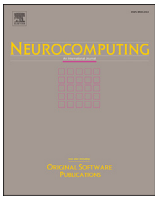
Neurocomputing 681 (2026) 133333

Contents lists available at ScienceDirect

Neurocomputing

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



DDCEFormer: Dual-domain cross enhanced transformer for 3D human pose estimation

Deliang Yang [a], Yanrong Ge [b,*], Ning Xu [b], Rui Shi [c]

[a] School of Mechanical and Electrical Engineering, Beijing Polytechnic College, Beijing, 100042, China
[b] College of Physics, Hebei Normal University, Shijiazhuang, Hebei, 050024, China
[c] School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China

HIGHLIGHTS

- A novel Dual-Domain Cross Enhanced Transformer is proposed for 3D human pose estimation.
- Our EMA mechanism balances global and local features via variable dimensions and CNNs.
- A Cross Enhanced Transformer Block achieves deep fusion of spatial and temporal features.
- The method excels at estimating poses with large limb motions and demonstrates superior generalization.

ARTICLE INFO

Communicated by J. Yu

Keywords:
3D human pose estimation
Cross enhanced transformer
Enhanced multi-head attention
Spatio-temporal feature fusion

ABSTRACT

Estimating 3D human poses from monocular videos remains challenging, especially under large limb movements, due to depth ambiguity and self-occlusion. Existing methods often struggle to simultaneously capture global dependencies and local structural details during spatio-temporal feature modeling, limiting pose estimation accuracy. To address this, we propose a Dual-Domain Cross Enhanced Transformer (DDCEFormer) that enhances spatio-temporal representation by jointly modeling spatial and temporal domain features. Specifically, we design an Enhanced Multi-head Attention (EMA) mechanism that integrates variable-dimensional multi-head attention with convolutional layers to capture long-range global dependencies while reinforcing local structural features among joints. Based on EMA, we construct a Spatial Enhanced Transformer Block (S-ETB) and a Temporal Enhanced Transformer Block (T-ETB) to meticulously model spatial structural relationships and temporal dynamic evolution, respectively. Furthermore, we introduce an Enhanced Multi-head Cross-Attention (EMCA) module and build a Cross Enhanced Transformer Block (C-ETB) to achieve cross-enhancement and deep fusion of spatial and temporal features, thereby balancing global spatio-temporal correlations with local motion details. The model is optimized using a composite loss function consisting of three error components to improve estimation accuracy and robustness. Experimental results show that DDCEFormer achieves MPJPE and P-MPJPE of 39.1mm and 30.8mm on the Human3.6M dataset, and PCK, AUC, and MPJPE of 99.3%, 88.8%, and 13.4mm on the MPI-INF-3DHP dataset, significantly improving overall pose estimation performance and demonstrating superior accuracy, especially for upper-body joints, and exhibits strong generalization capability. The code and model are available at: https://github.com/yangdl8/DDCEFormer.

1. Introduction

Monocular 3D Human Pose Estimation (3D HPE) from video is a fundamental task in computer vision [1–3], with wide applications in action recognition [4], human-computer interaction [5], and augmented/virtual reality [6]. Recovering accurate 3D joint locations from video sequences captures rich spatio-temporal information about human motion, enabling a deeper understanding of actions and interactions. Current mainstream methods typically follow a pipeline for lifting 2D poses to 3D space. However, the inherent depth ambiguity and self-occlusion in monocular 2D observations pose a core challenge for efficient and accurate 3D reconstruction from 2D sequences.

* Corresponding author.
Email address: geyr@hebtu.edu.cn (Y. Ge).

https://doi.org/10.1016/j.neucom.2026.133333
Received 15 January 2026; Received in revised form 24 February 2026; Accepted 10 March 2026
Available online 11 March 2026

To overcome the limitations of single-frame estimation [7], many studies leverage temporal information in videos to enhance prediction stability [8–18]. Early works employed temporal Convolutional Neural Networks (CNNs) to capture motion patterns within fixed receptive fields [8–10]. Other methods utilized Graph Convolutional Networks (GCNs) for feature propagation on spatio-temporal graphs constructed from human skeletons [11,12]. Long Short-Term Memory networks (LSTMs) were also introduced to estimate 3D poses from 2D keypoint sequences [13,14], though their sequence modeling efficiency is limited. Recently, the self-attention mechanism in Transformers has demonstrated strong performance in various vision tasks [19] and has been rapidly applied to 3D HPE, improving estimation accuracy by modeling long-range spatio-temporal dependencies [15–18]. In the context of 3D HPE, Transformer-based models directly model joint sequences and pose representations, capturing global dependencies across joints and time, which effectively alleviates depth ambiguity and temporal inconsistency in human motion.

These video-based methods can be broadly categorized into two architectural types: sequence-to-frame (seq2frame) methods [8–11,14–17] predict the central frame of an input clip to suppress noise and obtain robust outputs, but often introduce redundancy through per-frame computation; in contrast, sequence-to-sequence (seq2seq) methods [12,13,18] reconstruct the entire sequence at once, improving computational efficiency while maintaining pose coherence. Different network architectures exhibit distinct characteristics: LSTMs and Transformers excel at capturing long-term dependencies, while CNNs and GCNs are better at extracting local features. Human motion inherently combines structured local details with global spatio-temporal correlations, making the fusion of multiple architectural advantages a recent research trend [20–25].

However, existing fusion methods still exhibit notable shortcomings: Firstly, inadequate extraction and fusion of spatio-temporal features. Some methods focus on local and global modeling in the temporal dimension while underperforming in spatial feature representation [20,21,25], or they model both global and local spatio-temporal features but fail to fuse them effectively [23]. Secondly, redundant computation. For instance, certain seq2frame methods incur larger estimation errors due to structural design [20,22]. Thirdly, high model complexity, such as models combining multi-head attention and graph convolutions [24]. These issues become especially pronounced under large-amplitude human motions, where inaccurate local modeling or ineffective fusion can significantly degrade pose estimation performance.

In summary, although existing methods have made significant progress in spatio-temporal feature modeling, most fail to balance and effectively fuse global and local information [20,21,23], leading to decreased estimation accuracy during large human motions. This indicates that effective 3D HPE requires not only strong global dependency modeling, as provided by Transformers, but also precise preservation of local joint structures and their coordinated fusion across spatial and temporal domains.

To address these challenges, we introduce a novel Dual-Domain Cross Enhanced Transformer Block (DDC-ETB) to jointly and thoroughly model spatial and temporal domain features in human motion. This module consists of three core sub-modules: The Spatial Enhanced Transformer Block (S-ETB) focuses on learning spatial structural relationships among joints within a single frame; the Temporal Enhanced Transformer Block (T-ETB) models the motion trajectory of the same joint across consecutive frames; and the Cross Enhanced Transformer Block (C-ETB) performs deep fusion of the aforementioned spatio-temporal features. By explicitly decomposing pose modeling into spatial-domain, temporal-domain, and cross-domain enhancement, the proposed design directly addresses the aforementioned limitations of existing Transformer-based approaches. By stacking multiple DDC-ETB modules, we construct an efficient seq2seq architecture—the Dual-Domain Cross Enhanced Transformer (DDCEFormer)—that reduces redundant computation while effectively preserving the temporal coherence of the pose sequence.

To more balancedly represent both global and local spatio-temporal dependencies of human motion, we design an Enhanced Multi-head Attention (EMA) mechanism. EMA innovatively combines variable-dimensional multi-head attention with Convolutional Neural Networks (CNNs). The former adaptively captures long-range global context and controls computational cost, while the latter precisely extracts dynamic structural features within local windows. This design is specifically motivated by the need to enhance joint-level local structures while retaining the global modeling capability of Transformers in 3D HPE. Based on EMA, we build Enhanced Multi-head Self-Attention (EMSA) and Enhanced Multi-head Cross-Attention (EMCA) modules, enabling comprehensive modeling and efficient fusion of spatio-temporal features.

Our contributions are summarized as follows:

(1) **DDCEFormer Model Architecture**: We present a novel 3D HPE architecture, DDCEFormer. This architecture employs S-ETB and T-ETB to learn spatial and temporal features of joint motion, respectively, and leverages C-ETB for cross-fusion of spatio-temporal information, thereby thoroughly modeling the spatio-temporal correlations of joint movements.
(2) **Enhanced Multi-head Attention Mechanism**: We design an Enhanced Multi-head Attention (EMA) mechanism and construct EMSA and EMCA modules based on it. This mechanism combines variable-dimensional multi-head attention with CNNs, employing EMSA to jointly model global and local features, and using EMCA to achieve efficient fusion of spatio-temporal context. It achieves a good balance between model complexity and estimation performance while ensuring representational capacity.
(3) **Experiments and Performance Validation**: A composite loss function comprising three error terms is adopted as the optimization objective to improve estimation accuracy and robustness. Extensive experiments and comparative analyses are conducted on two large-scale datasets, Human3.6M [26] and MPI-INF-3DHP [27]. The results demonstrate that the proposed DDCEFormer method achieves improved accuracy and robustness in 3D HPE.

Notations. The following notations are used in this paper. The superscript T denotes the transpose of a matrix. $R^{n\times m\times p}$ stands for the real vector space of size $n\times m\times p$. FC(·), LN(·), RESHAPE(·), MLP(·), Softmax(·), and Conv2d(·) denote linear transformation, layer normalization, dimension reshaping, multi-layer perceptron, Softmax normalization, and 2D convolution operations, respectively.

2. Related work

2.1. 3D human pose estimation

In monocular video/images, 3D HPE can be categorized into direct estimation methods [28–30] and 2D-to-3D lifting methods [31–33]. Direct estimation methods infer 3D pose directly from 2D images. 2D-to-3D lifting methods first obtain 2D joint locations using a pre-trained 2D pose detector, then feed these locations into a 2D-to-3D lifting network to complete 3D pose estimation. Benefiting from rapidly developed 2D human pose estimation algorithms such as SHN [34], CPN [35], AlphaPose [36], and HRNet [37], 2D-to-3D lifting methods can perform 3D pose estimation efficiently and accurately. This paper uses monocular video as input and follows the 2D-to-3D lifting paradigm.

2.2. Traditional neural network-based methods

Zhou et al. [38] first trained a deep fully convolutional network to predict uncertainty maps for 2D joint locations, then used a proposed Expectation-Maximization algorithm to complete 2D pose estimation. Pavlakos et al. [39] used a CNN for end-to-end training, taking color images as input and directly outputting 3D pose information. For 3D pose estimation from a single RGB image, temporal information from consecutive video frames helps improve accuracy and robustness. To effectively

utilize temporal information in videos, Pavllo et al. [10] employed temporal CNNs to capture global dependencies between adjacent frames. Liu et al. [9] introduced attention mechanisms focusing on temporal context to adaptively identify key frames and the tensor output of each DNN layer for more accurate estimation. Existing CNN-based methods primarily rely on convolution operations to model temporal information, which has limitations in capturing global relationships in human motion. To obtain richer semantic features, stacking multiple convolutional layers is needed to enlarge the CNN's receptive field. Although dilated temporal CNNs can capture global dependencies, their internal connectivity is still somewhat limited.

To address spatial dependencies and temporal consistency, Dabral et al. [40] introduced bone length constraints in temporal networks. Chen et al. [8] constrained human structure via bone orientation and length to ensure the temporal consistency of human anatomy in videos. Cai et al. [11] incorporated human structural priors into graph convolutional networks (GCNs). Xu et al. [41] proposed graph-stacked hourglass networks for multi-scale human skeleton representation. Wang et al. [12] proposed a U-shaped graph convolutional network (UGCN) to capture both short-term and long-term motion information. Hu et al. [42] further proposed a spatial-temporal conditional directed graph convolution, constructing a U-shaped conditional directed graph convolutional network. Hossain et al. [13] proposed a seq2seq network composed of layer-normalized LSTM units. These traditional neural network-based methods still have limitations in capturing long-term dependencies. This paper uses the Transformer model to better capture global temporal correlations due to its stronger global relational capacity. Additionally, we design an enhanced multi-head self-attention block that fuses variable-dimensional multi-head self-attention with CNN convolutions, facilitating simultaneous focus on global and local spatio-temporal correlations.

