates 2D pose used from external model, and occlusion label used from Feature-Net. Our full model is denoted as DG-CNN. The performance comparison between previous methods are shown in Table 2. The values expressed in all action scenarios are MPJPE, and the values in the MPJPE column indicate the average error for all scenarios. We also report PSIM$^+$ values to validate the perceptual quality of the estimated 3D poses.

The proposed method, DG-CNN, achieves a performance improvement of 0.6 *mm* over the state-of-the-art [17], and remains comparable even to a recent method trained with augmented data [16], with a difference of 1.5 *mm*. In PSIM$^+$, we observed consistent performance improvements across all methods, demonstrating the effectiveness of our perceptually motivated design and incorporation of similarity loss. An interesting observation is that our model exhibits only a minimal performance gap (< 3 *mm* P-MPJPE) between estimated and ground-truth 2D inputs, suggesting strong robustness to noisy input 2D poses. Table 3 presents the results on the MPI-INF-3DHP dataset. The upper part of the table presents cross-dataset evaluation using the model trained on Human3.6M, while methods marked with * in the lower part follow the evaluation protocol of [13]. In line with the results observed on the Human3.6M dataset, the proposed method achieves comparable performance to state-of-the-art methods in conventional metrics (PCK, AUC, and MPJPE), even without using ground-truth 2D poses as input. Importantly, it demonstrates significantly stronger performance in perceptual metrics, including PSIM and PSIM$^+$.

Table 3
Performance Comparison on MPI-INF-3DHP.

Method		PCK[↑]	AUC[↑]	MPJPE[↓]	**PSIM**[4][↑]	**PSIM**$^{+}$[↑]
Zhao et al. [11]	(SH)	71.0	33.6	104.2	0.489	0.477
Pavllo et al. [2]	(GT)	73.4	34.0	100.9	0.512	0.493
Lee et al. [4]	(SH)	82.6	48.8	81.3	0.583	0.579
DG-CNN	(SH)	**84.5**	**51.8**	**74.1**	**0.668**	**0.671**
Yu et al. [13]*	(GT)	98.5	79.1	27.7	0.858	0.843
Lang et al. [17]*	(Im)	98.5	84.7	17.9	0.909	0.875
Peng et al. [16]*	(GT)	**98.9**	**85.9**	**16.7**	0.915	0.899
DG-CNN*	(SH)	98.7	84.8	18.2	**0.926**	**0.911**

5.5. Ablation studies

The ablation study results are presented in Table 4. The first group evaluates our model architecture. We first introduced a single-graph baseline, referred to as the Single Graph Convolutional Neural Network (SG-CNN), with three variants: Baseline (BL), Position-Aware (PA), and Occlusion-Aware (OA). The BL corresponds to a standard graph convolutional network, while the PA and OA variants individually incorporate our position-aware graph structure and occlusion-aware branch, respectively. DG-CNN† denotes the dual-graph based approach, with three variants: without Occlusion Labels (OL), with Temporal Positional Encoding (TP), and with Similarity Loss (SL). The OL variant excludes occlusion labels from Lifting-Net, while the TP and SL variants individually incorporate the proposed temporal positional encoding and the PSIM^{+}-based similarity loss, respectively.

Results show that DG-CNN outperforms all SG-CNN variants, confirming the benefit of dual-graph modeling of both position- and occlusion-aware graphs. DG-CNN†(OL) performs worse than SG-CNN (OA), suggesting that incorporating information on visually sensitive joints has a greater impact than the additional architectural complexity of dual-graph modeling. Furthermore, adding temporal and perceptual modules incrementally improves both MPJPE and PSIM^{+}, highlighting the importance of sensitivity-aware supervision and temporal modeling. The second group investigates the effect of temporal input length by varying the number of input frames from 1 to 60. Performance improves with longer sequences, peaking at 40 frames before slightly declining, suggesting a balance between context and redundancy. The third group explores different configurations of the combined loss function. We ablate the similarity loss components $\mathcal{L}_{\text{sim}}(w_{\text{p}}, w_{\text{s}}, w_{\text{v}})$ by varying their weights, along with different ratios between $\mathcal{L}_{3D}$ and $\mathcal{L}_{\text{sim}}$. The term $\mathcal{L}_{\text{sim}}$ proves especially effective in penalizing errors in visually sensitive joints, leading to improved perceptual quality. Additionally, assigning higher weight to the similarity loss consistently increases PSIM^{+} while maintaining competitive MPJPE. These results highlight the importance of similarity loss in enhancing perceptual quality without compromising numerical accuracy.

