Multimedia Systems (2024) 30:243
https://doi.org/10.1007/s00530-024-01451-4

RESEARCH

Exploring multi-level transformers with feature frame padding network for 3D human pose estimation

Sathiyamoorthi Arthanari[1] · **Jae Hoon Jeong**[1] · **Young Hoon Joo**[1]

Received: 19 June 2024 / Accepted: 6 August 2024

Abstract
Recently, transformer-based architecture achieved remarkable performance in 2D to 3D lifting pose estimation. Despite advancements in transformer-based architecture they still struggle to handle depth ambiguity, limited temporal information, lacking edge frame details, and short-term temporal features. Consequently, transformer architecture encounters challenges in preciously estimating the 3D human position. To address these problems, we proposed Multi-Level Transformers with a Feature Frame Padding Network (MLTFFPN). To do this, we first propose the frame-padding network, which allows the network to capture longer temporal dependencies and effectively address the lacking edge frame information, enabling a better understanding of the sequential nature of human motion and improving the accuracy of pose estimation. Furthermore, we employ a multi-level transformer to extract temporal information from 3D human poses, which aims to improve the short-range temporal dependencies among keypoints of the human pose skeleton. Specifically, we introduce the Refined Temporal Constriction and Proliferation Transformer (RTCPT), which incorporates spatio-temporal encoders and a Temporal Constriction and Proliferation (TCP) structure to reveal multi-scale attention information and effectively addresses the depth ambiguity problem. Moreover, we incorporate the Feature Aggregation Refinement (FAR) module into the TCP block in a cross-layer manner, which facilitates semantic representation through the persistent interaction of queries, keys, and values. We extensively evaluate the efficiency of our method through experiments on two well-known benchmark datasets: Human3.6M and MPI-INF-3DHP.

Keywords 3D human pose estimation · Frame-padding network · Spatio-temporal transformer · Temporal constriction and proliferation transformer · Feature aggregation refinement module

1 Introduction

In the development of computer vision, several fields play a crucial role, including object tracking, and 3D human pose estimation. Object tracking involves continuously identifying and following the trajectory of specific objects across video frames, crucial for applications from autonomous vehicles to surveillance [1–6]. On the other hand, human pose estimation is a fundamental task in computer vision that aims to infer the 3D positions of a human body from images or videos [7–9]. In recent years, advanced deep learning techniques have led to significant progress in this field, particularly with the emergence of 3D human pose estimation. It has achieved significant attention due to its pivotal role in diverse applications, including virtual reality, action recognition, and human-robot interaction. Specifically, the 3D human pose estimation offers numerous distinct advantages. It allows for more accurate and robust gesture recognition, enhancing human-computer interaction experiences. In robotics, it improves human-robot collaboration by enabling robots to better understand and respond to human actions. For autonomous vehicles, it enhances safety by accurately predicting pedestrian movements. Overall, 3D pose estimation drives innovation and efficiency across various sectors by offering a deeper understanding of human movement. The 3D pose estimation field can be categorized into two primary

Communicated by Yongdong Zhang.

✉ Young Hoon Joo
yhjoo@kunsan.ac.kr

Sathiyamoorthi Arthanari
sathyainfotech005@gmail.com

Jae Hoon Jeong
jh7129@kunsan.ac.kr

[1] School of IT Information and Control Engineering, Kunsan National University, 558 Daehak-ro, Gunsan-si, Jeollabuk-do 54150, South Korea

Published online: 13 August 2024

approaches: the end-to-end approach and the lifting-based approach. The end-to-end approach aims to directly predict the 3D coordinates of human body joints from input images or video frames without relying on intermediate representations. In contrast, lifting approaches employ a two-stage pipeline where the initial stage extracts 2D keypoints, and the subsequent stage lifts the 2D coordinates into 3D space. Nowadays, several cutting-edge approaches embrace the 2D to 3D lifting-based techniques. In this study, we adopt the lifting-based approach due to its capacity to utilize well-established and precise 2D pose detectors, which simplifies the process of inferring 3D human posture from easily accessible 2D keypoint annotations. However, 3D human pose estimation presents unique challenges due to the inherent ambiguity in mapping 2D image features to 3D space, occlusions, depth ambiguity, and appearance variations, as well as the complexity of human motion. To address these challenges, human pose estimation has seen the development of techniques, with graph convolutional networks and transformer-based approaches playing pivotal roles.