2.3. Transformer-based methods

The Transformer model, with its powerful self-attention mechanism, has been introduced to 3D HPE [43]. Zheng et al. proposed PoseFormer [15], first using separate spatial and temporal Transformers to model joint correlations in different dimensions. However, this method overlooks motion differences among joints, leading to insufficient spatio-temporal correlation learning. Li et al. proposed StridedFormer [16], which replaces fully connected layers in the Transformer encoder's feed-forward network with strided convolutions to gradually shorten sequence length, simply and effectively mapping long sequences of 2D joint positions to a single 3D pose. Einfalt et al. [44] used masked token modeling in Transformers for upsampling temporal sequence representations, decoupling the sampling rate of input 2D poses from the video's target frame rate to reduce overall computational complexity.

Since motion patterns of different body parts are inconsistent, Xue et al. [21] proposed a part-aware temporal attention module to extract temporal dependencies for each part separately. Shan et al. [45] proposed a pre-trained spatial-temporal many-to-one model to reduce the difficulty of capturing spatial and temporal information. To address ambiguity and occlusion, Li et al. proposed MHFormer [46], learning multiple plausible pose hypotheses to synthesize more accurate 3D poses. Li et al. further improved MHFormer++ [17], replacing the standard Transformer encoder in the original MHG module with a graph Transformer encoder to better constrain the spatial structure of human joints. These seq2frame methods require the repeated input of large, overlapping 2D keypoint sequences to obtain 3D poses for all frames. Zhang et al. proposed the seq2seq MixSTE method [18], which employs a Transformer-based mixed spatio-temporal encoder to alternately obtain spatial and temporal feature encodings for joints, improving 3D pose reconstruction accuracy. Wang et al. proposed a DBSCAN-based clustering module [47] to detect noisy temporal features and designed an adjacency matrix masking mechanism to suppress their influence.

To simultaneously focus on local spatio-temporal correlations for each joint, Wang et al. proposed a global-local spatio-temporal encoder [20]. Li et al. proposed GraphMLP [22], integrating the graph structure of the human skeleton into an MLP model to allow local and global spatial interactions. Tang et al. proposed STCFormer [23], decoupling space and time to build parallel spatio-temporal criss-cross attention blocks, while embedding spatio-temporal convolutions of adjacent joints and joint grouping information into the attention blocks. Liu et al. proposed STGFormer [24] based on a spatio-temporal interlaced graph attention mechanism, directly integrating graph information into corresponding attention layers to learn long-range dependencies across time and space. Zhong et al. proposed FMFormer [25], which includes a frame padding preprocessing step and a multi-scale temporal transformer backbone to effectively establish temporal dependencies. Diaz-Arias et al. proposed ConvFormer [48], which leverages a dynamic multi-head convolutional self-attention mechanism to achieve 3D human pose estimation of the central frame from monocular videos. Li et al. proposed UniFormer [49], which integrates the advantages of convolution and self-attention for image and video recognition and classification.

Existing Transformer methods excel at capturing global dependencies among joints but still have limitations in precisely modeling local dependencies. Moreover, these methods lack an in-depth consideration of fusing global and local spatio-temporal features, leading to significant deviations between estimates and ground truth. To address this, this paper introduces a novel architecture where parallel S-ETB and T-ETB extract spatial and temporal features of joints, respectively, and fuse them via C-ETB, enabling cross-focus on local and global spatio-temporal information, effectively improving pose estimation accuracy and robustness.

3. DDCEFormer method for 3D HPE

We present the proposed DDCEFormer model, which is designed to reconstruct 3D human poses from input 2D pose sequences. First, the overall framework is reviewed, followed by detailed descriptions of key modules.

3.1. Overall architecture

The overall architecture of DDCEFormer is illustrated in Fig. 1(a). It takes a 2D pose sequence as input, performs spatio-temporal feature extraction and fusion, and finally outputs a 3D pose sequence. The DDCEFormer consists of the following components: (i) **Backbone Module**: A DDC-ETB, which is composed of three sub-modules: S-ETB, T-ETB, and C-ETB. (ii) **Auxiliary Modules**: Linear Mapping, Spatial Position Embedding (SPE), Temporal Position Embedding (TPE), and a Regression Head. (iii) **Dimension Transformation**: A RESHAPE operation.

The specific workflow is as follows. First, the Linear Mapping module projects the original 2D pose sequence $X \in R^{T\times N\times 2}$ into a high-dimensional feature space. Next, the SPE and TPE modules embed spatial position information of joints and temporal position information per frame into the high-dimensional features, generating the initial feature representation $X_{\mathrm{I}} \in R^{T\times N\times d_m}$. This feature is then fed into L stacked layers of DDC-ETB modules for deep feature extraction and fusion, yielding the output feature representation $X_{\mathrm{O}} \in R^{T\times N\times d_m}$. The DDC-ETB module is constructed by the parallel connection of S-ETB and T-ETB, followed by a serial connection with C-ETB.

Here, the S-ETB module encodes correlations among joints within the spatial domain, producing spatial features $X_{\mathrm{OS}} \in R^{T\times N\times d_m}$. The T-ETB module captures temporal correlations in the time domain, producing temporal features $X_{\mathrm{OT}} \in R^{N\times T\times d_m}$. The C-ETB module fuses these spatio-temporal features, generating a unified spatio-temporal feature representation $X_{\mathrm{OC}} \in R^{N\times T\times d_m}$. The RESHAPE operation is used for tensor dimension transformation during processing. This stacked parallel architecture aims to enhance the coherence of spatio-temporal feature encoding. Finally, the Regression Head module performs regression on

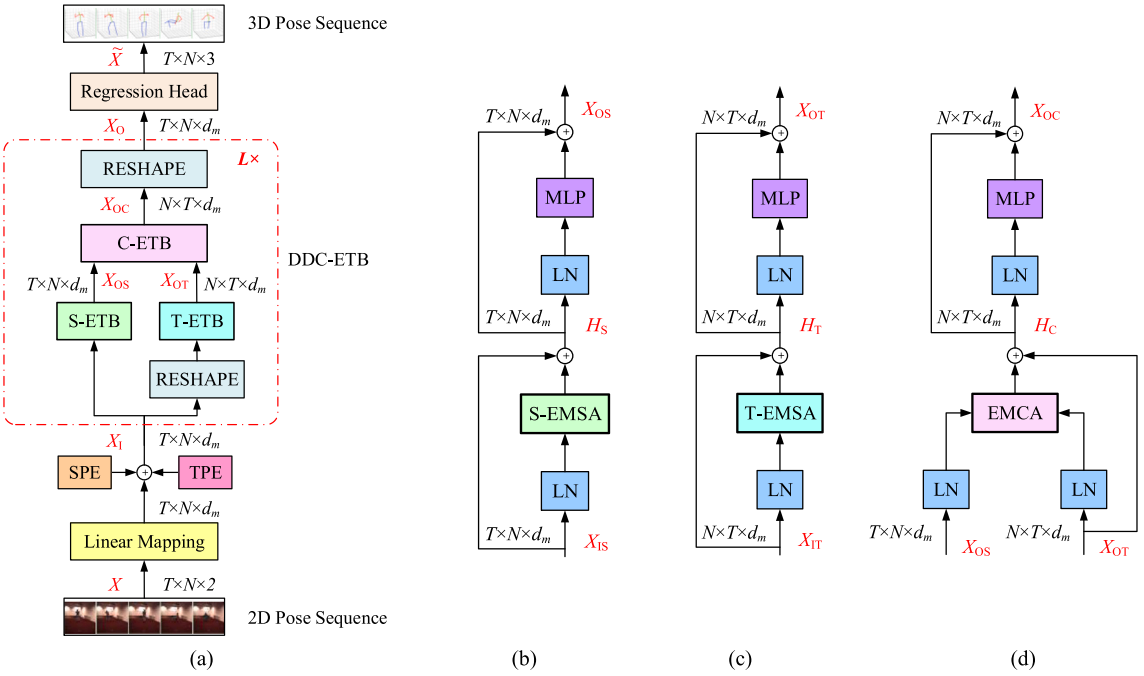


Fig. 1. Overall framework of DDCEFormer and flowcharts of DDC-ETB. (a) Overall framework of DDCEFormer. (b), (c), (d) Flowcharts of S-ETB, T-ETB, and C-ETB, respectively.

the refined features to predict the final 3D pose sequence $\tilde{X} \in R^{T\times N\times 3}$. In the model, T denotes sequence length, N denotes the number of body joints, and 2, d_m, and 3 denote the channel numbers of input, internal features, and output, respectively.

Figs. 1(b)–(d) further illustrate the internal structures of the three core sub-modules. Each follows an enhanced Transformer design with residual connections, containing an Enhanced Multi-head Attention (EMA) module and a Multi-Layer Perceptron (MLP), with features being Layer Normalized (LN) before input. The three sub-modules achieve functional differentiation through different attention mechanisms: S-ETB employs Spatial Enhanced Multi-head Self-Attention (S-EMSA), T-ETB adopts Temporal Enhanced Multi-head Self-Attention (T-EMSA), and C-ETB utilizes Enhanced Multi-head Cross-Attention (EMCA) to promote spatio-temporal feature interaction.

3.2. Feature mapping and position embedding

The modeling approach considers joints from both spatial and temporal dimensions: the spatial dimension captures structural relationships among joints within each frame, while the temporal dimension captures the motion trajectory of individual joints across consecutive frames. The DDCEFormer takes 2D keypoints as input, first maps them to high-dimensional features via a linear embedding layer, then introduces learnable SPE and TPE to fuse spatial and temporal position information, obtaining the initial features X_{I}, as shown in Eq. (1).

$$X_{\mathrm{I}} = \mathrm{FC}(X) + E_{\text{S-POS}} + \mathrm{RESHAPE}(E_{\text{T-POS}}) \tag{1}$$

where $X \in R^{T\times N\times 2}$ is the input continuous 2D pose sequence, $X_{\mathrm{I}} \in R^{T\times N\times d_m}$ is the output high-dimensional feature after feature mapping and position embedding. $E_{\text{S-POS}} \in R^{T\times N\times d_m}$ and $E_{\text{T-POS}} \in R^{N\times T\times d_m}$ are the spatial position embedded feature and temporal position embedded feature, respectively.

3.3. S-ETB and T-ETB modules

Within the DDC-ETB module, spatial motion relationships among joints and temporal motion relationships of joints are effectively modeled by the parallel-connected S-ETB and T-ETB. To fully extract global and local spatial and temporal features, the EMSA mechanism is proposed, based on conventional MSA and CNN. This mechanism is applied in S-ETB and T-ETB as S-EMSA and T-EMSA, respectively.

3.3.1. Enhanced multi-head self-attention (EMSA)

The architecture of conventional MSA is illustrated in Fig. 2(a). The MSA employs a multi-head joint mechanism to precisely model information from different representation subspaces. First, input vectors are linearly mapped to generate the query matrix Q, the key matrix K and the value matrix V adapted for multiple heads. For each attention head, scaled dot-product attention computes attention scores, which are multiplied by the corresponding V to obtain the single head's output. Finally, outputs from all heads are concatenated and linearly transformed to produce a vector with the same dimension as the original input. The core of MSA is the dynamic weighting of V via attention scores, which fuses multi-dimensional information for global feature representation.