For model efficiency analysis, the proposed DG-CNN module contains approximately 1.5 million parameters with a computational cost of 2.5 GFLOPs. In contrast, the 2D backbone, based on an occlusion-aware Stacked Hourglass network, accounts for 32 million parameters and approximately 45 GFLOPs. When combined, the full model has a total of approximately 35 million parameters and a computational cost of 51 GFLOPs. This overall model size is comparable to ResNet-50 to ResNet-101 and is roughly half the size of a typical Vision Transformer (ViT), demonstrating the model's efficiency relative to its strong performance and occlusion robustness.

5.6. Occlusion analysis

We conducted experiments on the Occlusion Analysis Dataset to evaluate the robustness of 3D HPE methods under conditions where visually sensitive joints are frequently observed. Table 5 compares various models under these conditions. While existing methods perform reasonably under joint-level occlusion, their accuracy deteriorates significantly with larger occlusions. For example, the model in [5] performs well on Human3.6M but suffers under occlusion, and even recent methods [13,16] show substantial degradation. In contrast, our method consistently outperforms all baselines across occlusion levels, demonstrating superior robustness through explicit modeling of visually sensitive joints. Finally, Fig. 9 presents the qualitative results on various occlusion cases. The 3D poses predicted from the proposed model showed qualitatively fair results in both joint and body part occlusion cases. Similar to quantitative results, our method reconstructs abnormal 3D human poses in the half-body occlusion.

Table 4
Ablation Studies on Human3.6M.

Model Arch.	MPJPE	P-MPJPE	PSIM +	Window Size	MPJPE	P-MPJPE	PSIM +	Loss Configuration	MPJPE	P-MPJPE	PSIM +
SG-CNN (BL)	59.6	50.6	0.532	*DG-CNN* (1)	45.4	36.8	0.739	$\mathcal{L}_{Lift} = \mathcal{L}_{sim}(0.5, 0.5, 0.1)$	50.8	41.1	0.745
SG-CNN (PA)	48.4	38.3	0.637	*DG-CNN* (10)	38.5	32.1	0.765	$\mathcal{L}_{Lift} = \mathcal{L}_{sim}(0.3, 0.3, 0.3)$	48.6	40.8	0.739
SG-CNN (OA)	42.6	34.2	0.736	*DG-CNN* (20)	37.8	30.4	0.799	$\mathcal{L}_{Lift} = \mathcal{L}_{sim}(0.6, 0.15, 0.25)$	46.9	39.9	0.760
DG-CNN† (OL)	48.5	38.4	0.635	*DG-CNN* (30)	37.4	29.1	0.813	$\mathcal{L}_{Lift} = 0.5\mathcal{L}_{3D} + 0.5\mathcal{L}_{sim}$	40.2	31.5	0.841
DG-CNN† (TP)	36.8	28.1	0.851	*DG-CNN* (40)	**35.8**	**27.7**	**0.874**	$\mathcal{L}_{Lift} = 0.3\mathcal{L}_{3D} + 0.7\mathcal{L}_{sim}$	36.6	28.3	0.858
DG-CNN† (SL)	36.6	28.3	0.858	*DG-CNN* (50)	37.1	28.5	0.832	$\mathcal{L}_{feat} = 0.2\mathcal{L}_{3D} + 0.8\mathcal{L}_{sim}$	37.8	28.0	0.861
DG-CNN	**35.8**	**27.7**	**0.874**	*DG-CNN* (60)	37.3	28.9	0.825	$\mathcal{L}_{Lift} = 0.1\mathcal{L}_{3D} + 0.9\mathcal{L}_{sim}$	**35.8**	**27.7**	**0.874**

Table 5
Performance Comparison on Occlusion Analysis Dataset.