Unlike the conventional MSA, we propose an EMSA, shown in Fig. 2(b). EMSA employs two parallel operations: one is variable-dimensional scaled dot-product multi-head self-attention, focusing on global information while reducing redundancy. The other is a convolution operation, treating V as a 2D feature map in space and time, applying a 2D convolution over adjacent joints to focus on local spatial and temporal motion relationships contained in the skeleton sequence. The fusion of variable-dimensional scaled dot-product multi-head self-attention with convolution allows the model to more effectively balance the integration of local and global information, improving prediction accuracy.

The attention operation in EMSA uses Q, K, and V matrices of heterogeneous dimensions. Specifically, the dimensions of Q and K are set to $M\times kd_m$, and V to $M\times gd_m$, where M is the sequence length, d_m is the feature dimension, and k, g are independent scaling factors.

The scaling factor k is designed to control the degree of dimensionality reduction applied to Q and K. Its specific functions include:

(1) Dimensionality Reduction: By reducing k, the dimensions of Q and K are effectively compressed, thereby decreasing the

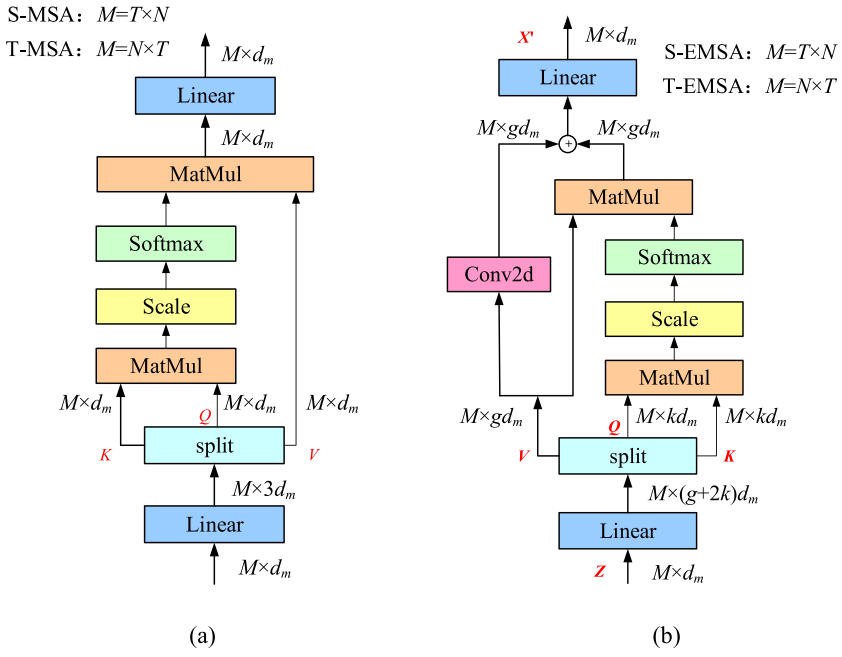


Fig. 2. Multi-head self-attention architectures. (a) Conventional Multi-head Self-Attention (MSA). (b) Enhanced Multi-head Self-Attention (EMSA), adopted in S-ETB and T-ETB as S-EMSA and T-EMSA, respectively. The fusion of variable-dimensional multi-head self-attention and convolution enables the model to focus on both global and local spatial and temporal information.

computational cost of the attention matrix QK^{T} and reducing the overall computational complexity of the model.

(2) Redundancy Elimination: The compressed Q and K mitigate redundant features while retaining the most critical information for global attention. This enables the model to focus on more informative dimensions, thereby improving the effectiveness of global information modeling.

The scaling factor g is designed to regulate the information capacity of V. Its primary roles include:

(1) Enhanced Feature Capacity: Increasing g results in a higher dimensionality for V compared to Q and K, allowing it to carry richer information and strengthen the expressive power of output features, which benefits subsequent tasks.

(2) Support for Convolutional Information Fusion: In the EMSA mechanism, a parallel branch performs convolution operations on V. A higher-dimensional V facilitates the capture of local dynamic features in both spatial and temporal domains, thereby improving the extraction of effective local information via convolution.

(3) Dimensional Decoupling: By separately controlling the dimensions of Q/K and V through the scaling factors k and g, the modeling of attention weights and the optimization of output representations are decoupled. This enhances the flexibility and representational capacity of the model.

In designing the local feature extraction path, only a convolution layer is introduced, omitting the activation, pooling, and fully connected layers commonly found in CNNs. This is because the MLP module within the Transformer already provides nonlinear mapping and fully connected operations; thus, retaining only the convolution layer can effectively fuse local information without reducing feature dimensionality.

In summary, the proposed EMSA achieves a good balance between global-local information integration and computational efficiency. Setting $k < 1$ compresses the Q and K dimensions, significantly improving computational efficiency while maintaining global modeling capability. Setting > 1 expands the V dimension combined with convolution, enhancing the expression of local dynamic features. Independently adjusting k and g allows the module to adapt to different task requirements, improving modeling capability for complex spatio-temporal sequences.

Taking S-EMSA in S-ETB as an example, the EMSA computation is explained. Input features Z_{S} for S-EMSA are linearly mapped to obtain the Q, K, V matrices of different dimensions: $Q_{\mathrm{S}} \in R^{N\times kd_m}$, $K_{\mathrm{S}} \in R^{N\times kd_m}$, $V_{\mathrm{S}} \in R^{N\times gd_m}$, as in Eq. (2).

$$Q_{\mathrm{S}} = Z_{\mathrm{S}}W_{\mathrm{QS}}, K_{\mathrm{S}} = Z_{\mathrm{S}}W_{\mathrm{KS}}, V_{\mathrm{S}} = Z_{\mathrm{S}}W_{\mathrm{VS}} \tag{2}$$

where $W_{\mathrm{QS}} \in R^{d_m\times kd_m}$, $W_{\mathrm{KS}} \in R^{d_m\times kd_m}$, and $W_{\mathrm{VS}} \in R^{d_m\times gd_m}$ are linear transformation matrices, and $Z_{\mathrm{S}} = \mathrm{LN}(X_{\mathrm{IS}})$ is the layer-normalized input feature of S-ETB.

S-EMSA uses variable-dimensional multi-head self-attention to extract global features and convolution to extract local features. The two outputs are linearly combined and transformed to produce the S-EMSA output, achieving a fusion of spatial global and local features. The process is shown in Eq. (3).

$$X'_{\mathrm{S}} = [\mathrm{Softmax}(Q_{\mathrm{S}}K_{\mathrm{S}}^{\mathrm{T}}/\sqrt{kd_m})V_{\mathrm{S}} + \mathrm{Conv2d}(V_{\mathrm{S}})]W_{\mathrm{OS}} \tag{3}$$

where k is the scaling factor, $Q_{\mathrm{S}} \in R^{T\times N\times kd_m}$, $K_{\mathrm{S}} \in R^{T\times N\times kd_m}$, and $V_{\mathrm{S}} \in R^{T\times N\times gd_m}$ are the query, key, and value matrices, and $W_{\mathrm{OS}} \in R^{gd_m\times d_m}$ is a linear transformation matrix.

For simplicity, the S-EMSA computation is denoted as Eq. (4).

$$X'_{\mathrm{S}} = \text{S-EMSA}(Z_{\mathrm{S}}) \tag{4}$$

where $Z_{\mathrm{S}} \in R^{T\times N\times d_m}$ and $X'_{\mathrm{S}} \in R^{T\times N\times d_m}$ are the input and output features of S-EMSA, and S-EMSA$(\cdot)$ denotes the operation function.

The analysis for T-EMSA is similar to that of S-EMSA and is omitted for brevity.The difference is that variables in Fig. 2(b) are marked

with the subscript 'T', $M = N \times T$. The T-EMSA computation is denoted as Eq. (5).

$$X'_{\mathrm{T}} = \text{T-EMSA}(Z_{\mathrm{T}}) \tag{5}$$

where $Z_{\mathrm{T}} \in R^{N\times T\times d_m}$ and $X'_{\mathrm{T}} \in R^{N\times T\times d_m}$ are the input and output features of T-EMSA, $Z_{\mathrm{T}} = \mathrm{LN}(X_{\mathrm{IT}})$ is the layer-normalized input feature of T-ETB, and T-EMSA(·) denotes the operation function.

3.3.2. Multi-layer perceptron (MLP)

The MLP consists of two linear layers for non-linear transformation and feature mapping, as shown in Eq. (6).

$$H_{\mathrm{MLP}} = \sigma(HW_1 + b_1)W_2 + b_2 \tag{6}$$

where $H \in R^{M\times d_m}$ is the MLP input feature, $H_{\mathrm{MLP}} \in R^{M\times d_m}$ is the output. σ denotes the GELU activation function for nonlinearity. $W_1 \in R^{d_m\times d}$ and $W_2 \in R^{d\times d_m}$ are the weights of the two linear fully connected layers, and $b_1 \in R^d$, $b_2 \in R^{d_m}$ are the bias terms.

3.3.3. Computation of S-ETB and T-ETB

As shown in Fig. 1(b), S-ETB consists of a multi-head self-attention module and an MLP module. The LN is applied before each module. The S-ETB computation is shown in Eq. (7).

$$\begin{aligned} H_{\mathrm{S}} &= \text{S-EMSA}(\mathrm{LN}(X_{\mathrm{IS}})) + X_{\mathrm{IS}}, \\ X_{\mathrm{OS}} &= \mathrm{MLP}(\mathrm{LN}(H_{\mathrm{S}})) + H_{\mathrm{S}}. \end{aligned} \tag{7}$$

where $X_{\mathrm{IS}} \in R^{T\times N\times d_m}$, $H_{\mathrm{S}} \in R^{T\times N\times d_m}$, and $X_{\mathrm{OS}} \in R^{T\times N\times d_m}$ are the input, intermediate, and output features of S-ETB, respectively.

T-ETB, shown in Fig. 1(c), is computed as in Eq. (8).

$$\begin{aligned} H_{\mathrm{T}} &= \text{T-EMSA}(\mathrm{LN}(X_{\mathrm{IT}})) + X_{\mathrm{IT}}, \\ X_{\mathrm{OT}} &= \mathrm{MLP}(\mathrm{LN}(H_{\mathrm{T}})) + H_{\mathrm{T}}. \end{aligned} \tag{8}$$

where $X_{\mathrm{IT}} \in R^{N\times T\times d_m}$, $H_{\mathrm{T}} \in R^{N\times T\times d_m}$, and $X_{\mathrm{OT}} \in R^{N\times T\times d_m}$ are the input, intermediate, and output features of T-ETB, respectively.

3.4. Enhanced multi-head cross-attention (EMCA) and C-ETB

The S-ETB and T-ETB modules are designed to achieve comprehensive representations of spatial and temporal features, respectively. Existing studies typically fuse these two types of features through strategies such as feature concatenation [18], simple addition [23], or adaptive weighting [50]. However, these approaches treat spatial and temporal features equally without interactive refinement, failing to capture the deep dependencies between the two domains and thus limiting the context-aware capability of the fused representations.

To address this limitation, we propose the C-ETB module based on an enhanced attention mechanism. The core of this module lies in the introduction of the EMCA mechanism to achieve dynamic and deep fusion of spatial and temporal features. The fundamental concept can be summarized as cross-domain modulation: temporal features X_{OT} serve as dynamic filters to reweight and selectively enhance spatial features X_{OS}. Through this process, the temporal context adaptively modulates the spatial features, enabling the fused results to be more coherent and context-aware under the guidance of temporal information.

The specific architecture of EMCA is illustrated in the dashed box of Fig. 3, with a design that balances global interaction and local detail preservation. Specifically, EMCA employs a variable-dimension multi-head attention mechanism to achieve global cross-attention: the temporal features X_{OT} are projected into the query matrix Q and the key matrix K, while the spatial features X_{OS} serve as the value matrix V in the attention computation. This design enables temporal information to guide the weighting of spatial features from a global perspective. Simultaneously, to preserve local structural information within the spatial features, EMCA introduces a parallel convolutional path on the

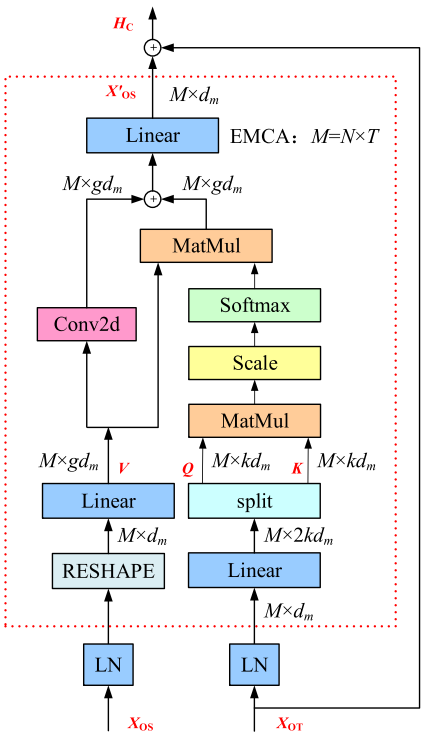


Fig. 3. Spatio-temporal information fusion. The dashed box shows the EMCA. The input of EMCA comes from X_{OT} and X_{OS}, and its output is X'_{OS}, which realizes the weighting of spatial features. X'_{OS} and X_{OT} are summed, achieving weighted fusion of X_{OT} and X_{OS}.

V, capturing local details through convolution operations. Finally, the outputs from the attention mechanism and the convolutional path are combined through linear addition and transformation to produce the final EMCA output, thereby achieving dynamic weighted optimization of the spatial features. This mechanism effectively enhances the representation capability of cross-domain fusion by integrating global modulation with local feature preservation.

The Q and K matrices of EMCA have the same dimension $M\times kd_m$, and the V matrix has the dimension $M\times gd_m$, with k and g as scaling factors, $M = N\times T$. As shown in Eq. (9).

$$\begin{aligned} Q_{\mathrm{C}} &= \mathrm{LN}(X_{\mathrm{OT}})W_{\mathrm{QC}}, \\ K_{\mathrm{C}} &= \mathrm{LN}(X_{\mathrm{OT}})W_{\mathrm{KC}}, \\ V_{\mathrm{C}} &= \mathrm{RESHAPE}(\mathrm{LN}(X_{\mathrm{OS}}))W_{\mathrm{VC}}. \end{aligned} \tag{9}$$

where $W_{\mathrm{QC}} \in R^{d_m\times kd_m}$, $W_{\mathrm{KC}} \in R^{d_m\times kd_m}$, and $W_{\mathrm{VC}} \in R^{d_m\times gd_m}$ are linear transformation matrices; $\mathrm{LN}(X_{\mathrm{OT}})$ and $\mathrm{RESHAPE}(\mathrm{LN}(X_{\mathrm{OS}}))$ are derived from the output features of T-ETB and S-ETB respectively.

EMCA computation is shown in Eq. (10).

$$X'_{\mathrm{OS}} = [\mathrm{Softmax}(Q_{\mathrm{C}}K_{\mathrm{C}}^{\mathrm{T}}/\sqrt{kd_m})V_{\mathrm{C}} + \mathrm{Conv2d}(V_{\mathrm{C}})]W_{\mathrm{OC}} \tag{10}$$

where k is the scaling factor; $Q_{\mathrm{C}} \in R^{N\times T\times kd_m}$, $K_{\mathrm{C}} \in R^{N\times T\times kd_m}$, $V_{\mathrm{C}} \in R^{N\times T\times gd_m}$, and $W_{\mathrm{OC}} \in R^{gd_m\times d_m}$ denote the query, key, value, and output transformation matrices in EMCA, respectively.

From Eqs. (9) and (10), X'_{OS} is the weighted feature of X_{OS}, denoted as $X'_{\mathrm{OS}} = \alpha X_{\mathrm{OS}}$, where α is the weighting coefficient.

The output X'_{OS} is summed with temporal correlation features X_{OT}, achieving weighted fusion of spatio-temporal features X_{OT} and X_{OS}. Therefore, the fused feature H_{C} of C-ETB is given by Eq. (11).

$$H_{\mathrm{C}} = X'_{\mathrm{OS}} + X_{\mathrm{OT}} = \alpha X_{\mathrm{OS}} + X_{\mathrm{OT}} \tag{11}$$

In Fig. 1(d), the fused feature H_{C} undergoes MLP feature extraction to effectively capture joint spatio-temporal motion relationships,

yielding the C-ETB output features, as in Eq. (12).

$$X_{\mathrm{OC}} = \mathrm{MLP}(\mathrm{LN}(H_{\mathrm{C}})) + H_{\mathrm{C}}. \tag{12}$$

where $H_{\mathrm{C}} \in R^{N\times T\times d_m}$ and $X_{\mathrm{OC}} \in R^{N\times T\times d_m}$ are the intermediate and output features of C-ETB.

3.5. Regression head

As illustrated in Fig. 1(a), the output X_{O} of DDC-ETB passes through a linear regression head to obtain the 3D estimated pose sequence $\tilde{X}$. The regression head computation is shown in Eq. (13).

$$\tilde{X} = \mathrm{FC}(\mathrm{LN}(X_{\mathrm{O}})) \tag{13}$$

where $X_{\mathrm{O}} \in R^{T\times N\times d_m}$ and $\tilde{X} \in R^{T\times N\times 3}$ are the inputs and outputs of the linear regression head.

3.6. Loss function

The entire model is trained using an end-to-end approach. The loss function is designed to consider both positional accuracy and velocity accuracy, aiming to enhance the model's predictive capability in dynamic environments, generate smoother and more natural motion trajectories, improve its responsiveness to rapid changes, and strengthen overall robustness.

The defined loss function consists of three parts: L_p, L_s, and L_v, where L_p is the Mean Per Joint Position Error (MPJPE), L_s is the Normalized MPJPE (N-MPJPE) after normalizing the predicted position by scale, and L_v is the Mean Per Joint Velocity Error (MPJVE). Specifically, it is shown in Eq. (14).

$$L = L_p + \lambda_s L_s + \lambda_v L_v \tag{14}$$

where λ_s and λ_v are the weighting coefficients. L_p, L_s, and L_v are given in Eqs. (15–17).

$$L_p = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{n=1}^{N}\|Y_n^t - \tilde{X}_n^t\|_2\right) \tag{15}$$

$$L_s = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{n=1}^{N}\|Y_n^t - s(t)\tilde{X}_n^t\|_2\right) \tag{16}$$

where $s(t) = \left(\frac{1}{N}\sum_{n=1}^{N}(Y_n^t \cdot \tilde{X}_n^t)\right)/\left(\frac{1}{N}\sum_{n=1}^{N}\|\tilde{X}_n^t\|_2^2\right)$, $s(t)$ denotes the coefficients of the predicted position at frame t normalized by scale, and the operator $(\cdot)$ denotes the inner product of the two vectors.

$$L_v = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(\frac{1}{N}\sum_{n=1}^{N}\|(Y_n^{t+1} - Y_n^t) - (\tilde{X}_n^{t+1} - \tilde{X}_n^t)\|_2\right) \tag{17}$$

where $Y \in R^{T\times N\times 3}$, $\tilde{X} \in R^{T\times N\times 3}$, Y_n^t and $\tilde{X}_n^t$ are the 3D pose ground truth and predicted values of the nth joint at frame t, respectively. Y_n^{t+1} and $\tilde{X}_n^{t+1}$ are the 3D pose ground truth and predicted values of the nth joint at frame $t+1$, respectively.

4. Experimental results and analysis

4.1. Datasets and evaluation metrics

The effectiveness of DDCEFormer is validated through experiments on two large-scale datasets, Human3.6M and MPI-INF-3DHP, and comparative analysis against other methods.

Human3.6M contains 3.6 million video frames, capturing 11 professional subjects performing 15 daily activities. Data is recorded by 4 synchronized cameras at 50Hz, making it the most common indoor dataset for 3D HPE. Following the same protocol as compared methods [18,23–25], we employ the widely-adopted 2D pose estimator CPN to extract 17 2D joints. Data from 5 subjects (S1, S5, S6, S7, S8) are used for training and 2 subjects (S9, S11) for testing.

Two common evaluation protocols are used [18,23–25]. Protocol 1 (P1) is Mean Per Joint Position Error (MPJPE) after aligning the root joint (pelvis), computing the average Euclidean distance between estimated and ground truth 3D joints. Protocol 2 (P2) is Reconstruction MPJPE (P-MPJPE), the average position error after performing Procrustes alignment (rigid transformation) between estimates and ground truth. Both metrics are in mm; lower values indicate smaller errors and higher accuracy.

MPI-INF-3DHP contains over 1.3 million frames captured from 14 camera views, involving 8 subjects performing 8 activities in both indoor and outdoor environments. Although smaller than Human3.6M, its diverse scenes, viewpoints, and actions make it more challenging. The dataset includes complex outdoor scenarios and frequent occlusions, partially simulating in-the-wild conditions and thereby providing a robust test of model generalization. Following prior work [18,23–25], ground truth 2D poses are used as input. Performance is evaluated using MPJPE, Percentage of Correct Keypoints within 150mm (PCK), and Area Under the Curve (AUC), where higher PCK and AUC indicate better generalization and robustness, and lower MPJPE reflects higher prediction accuracy.

4.2. Experimental environment and parameter settings

Experiments are conducted on a hardware platform with an Intel Core i9-9900k CPU@3.60GHz, NVIDIA GeForce RTX 3090 GPU, 64GB DDR4 2666 MHz RAM, implemented based on PyTorch.

Settings for Human3.6M: Optimized with Adam, trained for 180 epochs, initial learning rate 4e-5, decay factor 0.99 per epoch. Parameters: input sequence length $T=243$, joint feature dimension $d_m=512$, number of layers $L=8$, number of attention heads $h=8$, loss weights $\lambda_s=0.5$, $\lambda_v=20$, convolutional kernel size 3×3, and scaling factors $k=0.75$, $g=1.5$ for EMSA and EMCA.

Settings for MPI-INF-3DHP: Sequence lengths are shorter, so $T=27$ and $T=81$ are selected; other parameters are consistent with Human3.6M.

4.3. Analysis of experimental results

This section validates the method's effectiveness through comparisons and detailed analysis.

4.3.1. Comparative analysis on human3.6M

On Human3.6M, DDCEFormer is compared with state-of-the-art methods using P1 and P2 metrics. Detailed results are in Tables 1 and 2. These tables cover the evaluation of 15 daily activity actions, with the best values highlighted in bold and the second-best values underlined. The last row of the table shows the difference between our method's estimates and the best-performing estimates from other methods. The unit of the data is millimeters (mm).

The comparative results in Table 1 (Protocol 1, MPJPE) indicate that our method achieves an average MPJPE of 39.1mm across all actions, obtaining the best MPJPE for 11 actions and the second-best for 3. Similarly, as detailed in Table 2 (Protocol 2, P-MPJPE), our method attains an average P-MPJPE of 30.8mm, with the best for 11 actions and the second-best for 4. The last rows show that our method performs particularly well in actions with large limb movements like Eat, Photo, Sit, and Smoke. MPJPE reductions are 1.4mm (4.0%), 3.3mm (7.3%), 2.1mm (4.3%), and 1.2mm (3.0%) compared to the second-best, and P-MPJPE reductions are 1.1mm (3.8%), 2.5mm (7.1%), 2.5mm (6.5%), and 1.0mm (3.1%).