Method		Occlusion Case											
		Joint			Body part			Half body			Average		
		MPJPE	PSIM[4]	PSIM^{+}	MPJPE	PSIM[4]	PSIM^{+}	MPJPE	PSIM[4]	PSIM^{+}	MPJPE	PSIM[4]	PSIM^{+}
Zhao et al. [11]	(SH)	32.1	0.883	0.835	68.1	0.529	0.483	118.6	0.378	0.332	74.1	0.596	0.464
Pavllo et al. [2]	(CPN)	30.8	0.898	0.841	60.3	0.568	0.532	109.3	0.436	0.342	68.3	0.634	0.491
Cai et al. [12]	(CPN)	31.7	0.877	0.836	62.6	0.541	0.558	104.7	0.465	0.350	65.2	0.627	0.513
Lee et al. [4]	(SH)	31.5	0.875	0.859	58.3	0.571	0.596	96.9	0.483	0.361	62.2	0.643	0.605
Zhu et al. [5]	(SH)	32.1	0.868	0.848	59.6	0.563	0.581	98.9	0.467	0.363	63.5	0.632	0.597
Yu et al. [13]	(CPN)	31.8	0.861	0.849	57.9	0.570	0.577	96.5	0.487	0.381	62.1	0.639	0.602
Peng et al. [16]	(CPN)	31.4	0.870	0.856	54.7	0.575	0.593	94.8	0.484	0.397	60.3	0.643	0.616
DG-CNN	(SH)	30.3	0.881	0.863	52.8	0.594	0.641	82.5	0.510	0.434	56.2	0.661	0.646

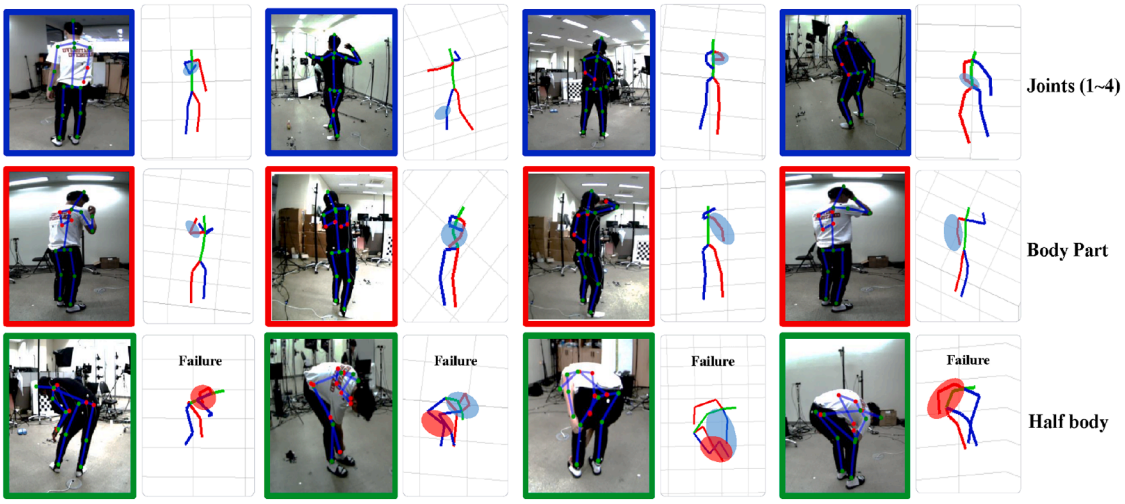

Fig. 9. Qualitative results on the Occlusion Analysis Dataset. The red point marks an occluded joint, while blue and red circles indicate the best and worst reconstructed joints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

In this study, we investigated the significance of structure and similarity in 3D human pose similarity quantification and estimation. We proposed a novel enhanced similarity metric and estimation model for 3D HPE model building on the discovery that structure and sensitivity are key elements in 3D human pose perception. We developed PSIM$^+$ to explicitly measure these components and designed the position-aware and occlusion-aware branches within the DG-CNN framework. Furthermore, by using PSIM$^+$ as a similarity loss, we were able to incorporate structure and sensitivity awareness both implicitly and explicitly into the DG-CNN model. By integrating fundamental elements from the perspective of human perception, we significantly improved both the quantitative and qualitative performance of the similarity metric and the estimation model. Nonetheless, our method still faces limitations in extreme occlusion scenarios, ambiguous poses with minimal visual cues, and in-the-wild settings where occlusion labels are unavailable or difficult to obtain. Incorporating stronger priors, diffusion-based augmentation, or multi-view cues may be necessary to address such challenging cases, remaining as important avenues for future work.

CRediT authorship contribution statement

Kyoungoh Lee: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization; **Jungwoo Huh:** Writing – review & editing, Validation, Resources, Investigation, Data curation; **Jiwoo Kang:** Supervision; **Sanghoon Lee:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Data availability

The data that has been used is confidential.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the Culture, Sports and Tourism Research and Development Program through Korea Creative Content Agency Grant funded by the Ministry of Culture, Sports and Tourism in 2024 under Grant RS-2024-00398413, Contribution Rate: 90%; in part by the National Research Foundation of Korea (NRF) Grant through Korean Government [Ministry of Science and ICT (MSIT)] under Grant RS-2025-02216328, Contribution Rate: 10%; and in part by the Yonsei Signature Research Cluster Program of 2025 under Grant 2025-22-0013.

References

[1] K. Lee, I. Lee, S. Lee, Propagating lstm: 3d pose estimation based on joint interdependency, in: Proc. Eur. Conf. Comput. Vis, Eur. Conf. Comput. Vis, 2018, pp. 119–135.
[2] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2019, pp. 7753–7762.
[3] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6m: large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2013) 1325–1339.
[4] K. Lee, W. Kim, S. Lee, From human pose similarity metric to 3d human pose estimator: temporal propagating lstm networks, IEEE Trans. Pattern Anal. Mach. Intell. 2022.
[5] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, Y. Wang, Motionbert: a unified perspective on learning human motion representations, in: Proc. IEEE Int. Conf. Comput. Vis, IEEE Int. Conf. Comput. Vis, 2023.
[6] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, W. Zhang, Saliency-guided quality assessment of screen content images, IEEE Trans. Multimed. 18 (6) (2016) 1098–1110.
[7] H. Kim, S. Lee, A.C. Bovik, Saliency prediction on stereoscopic videos, IEEE Trans. Image Process. 23 (4) (2014) 1476–1490.
[8] J.C. Gower, Generalized procrustes analysis, Psychometrika 40 (1) (1975) 33–51.
[9] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3d human pose estimation in the wild using improved cnn supervision, in: Proc. Int. Conf. on 3D Vis, Int. Conf. on 3D Vis, 2017, pp. 506–516. 3DV.
[10] Y. Bin, Z.-M. Chen, X.-S. Wei, X. Chen, C. Gao, N. Sang, Structure-aware human pose estimation with graph convolutional networks, Pattern Recognit. 106 (2020) 107410.
[11] L. Zhao, X. Peng, Y. Tian, M. Kapadia, D.N. Metaxas, Semantic graph convolutional networks for 3d human pose regression, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2019, pp. 3425–3435.
[12] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, N.M. Thalmann, Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks, in: Proc. IEEE Int. Conf. Comput. Vis, IEEE Int. Conf. Comput. Vis, 2019, pp. 2272–2281.
[13] B.X. Yu, Z. Zhang, Y. Liu, S.H. Zhong, Y. Liu, C.W. Chen, Gla-gcn: global-local adaptive graph convolutional network for 3d human pose estimation from monocular video, in: Proc. IEEE Int. Conf. Comput. Vis, IEEE Int. Conf. Comput. Vis, 2023, pp. 8818–8829.
[14] W. Li, H. Liu, H. Tang, P. Wang, L.V. Gool, Mhformer: multi-hypothesis transformer for 3d human pose estimation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2022, pp. 13147–13156.
[15] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, Mixste: seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2022, pp. 13232–13242.
[16] J. Peng, Y. Zhou, P. Mok, Ktpformer: kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation, in: Proc. IEEE Conf. Com-

put. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2024, pp. 1123–1132.
[17] B. Lang, M.C. Chuah, Event-guided video transformer for end-to-end 3d human pose estimation, in: IEEE Winter Conf. on Appl. of Comput. Vis, IEEE, 2025, pp. 5114–5124.
[18] Z. Chen, J. Dai, J. Bai, J. Pan, Dgformer: dynamic graph transformer for 3d human pose estimation, Pattern Recognit. 152 (2024) 110446.
[19] W. Li, M. Liu, H. Liu, T. Guo, T. Wang, H. Tang, N. Sebe, Graphmlp: a graph mlp-like architecture for 3d human pose estimation, Pattern Recognit. 158 (2025) 110925.
[20] Y. Cheng, B. Yang, B. Wang, W. Yan, R.T. Tan, Occlusion-aware networks for 3d human pose estimation in video, in: Proc. IEEE Int. Conf. Comput. Vis, IEEE Int. Conf. Comput. Vis, 2019, pp. 723–732.
[21] C. Han, X. Yu, C. Gao, N. Sang, Y. Yang, Single image based 3d human pose estimation via uncertainty learning, Pattern Recognit. 132 (2022) 108934.
[22] W. Li, H. Liu, H. Tang, P. Wang, Multi-hypothesis representation learning for transformer-based 3d human pose estimation, Pattern Recognit. 141 (2023) 109631.
[23] L. Bragagnolo, M. Terreran, D. Allegro, S. Ghidoni, Multi-view pose fusion for occlusion-aware 3d human pose estimation, in: Proc. Eur. Conf. Comput. Vis, Eur. Conf. Comput. Vis, Springer, 2024, pp. 117–133.
[24] L. Zhang, K. Zhou, F. Lu, Z. Li, X. Shao, X.-D. Zhou, Y. Shi, Esmformer: error-aware self-supervised transformer for multi-view 3d human pose estimation, Pattern Recognit. 158 (2025) 110955.
[25] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[26] A. Galdran, T. Araújo, A.M. Mendonça, A. Campilho, Retinal image quality assessment by mean-subtracted contrast-normalized coefficients, in: Proc, null, 2017, pp. 844–853.
[27] L. Zhang, Y. Shen, H. Li, Vsi: a visual saliency-induced index for perceptual image quality assessment, IEEE Trans. Image Process. 23 (10) (2014) 4270–4281.
[28] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Proc. Eur. Conf. Comput. Vis, Eur. Conf. Comput. Vis, 2016, pp. 483–499.
[29] K. Gu, L. Yang, M.B. Mi, A. Yao, Bias-compensated integral regression for human pose estimation, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10687–10702.
[30] J. Wang, S. Huang, X. Wang, D. Tao, Not all parts are created equal: 3d pose estimation by modeling bi-directional dependencies of body parts, in: Proc. IEEE Int. Conf. Comput. Vis, IEEE Int. Conf. Comput. Vis, 2019, pp. 7771–7780.
[31] X. Li, C. Lv, W. Wang, G. Li, L. Yang, J. Yang, Generalized focal loss: towards efficient representation learning for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (3) (2022) 3139–3153.
[32] H. Oh, S. Lee, Visual presence: viewing geometry visual information of uhd s3d entertainment, IEEE Trans. Image Process. 25 (7) (2016) 3358–3371.
[33] W. Kim, S. Lee, A.C. Bovik, Vr sickness *versus* vr presence: a statistical prediction model, IEEE Trans. Image Process. 30 (2020) 559–571.
[34] T. Tieleman, G. Hinton, Lecture 6.5-Rmsprop: divide the gradient by a running average of its recent magnitude, COURSERA: Neural Netw. Mach. Learn., 2012.
[35] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, IEEE Conf. Comput. Vis. Pattern Recognit, 2018, pp. 7103–7112.