These results indicate that by fully fusing global and local spatio-temporal features, our method achieves higher accuracy and stronger robustness for poses with large limb movements. To evaluate the statistical stability of our results, we conducted each experiment five times with different random seeds. Our method attained a mean MPJPE of 39.1mm with a standard deviation of 0.2mm.

Table 1
Comparative analysis of MPJPE for P1 on the Human3.6M dataset in millimeters.

Method	T	Publication	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
GraphMLP[22]	243	PR'25	40.0	44.2	39.9	43.4	46.5	52.2	42.3	40.9	55.8	59.5	45.1	42.1	45.2	29.5	30.3	43.8
StridedFormer[16]	243	TMM'22	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
ConvFormer[48]	243	VC'24	41.0	43.2	39.0	42.4	44.5	52.2	41.7	40.8	53.0	60.6	44.8	41.3	43.7	29.6	30.9	43.2
PATA[21]	243	TIP'22	39.9	42.7	40.3	42.3	45.0	52.8	40.4	39.3	56.9	61.2	44.1	41.3	42.8	28.4	29.3	43.1
MHFormer[46]	351	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
MHFormer++[17]	351	PR'23	39.1	42.7	38.7	40.3	44.1	50.0	41.4	38.7	53.9	61.6	43.6	40.8	42.5	29.6	30.6	42.5
P-STMO[45]	243	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
GLSTE[20]	243	TMM'24	39.2	42.3	39.6	41.0	44.0	49.6	41.0	39.9	53.2	59.2	43.2	41.2	42.1	29.0	29.2	42.2
MixSTE[18]	243	CVPR'22	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
STCFormer[23]	243	CVPR'23	38.4	41.2	36.8	38.0	42.7	50.5	38.7	**38.2**	52.5	56.8	41.8	38.4	40.2	**26.2**	**27.7**	40.5
STGFormer[24]	243	PR'26	38.4	40.7	37.1	38.1	42.7	49.9	39.0	39.3	50.4	56.0	41.4	38.3	**39.1**	26.9	28.1	40.3
FMFormer[25]	243	TMM'24	**36.9**	39.6	36.9	39.3	41.8	48.3	38.4	38.7	51.1	53.7	41.9	38.7	40.4	27.7	27.9	40.1
Ours($k=0.75$, $g=1.5$)	243		37.6 +0.7	**39.4** −0.2	**35.4** −1.4	**37.7** −0.3	**41.2** −0.6	**45.0** −3.3	**37.9** −0.5	38.5 +0.3	**48.3** −2.1	**53.3** −0.4	**40.2** −1.2	**37.6** −0.7	39.9 +0.8	27.1 +0.9	**27.7** +0.0	**39.1** −1.0

Table 2
Comparative analysis of P-MPJPE for P2 on the Human3.6M dataset in millimeters.

Method	T	Publication	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
GraphMLP[22]	243	PR'25	32.1	34.9	32.4	35.7	36.2	41.5	33.2	31.4	44.4	47.7	36.9	32.8	35.5	23.9	25.0	34.9
StridedFormer[16]	243	TMM'22	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
ConvFormer[48]	243	VC'24	31.4	34.2	32.0	35.2	34.0	40.3	32.7	31.3	42.6	49.0	36.2	31.3	34.8	23.4	24.9	34.2
PATA[21]	243	TIP'22	31.2	34.1	31.9	33.8	33.9	39.5	31.6	30.0	45.4	48.1	35.0	31.1	33.5	22.4	23.6	33.7
MHFormer[46]	351	CVPR'22	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
MHFormer++ [17]	351	PR'23	31.6	34.8	32.2	33.2	34.7	39.7	33.0	31.0	43.5	49.6	36.1	32.4	33.8	23.9	24.7	34.2
P-STMO[45]	243	ECCV'22	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
GLSTE[20]	243	TMM'24	30.8	34.6	32.4	32.9	34.0	39.5	31.5	31.0	44.2	48.7	35.2	32.4	33.5	23.1	23.5	33.8
MixSTE[18]	243	CVPR'22	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
STCFormer[23]	243	CVPR'23	**29.3**	33.0	30.7	30.6	32.7	38.2	29.7	**28.8**	42.2	45.0	33.3	29.4	31.5	**20.9**	22.3	31.8
STGFormer[24]	243	PR'26	**29.3**	33.0	29.9	30.6	32.5	37.5	30.3	29.6	41.0	44.1	33.2	29.1	**30.9**	21.7	22.8	31.7
FMFormer[25]	243	TMM'24	30.1	32.4	30.2	31.9	32.4	37.8	30.4	29.9	41.2	**43.4**	34.0	29.8	32.3	21.5	22.4	32.0
Ours($k=0.75$, $g=1.5$)	243		**29.3** +0.0	**31.7** −0.7	**28.8** −1.1	**29.7** −0.9	**31.5** −0.9	**35.0** −2.5	**28.8** −0.9	29.2 +0.4	**38.5** −2.5	43.5 +0.1	**32.2** −1.0	**28.7** −0.4	31.5 +0.6	21.1 +0.2	**22.2** −0.1	**30.8** −0.9

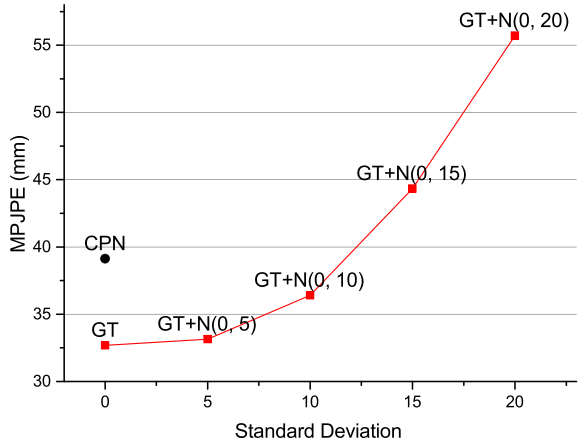


Fig. 4. Adding Gaussian noise to evaluate the robustness of DDCEFormer model.

Furthermore, we evaluated the robustness of our method to different noise levels by adding Gaussian noise with varying standard deviations to the 2D ground-truth joint coordinates during testing. The experimental results are presented in Fig. 4. It can be observed that the MPJPE of 3D pose estimation is positively correlated with the noise level of the 2D input, with a relatively gentle increase under low-noise conditions. This indicates that the proposed method exhibits a certain capacity to suppress input noise. These results validate the effectiveness and robustness of our method under noisy conditions.

We also systematically evaluate DDCEFormer against two representative seq2seq Transformer baselines across four aspects: computational complexity, inference efficiency, training resource consumption, and estimation accuracy. The results are presented in Table 3.

Table 3
Comparison of computational complexity and efficiency with state-of-the-art methods.

Method	Param (M)	FLOPs (G)	FPS	GPU Memory (G)	Training Time (min/epoch)	MPJPE (mm)
MixSTE[18]	33.8	139	9856	11.7	22.5	40.9
FMFormer[25]	42.3	193	7297	15.3	28.3	40.1
Ours	53.8	222	4858	19.5	36.2	39.1

As shown in Table 3, DDCEFormer achieves the best MPJPE. This performance gain is primarily attributed to its enhanced multi-head attention mechanism and spatio-temporal cross-augmentation fusion strategy. However, this improvement in accuracy comes with increased computational cost. Compared to the baselines, our model exhibits higher parameter counts and FLOPs, leading to greater GPU memory usage and longer training time per epoch, as well as a reduced inference speed. These results indicate a trade-off between performance and efficiency in the current model, suggesting that its deployment in real-time or resource-constrained scenarios requires further optimization.

4.3.2. *Comparative analysis on MPI-INF-3DHP*

Table 4 compares our method with others on MPI-INF-3DHP. For $T=81$, our method's PCK, AUC, and MPJPE improve by 0.6%, 4.9%, and reduce by 9.7mm compared to STCFormer, and by 0.5%, 4.7%, and reduce by 5.8mm compared to the second-best STGFormer. For $T=27$, compared to MixSTE, improvements are 4.8% in PCK, 21.7% in AUC, and a reduction of 40.4mm in MPJPE. By enhancing multi-head self-attention and cross-attention mechanisms, our method effectively fuses global and local spatio-temporal features. The results demonstrate the strong generalization ability of DDCEFormer.

Table 4
Results on MPI-INF-3DHP under three evaluation metrics.

Method	T	Publication	PCK↑ (%)	AUC↑ (%)	MPJPE↓ (mm)
PoseFormer[15]	9	ICCV'21	88.6	56.4	77.1
MHFormer[46]	9	CVPR'22	93.8	63.3	58.0
MixSTE[18]	27	CVPR'22	94.4	66.5	54.9
P-STMO[45]	81	ECCV'22	97.9	75.8	32.2
STCFormer[23]	81	CVPR'23	98.7	83.9	23.1
STGFormer[24]	81	PR'26	98.8	84.1	19.2
FMFormer[25]	81	TMM'24	98.6	71.5	39.4
Ours($k=0.75$, $g=1.5$)	27		99.2	88.2	14.5
	81		**99.3**	**88.8**	**13.4**

4.3.3. Detailed analysis

The core challenge of 3D human pose estimation lies in the accurate reconstruction of distal joints. Upper-body joints such as the wrists and elbows possess high degrees of freedom and are prone to self-occlusion and depth ambiguity during complex motions, which generally lead to significant estimation errors in existing methods. In contrast, although lower-body joints also exhibit high degrees of freedom, their movement patterns are relatively regular and they are less affected by occlusions. Therefore, conducting a fine-grained analysis of estimation performance across different body parts, particularly investigating the accuracy discrepancy between upper and lower limbs, is essential for gaining a deeper understanding of model behavior.

Accordingly, we perform a detailed analysis of MPJPE for all 17 joints on the Human3.6M dataset to comprehensively evaluate the per-joint estimation performance of our method, comparing it against FMFormer, MixSTE, and MHFormer. The results are presented in Fig. 5.

As observed in Fig. 5, the estimation errors for the ankles (joints 4 and 7) and wrists (joints 14 and 17) are consistently higher than those for other joints across all compared methods. This phenomenon aligns with the inherently high degrees of freedom of distal joints, further confirming the universality of the aforementioned challenge. Our method achieves the best overall average accuracy, with a mean MPJPE of 39.13mm, outperforming FMFormer (40.09mm), MixSTE (40.94mm), and MHFormer (42.95mm). Further analysis reveals that the advantages of our method are predominantly observed in the upper-body joints. Notably, it achieves the best estimates on the four most challenging upper-body distal joints: the right elbow (13), right wrist (14), left elbow (16), and left wrist (17). For lower-body joints, our method generally performs better than MixSTE and MHFormer, while achieving results comparable to FMFormer, with each method demonstrating strengths on different joints.

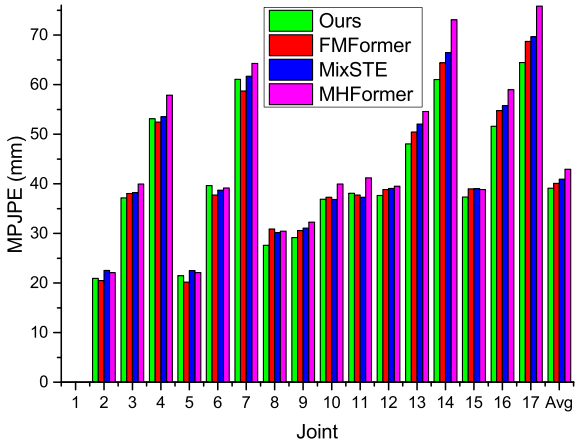


Fig. 5. Comparison of MPJPE for 17 human body joints.

Building on this observation, we conducted a detailed analysis of the four key upper limb joints to further quantify the improvement achieved by our method on upper-body joints, with the results reported in Table 5.

Table 5 demonstrates that our method achieves the best estimation accuracy across all four upper-body joints. Compared to the second-best method, FMFormer, our approach reduces the error by 2.4mm (right elbow), 3.4mm (right wrist), 3.1mm (left elbow), and 4.2mm (left wrist), corresponding to relative improvements ranging from 5.0% to 6.5%. The improvement is even more pronounced when compared to MHFormer, achieving a maximum reduction of 19.8%. These results robustly validate the effectiveness of our method in modeling complex upper-body motions. Notably, our method exhibits superior performance in actions involving substantial limb movement, such as Direction (joint 17), Eating (joints 13 and 14), Greeting (joints 13, 14 and 17), Photo (joints 13, 14, 16 and 17), Posing (joint 17), Purchases (joint 17), Smoking (joint 13), Sitting (joints 14, 16 and 17), and Waiting (joint 14). This further demonstrates its effectiveness in enhancing both the accuracy and robustness of 3D human pose estimation.

The above fine-grained analysis reveals the core strengths of our method. For the complex and varied motions of the upper-body joints,

Table 5
Detailed MPJPE analysis for four upper-body joints on the Human3.6M dataset in millimeters.

Joint	Method	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
right elbow (13)	MHFormer[46]	47.8	52.3	52.5	53.8	57.9	79.9	49.8	53.8	62.4	67.1	55.3	53.9	54.0	35.9	42.0	54.6
	MixSTE[18]	47.0	52.9	51.1	49.4	52.5	75.9	49.5	55.3	66.6	61.3	53.4	52.2	49.7	30.1	34.1	52.1
	FMFormer[25]	**45.7**	50.2	48.5	49.1	53.8	70.6	**46.4**	**50.2**	69.4	58.7	51.6	49.1	**47.6**	30.3	35.2	50.4
	Ours	47.3	**49.0**	**43.3**	**45.1**	**49.5**	**64.9**	48.4	51.2	**61.0**	**55.3**	**46.8**	**48.6**	48.9	**27.3**	**33.9**	**48.0**
		+1.6	−1.2	−5.2	−4.0	−3.0	−5.7	+2.0	+1.0	−1.4	−3.4	−4.8	−0.5	+1.3	−2.8	−0.2	−2.4
right wrist (14)	MHFormer[46]	67.7	67.4	77.3	77.6	86.3	97.4	66.4	72.0	85.9	84.3	78.8	66.5	69.9	41.8	56.7	73.1
	MixSTE[18]	63.0	64.8	67.5	67.5	73.2	91.1	63.2	72.4	83.3	78.7	68.6	62.5	63.4	35.0	43.0	66.5
	FMFormer[25]	59.5	62.4	66.2	67.6	72.1	85.8	59.8	67.7	81.6	74.9	70.2	59.4	**61.4**	34.4	42.7	64.4
	Ours	**58.3**	**60.4**	**60.5**	**60.6**	**71.6**	**78.4**	**59.1**	**63.8**	**75.9**	**72.0**	**65.9**	**54.9**	61.8	**32.5**	**39.6**	**61.0**
		−1.2	−2.0	−5.7	−6.9	−0.5	−7.4	−0.7	−3.9	−5.7	−2.9	−2.7	−4.5	+0.4	−1.9	−3.1	−3.4
left elbow (16)	MHFormer[46]	55.4	52.0	59.2	64.7	59.3	83.9	61.5	56.2	75.5	69.5	55.9	55.0	59.3	39.8	37.4	59.0
	MixSTE[18]	50.9	51.0	55.6	57.3	56.9	87.8	58.9	54.5	70.1	64.6	52.3	52.4	57.2	35.2	32.4	55.8
	FMFormer[25]	51.3	49.2	53.7	58.3	58.7	81.5	57.5	50.8	70.2	60.4	55.2	50.4	56.4	34.6	33.2	54.7
	Ours	**50.5**	**48.1**	**51.5**	**54.3**	**55.9**	**73.6**	**54.0**	**49.3**	**61.8**	**59.5**	**50.7**	**50.0**	**53.1**	**32.1**	**30.0**	**51.6**
		−0.4	−1.1	−2.2	−3.0	−1.0	−7.9	−3.5	−1.5	−8.3	−0.9	−1.6	−0.4	−3.3	−2.5	−2.4	−3.1
left wrist (17)	MHFormer[46]	84.6	68.6	76.9	91.2	71.9	101.4	71.6	76.5	90.6	87.9	75.8	65.7	84.0	42.7	48.1	75.8
	MixSTE[18]	78.3	64.3	68.4	81.6	69.8	103.2	67.8	75.7	80.8	75.6	67.9	63.0	74.8	35.4	38.2	69.7
	FMFormer[25]	76.2	62.1	68.0	80.1	66.8	98.6	66.3	74.2	86.1	**71.9**	70.5	58.5	76.9	35.7	38.4	68.7
	Ours	**71.2**	**61.6**	**64.8**	**73.2**	**66.5**	**87.1**	**60.3**	**68.1**	**72.0**	72.0	**66.4**	**58.4**	**73.5**	**33.8**	**38.1**	**64.5**
		−5.0	−0.5	−3.2	−6.9	−0.3	−11.5	−6.0	−6.1	−8.8	+0.1	−1.5	−0.1	−1.3	−1.6	−0.1	−4.2

our method effectively captures the movement characteristics of high degree of freedom joints through its efficient local feature extraction and spatio-temporal cross-fusion mechanism, significantly mitigating the large errors typically associated with distal upper limbs. For lower-body joints, which exhibit more regular movement patterns, our method maintains performance comparable to leading approaches. These results indicate that our method possesses a distinct advantage in handling joints with high degrees of freedom and large ranges of motion.

4.4. Parameter sensitivity analysis

The proposed EMSA and EMCA are variable-dimensional multi-head attention mechanisms, with the scaling factors k and g controlling dimension expansion for the Q, K and V matrices. Table 6 examines the impact of varying k with fixed $g=1.5$. Table 7 analyzes the impact of varying g with fixed $k=0.75$.

As shown in Table 6, increasing k leads to a higher number of parameters and greater storage requirements; however, the corresponding MPJPE and P-MPJPE do not exhibit a monotonically decreasing trend. When k is small (e.g., 0.25, 0.5), the feature extraction capability of the Q and K matrices in the attention mechanism is insufficient, preventing optimal performance. When k is large (e.g., 1.00, 1.25), the feature redundancy in the Q and K matrices also prevents optimal performance. At $k=0.75$, MPJPE and P-MPJPE reach their best values, indicating optimal feature selection. For g, Table 7 shows a similar trend, with optimal performance at $g=1.5$.

For comparison, using conventional fixed-dimension multi-head attention ($k=1.0$, $g=1.0$) yields MPJPE and P-MPJPE of 39.65mm and 31.15mm. From the sensitivity analysis, at $k=0.75$, $g=1.5$, MPJPE and P-MPJPE reduce to 39.13mm (reduction of 0.52mm, 1.33%) and 30.77mm (reduction of 0.38mm, 1.23%). This indicates that the variable-dimension method achieves a better balance between model complexity and accuracy.

Although the heterogeneous dimension scaling in EMA balances capacity and efficiency, it suffers from inherent limitations. The scaling factors rely on empirical tuning; excessive compression may create an information bottleneck, while expanding the V dimension directly increases the parameter count. Furthermore, the scaling factors k and g currently require manual selection, rendering the model sensitive to these two hyperparameters.

4.5. Ablation studies

All ablation studies are conducted on the Human3.6M dataset, using the typical 2D pose estimator CPN to extract 17 joint points of the 2D human body as inputs to the model.

Table 6
Experiments on the scaling factor k with $g=1.5$.

Scaling Factor k	Params (M)	Storage (M)	MPJPE (mm)	P-MPJPE (mm)
1.25	60.11	229	<u>39.34</u>	31.09
1.00	56.96	217	39.58	<u>31.00</u>
0.75	53.81	205	**39.13**	**30.77**
0.50	50.66	193	39.37	31.26
0.25	47.50	181	39.49	31.06

Table 7
Experiments on the scaling factor g with $k=0.75$.

Scaling Factor g	Params (M)	Storage (M)	MPJPE (mm)	P-MPJPE (mm)
2.00	60.17	230	40.10	31.59
1.75	56.99	218	39.99	31.33
1.50	53.81	205	**39.13**	**30.77**
1.25	50.63	193	39.69	<u>31.13</u>
1.00	47.45	181	39.53	31.53
0.75	44.27	169	<u>39.34</u>	31.38

Notations. In the ablation study results tables presented in the following sections, the best values are highlighted in bold, and the second-best values are underlined.

4.5.1. C-ETB module and CNN convolution in EMA

We incorporate convolutional operations into the EMA module to enhance local feature extraction. To determine the optimal kernel size, we conduct an ablation study comparing three different kernel sizes: 3×3, 5×5, and 7×7. The results demonstrate that the 3×3 kernel achieves the lowest MPJPE (39.13mm), outperforming the 5×5 (39.88mm) and 7×7 (39.71mm) kernels. This indicates that a 3×3 kernel is sufficient for effective local feature extraction, whereas larger kernels lead to degraded performance.

Building on this, we further evaluate the individual contributions of two key components: the C-ETB and the convolution in the EMA. The C-ETB is specifically designed for spatio-temporal fusion, whereas the convolution in the EMA aims to strengthen local information. Under the fixed settings of a 3×3 kernel, $k=0.75$, and $g=1.5$, we perform an ablation study by toggling the presence of the C-ETB and the EMA convolution. The results, also summarized in Table 8, confirm the effectiveness of both components.

The ablation results presented in Table 8 confirm that introducing CNN convolution or C-ETB improves accuracy. Adding CNN (G-2 vs. G-1) reduces MPJPE/P-MPJPE by 0.55mm/0.51mm, indicating that CNN extracts key local features. Adding C-ETB (G-3 vs. G-1) reduces by 0.81mm/ 1.12mm, showing C-ETB's advantage in fusing global and local spatio-temporal features. G-4 with both CNN and C-ETB further reduces by 2.44mm/2.47mm, indicating that their combination yields optimal feature fusion and significant performance gain.

4.5.2. Different network structures

The DDCEFormer parallelizes S-ETB and T-ETB, then cascades them with C-ETB for cross-fusion, improving accuracy. Different structures can be built by selecting modules within DDCE-ETB. Ablation results for different structures are in Table 9.

From Table 9, using only S-ETB or T-ETB yields weaker performance, with S-ETB better than T-ETB. The last three rows show that the parallel connection of S-ETB and T-ETB with feature fusion significantly improves performance. Comparing three fusion strategies, our C-ETB weighted fusion outperforms direct concatenation fusion [23] and adaptive fusion [50], verifying the effectiveness of the EMCA cross mechanism in fusing global and local spatio-temporal features.

Tables 8 and 9 reveal that the performance improvement stems from the synergistic effect of the EMA mechanism and the overall architecture, with the latter contributing more significantly. The heterogeneous dimension scaling and the local modeling capability of the convolutional branch in EMA, combined with the dynamic weighting-based spatio-temporal feature modulation in C-ETB, form the core foundation for

Table 8
Ablation study on the C-ETB module and CNN convolution in EMA.

Serial Number	C-ETB Module	CNN Convolution	MPJPE (mm)	P-MPJPE (mm)
G-1	×	×	41.57	33.24
G-2	×	✓	41.02	32.73
G-3	✓	×	<u>40.76</u>	<u>32.12</u>
G-4	✓	✓	**39.13**	**30.77**

Table 9
Ablation study of different network structure models.

S-ETB	T-ETB	Direct Fusion	Adaptive Fusion	C-ETB	MPJPE (mm)	P-MPJPE (mm)
✓	×	×	×	×	47.72	37.07
×	✓	×	×	×	49.51	38.85
✓	✓	✓	×	×	41.02	32.73
✓	✓	×	✓	×	<u>40.09</u>	<u>32.06</u>
✓	✓	×	×	✓	**39.13**	**30.77**

Table 10
Ablation study for varying layer numbers L.

Layer Numbers L	Params (M)	Storage (M)	MPJPE (mm)	P-MPJPE (mm)
10	67.26	257	39.55	31.07
9	60.53	231	39.54	31.38
8	53.81	205	**39.13**	**30.77**
7	47.08	180	39.91	31.36
6	40.36	154	39.67	31.53
5	33.63	128	40.84	31.98

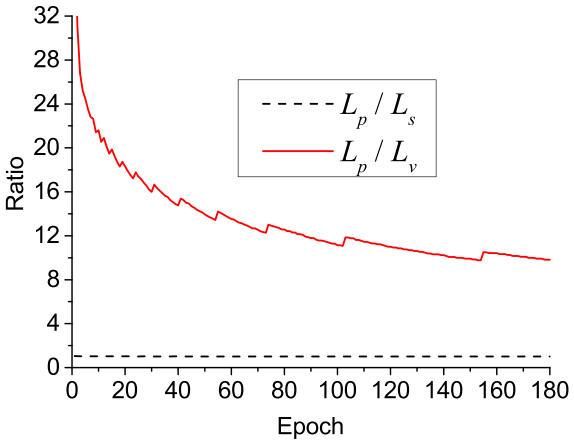


Fig. 6. Variation curves of the ratios L_p/L_s and L_p/L_v with epoch in the loss function.

the performance gain. In other words, EMA provides powerful feature representations, while the dual-domain cross-enhancement architecture establishes an efficient feature fusion framework. The combination of these two elements culminates in the final accuracy breakthrough.

4.5.3. Model layers and input sequence length

Table 10 presents ablation results for varying layer numbers L. As L increases, the number of parameters and the storage also increase, but MPJPE and P-MPJPE do not consistently decrease. Optimal performance is achieved at $L=8$.

Further experiments on the Human3.6M dataset demonstrate that as the input sequence length T decreases from 243 to 81 and 27, the MPJPE increases from 39.13mm to 42.52mm and 45.47mm, respectively, confirming the effectiveness of longer temporal contexts in improving pose estimation accuracy.

4.5.4. Weight coefficients in loss function

During the model training process for 180 epochs, the values of L_p, L_s and L_v in the loss function are recorded separately. The variation curves of the ratios L_p/L_s and L_p/L_v are shown in Fig. 6.

Based on the variation curves of the ratios L_p/L_s and L_p/L_v in Fig. 6, the ratio L_p/L_s fluctuates around 1. Therefore, the weight coefficient λ_s in the loss function is selected to be 0, 0.5, 1, and 1.5. Similarly, the ratio L_p/L_v varies between 9 and 32, so the weight coefficient λ_v in the loss function is selected to be 0, 10, 15, 20, 25, and 30. The ablation study results are shown in Table 11.

As illustrated in Table 11, incorporating either L_s or L_v effectively reduces MPJPE and P-MPJPE. Using only L_v is more effective than using only L_s, indicating that the velocity constraints play a more significant role in optimization. Analyzing different weight combinations: with fixed $\lambda_s=0.5$, adjusting λ_v from 0 to 30 gives the best performance at $\lambda_v=20$; with fixed $\lambda_v=20$, adjusting λ_s gives the best result at $\lambda_s=0.5$.

Table 11
Ablation study for weight coefficients in the loss function.

Serial Number	λ_s	λ_v	MPJPE (mm)	P-MPJPE (mm)
1	0.0	0.0	41.18	32.41
2	0.5	0.0	41.00	32.07
3	0.0	20.0	39.53	31.16
4	0.5	10.0	39.52	31.27
5	0.5	15.0	39.17	30.87
6	0.5	20.0	**39.13**	**30.77**
7	0.5	25.0	39.25	30.78
8	0.5	30.0	39.37	31.02
9	1.0	20.0	39.31	31.06
10	1.5	20.0	39.41	30.99

In summary, proper loss weighting significantly improves accuracy, with optimal MPJPE and P-MPJPE at $\lambda_s=0.5$, $\lambda_v=20$.

4.6. Visualization

The effectiveness of our proposed method is qualitatively demonstrated by comparing DDCEFormer with state-of-the-art approaches, including MHFormer, MixSTE, and FMFormer, on the Human3.6M dataset. The comparative results are presented in Fig. 7. In this figure, (a) shows the input 2D pose estimates, (b)-(e) illustrate the 3D pose estimations generated by the four methods, and (f) displays the corresponding ground-truth 3D poses. In the 3D estimation visualizations, the green, blue, and red lines represent the torso, lower limbs, and upper limbs, respectively.

A comparative analysis reveals that MHFormer, MixSTE, and FMFormer are prone to estimation errors when handling local occlusions. Typical error cases caused by occlusions of the upper or lower limbs are highlighted with purple arrows in Fig. 7(b)-(d). In contrast, by enhancing local feature representation and integrating spatio-temporal information, DDCEFormer effectively mitigates the interference caused by occlusions, thereby estimating more accurate 3D human poses.

We evaluate the robustness of DDCEFormer in real-world scenarios, with results shown in Fig. 8. The method estimates 3D human poses reasonably well under challenging conditions such as varying illumination, partial occlusions, and fast motion, as demonstrated in the first and second rows of Fig. 8. However, the accuracy of 3D pose prediction degrades significantly when the input 2D pose estimates contain large errors. As illustrated in the third row of Fig. 8, extensive occlusion of the legs by the skirt introduces inaccuracies in 2D joint detection, which subsequently compromises the quality of the final 3D pose estimation.

Furthermore, we reveal the global and local patterns captured by the EMA attention mechanism in DDCEFormer by visualizing and analyzing the spatial, temporal, and cross-attention maps across shallow, middle, and deep layers. The results are presented in Fig. 9.

Spatial attention evolves from simultaneously focusing on local joint connections and global structural relationships in shallow layers to refined attention with balanced local-global emphasis in deep layers. Temporal attention progressively expands its receptive field from adjacent frames in shallow layers to long-range temporal patterns in deep layers, while maintaining sparse characteristics throughout. Cross attention gradually strengthens from preliminary spatio-temporal fusion in shallow layers to high-level integration in deep layers, with consistent information flow across all layers. These visualizations demonstrate that all three attention mechanisms become increasingly focused with network depth, working synergistically to enhance 3D pose estimation accuracy.

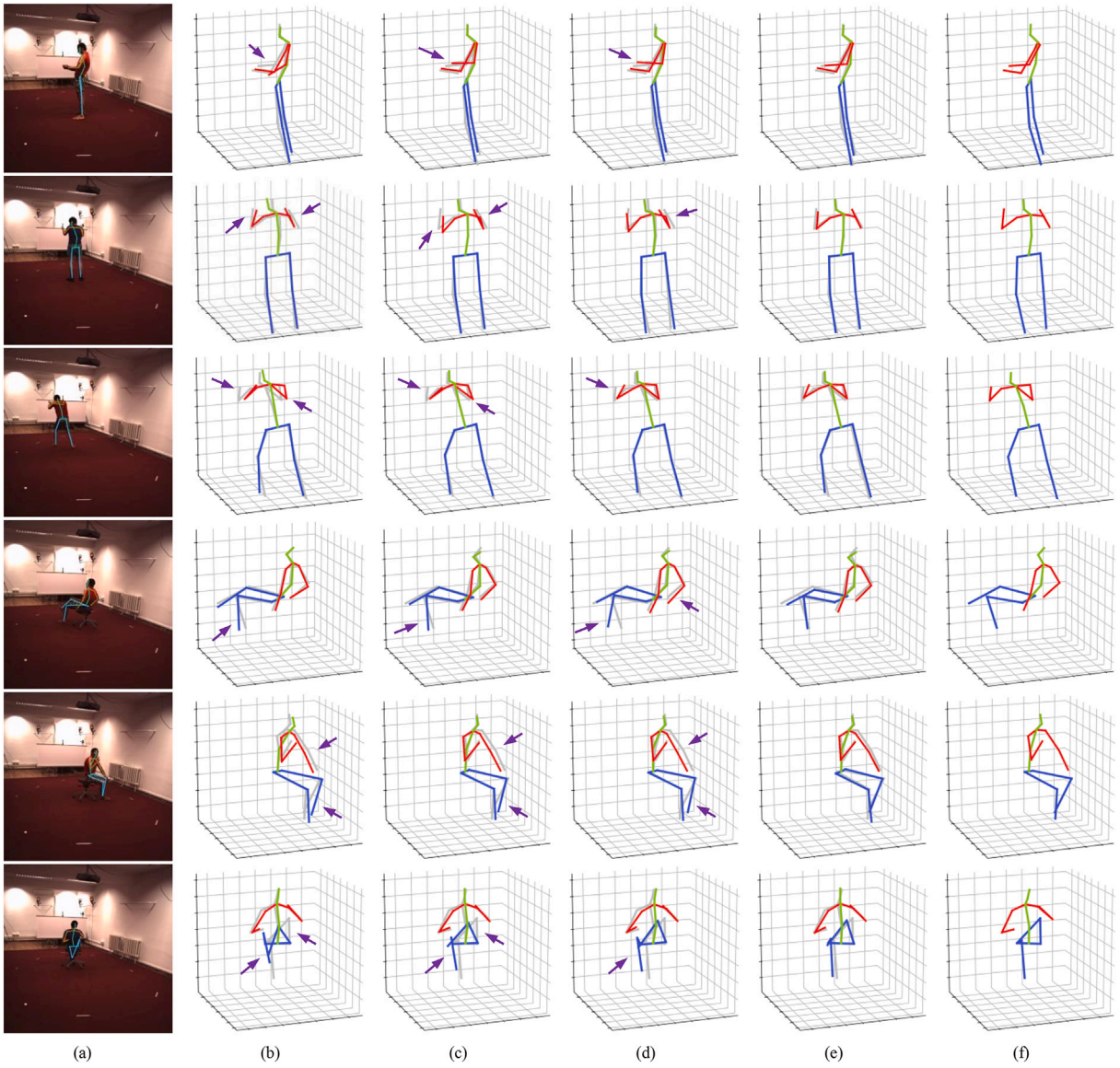


Fig. 7. Visual comparison of human pose estimation results across multiple methods. (a) Inputs. (b) MHFormer. (c) MixSTE. (d) FMFormer. (e) Ours. (f) Ground Truth. In (b)-(d), purple arrows highlight incorrectly estimated poses.

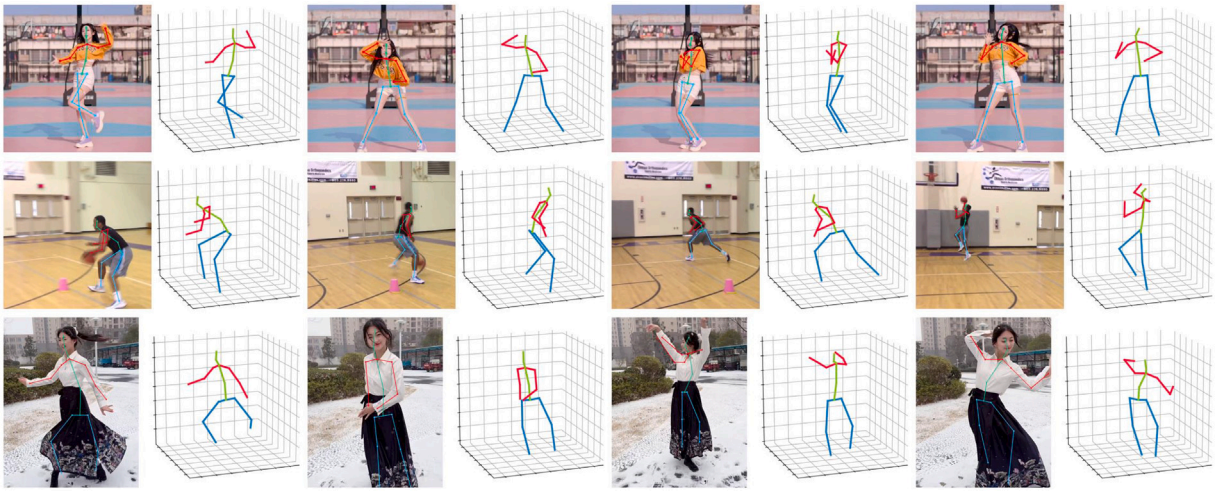


Fig. 8. Overview of qualitative results for the DDCEFormer algorithm from real-world scenarios.

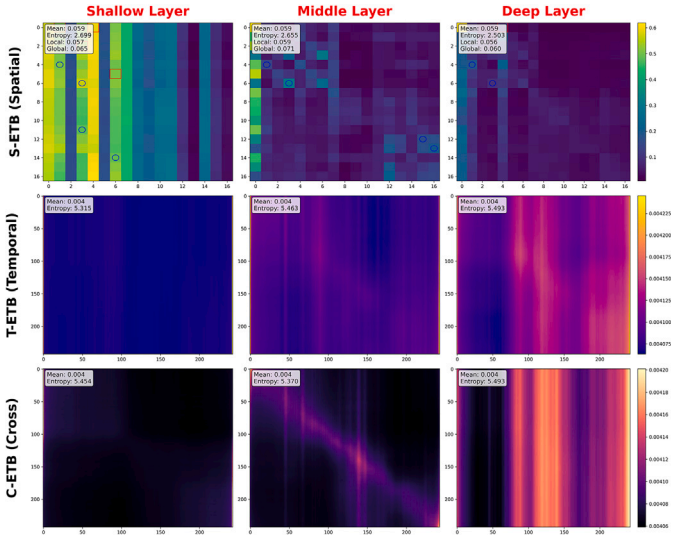


Fig. 9. Three layer visualization in EMA attention.

5. Conclusion

This paper proposes DDCEFormer, a novel approach for 3D human pose estimation from monocular videos. At its core lies an Enhanced Multi-head Attention (EMA) mechanism, which serves as the foundation of a parallel-serial hybrid architecture consisting of spatial (S-ETB), temporal (T-ETB), and cross (C-ETB) modules. Guided by a divide-and-conquer strategy, this architecture separately captures global dependencies and local details in both the spatial and temporal domains, and subsequently accomplishes deep fusion of spatio-temporal features through a cross-attention mechanism. Experimental evaluations on the Human3.6M and MPI-INF-3DHP datasets demonstrate that DDCEFormer achieves state-of-the-art performance in terms of overall estimation accuracy, with particularly strong modeling capabilities for complex actions involving large limb movements.

In summary, the main contributions of this work are the proposal and validation of a solution focused on the balanced and deep integration of global and local spatio-temporal features. The proposed DDCEFormer model, together with its core EMA mechanism, provides a novel technical paradigm for enhancing the accuracy of 3D human pose estimation. For future work, we will focus on developing adaptive methods for scaling factors, converting them into learnable parameters to achieve a dynamic balance between performance and efficiency. Furthermore, we intend to explore compression strategies such as structured pruning to reduce model complexity while maintaining estimation accuracy.

CRediT authorship contribution statement

Deliang Yang: Writing – original draft, Software, Funding acquisition, Data curation, Conceptualization. **Yanrong Ge:** Writing – review & editing, Supervision, Methodology. **Ning Xu:** Writing – review & editing, Visualization, Validation. **Rui Shi:** Writing – review & editing, Resources, Project administration.

Declaration of competing interest

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.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants 62403017, the Beijing Natural Science Foundation under Grant 4244088, and the Key Research Project of Beijing Polytechnic College under Grant BGY2022KY-01Z.

Data availability

Data will be made available upon request.

References

[1] Y. Liu, C. Qiu, Z. Zhang, Deep learning for 3D human pose estimation and mesh recovery: a survey, Neurocomputing. 596 (2024) 128049.
[2] Y. Zhang, B. Liu, J. Bao, Q. Huang, M. Zhang, J. Yu, Learnability matters: active learning for video captioning, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, 2024, pp. 37928–37954.
[3] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Trans. Image Process. 23 (2014) 2019–2032.
[4] P. Wang, W. Li, Z. Gao, C. Tang, P.O. Ogunbona, Depth pooling based large-scale 3D action recognition with convolutional neural networks, IEEE Trans. Multimed. 20 (2018) 1051–1061.
[5] R. Huo, Q. Gao, J. Qi, Z. Ju, 3D human pose estimation in video for human-computer/robot interaction, in: Proceedings of the International Conference on Intelligent Robotics and Applications, Springer, 2023, pp. 176–187.
[6] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, VNect: real-time 3D human pose estimation with a single RGB camera, ACM Trans. Graph. 36 (2017) 1–14.
[7] Y. Wang, P. Liu, H. Kang, D. Wu, D. Miao, ICFNet: interactive-complementary fusion network for monocular 3D human pose estimation, Neurocomputing. 616 (2025) 128947.
[8] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, J. Luo, Anatomy-aware 3D human pose estimation with bone-based pose decomposition, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 198–209.
[9] R. Liu, J. Shen, H. Wang, C. Chen, S.-C. Cheung, V. Asari, Attention mechanism exploits temporal contexts: real-time 3D human pose reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5064–5073.
[10] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 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, 2019, pp. 7753–7762.
[11] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272–2281.
[12] J. Wang, S. Yan, Y. Xiong, D. Lin, Motion guided 3D pose estimation from videos, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 764–780.
[13] M.R.I. Hossain, J.J. Little, Exploiting temporal information for 3D human pose estimation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 68–84.
[14] M. Lin, L. Lin, X. Liang, K. Wang, H. Cheng, Recurrent 3D pose sequence machines, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 810–819.
[15] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11656–11665.
[16] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, W. Yang, Exploiting temporal contexts with strided transformer for 3D human pose estimation, IEEE Trans. Multimedia 25 (2022) 1282–1293.
[17] W. Li, H. Liu, H. Tang, P. Wang, Multi-hypothesis representation learning for transformer-based 3D human pose estimation, Pattern Recognit. 141 (2023) 109631.
[18] J. Zhang, Z. Tu, J. Yang, Y. Chen, J. Yuan, 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, 2022, pp. 13232–13242.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[20] Y. Wang, H. Kang, D. Wu, W. Yang, L. Zhang, Global and local spatio-temporal encoder for 3D human pose estimation, IEEE Trans. Multimed. 26 (2024) 4039–4049.
[21] Y. Xue, J. Chen, X. Gu, H. Ma, H. Ma, Boosting monocular 3D human pose estimation with part aware attention, IEEE Trans. Image Process. 31 (2022) 4278–4291.
[22] 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 Recogn. 158 (2025) 110925.
[23] Z. Tang, Z. Qiu, Y. Hao, R. Hong, T. Yao, 3D human pose estimation with spatio-temporal criss-cross attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4790–4799.
[24] Y. Liu, Z. Zhang, STGFormer: spatio-temporal graphformer for 3D human pose estimation in video, Pattern Recognit. 171 (2026) 112239.
[25] Y. Zhong, G. Yang, D. Zhong, X. Yang, S. Wang, Frame-padded multiscale transformer for monocular 3D human pose estimation, IEEE Trans. Multimedia 26 (2024) 6191–6201.
[26] 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 (2014) 1325–1339.

[27] 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: Proceedings of the International Conference on 3D Vision (3DV), IEEE, 2017, pp. 506–516.
[28] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7307–7316.
[29] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, Integral human pose regression, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 529–545.
[30] K. Zhou, X. Han, N. Jiang, K. Jia, J. Lu, HEMlets PoSh: learning part-centric heatmap triplets for 3D human pose and shape estimation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2022) 3000–3014.
[31] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3D human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2640–2649.
[32] B. Wandt, B. Rosenhahn, RepNet: weakly supervised training of an adversarial reprojection network for 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7782–7791.
[33] C. Li, G.H. Lee, Generating multiple hypotheses for 3D human pose estimation with mixture density network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9887–9895.
[34] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 483–499.
[35] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7103–7112.
[36] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, AlphaPose: whole-body regional multi-person pose estimation and tracking in real-time, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 7157–7173.
[37] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.
[38] X. Zhou, M. Zhu, S. Leonardos, K.G. Derpanis, K. Daniilidis, Sparseness meets deepness: 3D human pose estimation from monocular video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 4966–4975.
[39] G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3D human pose, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 7025–7034.
[40] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, A. Jain, Learning 3D human pose from structure and motion, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 668–683.
[41] T. Xu, W. Takano, Graph stacked hourglass networks for 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16105–16114.
[42] W. Hu, C. Zhang, F. Zhan, L. Zhang, T.-T. Wong, Conditional directed graph convolution for 3D human pose estimation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 602–611.
[43] R.B. Neupane, K. Li, T.F. Boka, A survey on deep 3D human pose estimation, Artif. Intell. Rev. 58 (2024) 24.
[44] M. Einfalt, K. Ludwig, R. Lienhart, Uplift and upsample: efficient 3D human pose estimation with uplifting transformers, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2903–2913.
[45] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, W. Gao, P-STMO: pre-trained spatial temporal many-to-one model for 3D human pose estimation, in: Proceedings of the European Conference on Computer Vision, Springer, 2022, pp. 461–478.
[46] W. Li, H. Liu, H. Tang, P. Wang, L. Van Gool, MHFormer: multi-hypothesis transformer for 3D human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13147–13156.
[47] Y. Wang, M. Li, N. Meng, M. Xu, Optimizing the temporal adjacency matrix for 3D human pose estimation through clustering, Neurocomputing. 653 (2025) 131247.
[48] A. Diaz-Arias, D. Shin, ConvFormer: parameter reduction in transformer models for 3D human pose estimation by leveraging dynamic multi-headed convolutional attention, The Vis. Comput. 40 (2024) 2555–2569.
[49] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, UniFormer: unifying convolution and self-attention for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 12581–12600.
[50] W. Zhu, X. Ma, Z. Liu, et al , MotionBERT: a unified perspective on learning human motion representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15085–15099.

Author biography

Deliang Yang received the Ph.D. degree in pattern recognition and intelligent systems from Beijing University of Technology, Beijing, China. He conducted postdoctoral research with the Department of Automation, Tsinghua University, Beijing, China. He is currently an associate professor at Beijing Polytechnic College, Beijing, China. His research interests include human pose estimation, deep learning, and computer vision.

Yanrong Ge received the Ph.D. degree in pattern recognition and intelligent systems from Beijing University of Technology, Beijing, China. She is currently an associate professor with Hebei Normal University, China. Her research interests include human pose estimation, deep learning, and multi-agent collaborative control systems.

Ning Xu is currently pursuing a master's degree at the College of Physics, Hebei Normal University, China. Her research interests include human-centered computer vision and image processing.

Rui Shi received his Ph.D. degree in graphic and computer sciences from the University of Tokyo, Tokyo, Japan, in 2022. He is currently an associate professor in the School of Information Science and Technology, Beijing University of Technology, Beijing, China. He worked as a visiting researcher in the Department of General Systems Studies, the University of Tokyo. His current research interests include neural networks, autonomous driving, and explainable artificial intelligence.