Graph Convolutional Networks (GCNs) have emerged as a powerful tool in the field of 3D human pose estimation, revolutionizing the way researchers approach this complex problem. At the intersection of computer vision and machine learning, GCNs offer novel approaches [10–13] that leverages the inherent structural relationships within the human body to accurately infer 3D poses from 2D images or video frames. In this regard, the authors [10] have proposed a GraphMLP architecture that demonstrates robustness to noise and occlusions commonly encountered in real-world scenarios. By leveraging MLPs, which are inherently robust to noisy inputs, GraphMLP can effectively handle imperfect data and partial observations, enhancing the reliability of 3D pose estimation. In addition, the authors in [11] have presented the HPGCN method, which effectively learns discriminative features from hierarchical poselets, and empowers the model to adeptly capture intricate details and subtle variations in human poses. By leveraging both local and global context information encoded in poselets, the model achieves superior performance in 3D pose estimation tasks. Specifically, the GLA-GCN [12] approach have introduced that incorporates a global–local adaptive representation scheme, which dynamically adjusts the receptive fields of GCN based on the spatial relationships between body parts. This adaptive mechanism allows the model to focus on global context information and local details, enhancing the accuracy and robustness of 3D pose estimation. Moreover, the RS-Net [13] approach introduces a learnable modulation matrix that effectively adds extra edges to the skeleton graph, which helps the network discover additional connections between body parts that might be crucial for accurate pose estimation. Despite GCN-based approaches improving performance, they still face challenges in effectively handling self-occlusion and long-range dependencies. Transformer-based methods excel in addressing these concerns by effectively modeling the long-range dependencies among body joints.

On the other hand, transformer-based approaches outperform CNN methods in 3D human pose estimation due to their ability to capture long-range dependencies and complex relationships among body joints. By utilizing self-attention mechanisms, transformers can dynamically focus on relevant parts of the input, which is crucial for accurately predicting the positions of human joints in 3D space. Additionally, the architecture's inherent parallel processing capabilities facilitate efficient handling of high-dimensional data, resulting in faster training and inference times. This unified approach allows for simultaneous modeling of spatial and temporal information, enhancing the overall prediction accuracy and robustness. The versatility and efficiency of transformers make them a powerful tool for advancing the state-of-the-art in 3D human pose estimation. Researchers from various fields have increasingly focused on transformer-based methodologies [7–9, 14–19], recognizing their potential to revolutionize numerous domains. In this context, the PoseFormer [7] approach employs a hierarchical representation of the human body, where the input sequence is divided into smaller segments corresponding to different body parts or joints. This hierarchical structure helps the model to focus on specific regions of interest while maintaining a global understanding of the entire pose. Furthermore, the MixSTE [8] method adopts a sequence-to-sequence architecture, consisting of an encoder and a decoder. The encoder processes input video sequences, while the decoder generates corresponding 3D human pose estimations. However, MixSTE may face challenges in accurately capturing long-range dependencies and temporal dynamics in video sequences, leading to suboptimal pose estimations. MHFormer [9] addresses these issues through its multi-head attention mechanism, which enables the model to effectively capture both spatial and temporal information simultaneously. Specifically, the MotionAGFormer [14] incorporates an attention-based occlusion handling mechanism. This mechanism helps the network focus on reliable information and mitigate the effects of occlusions and depth ambiguity, leading to more comprehensive and accurate pose estimations. Moreover, the TransVOD [15] introduces an innovative approach to video object detection by leveraging Spatial-Temporal Transformers, which effectively model both spatial and temporal features across video frames. This method enhances detection accuracy by capturing complex interactions between objects over time, leading to improved performance in dynamic environments. Additionally, the authors in [16] have presented a novel local–global transformer neural network for temporal action segmentation, offering improved accuracy by capturing both local fine-grained details and global

contextual information. It enhances robustness to varying action durations and complex temporal dependencies, outperforming existing methods. In this study, we employ a transformer-based architecture as our baseline model for 3D human pose estimation, leveraging its potent capacity to model sequential data.

Specifically, the MixSTE [8] method utilizes the seq2seq feature encoder to extract the temporal information from the 2D keypoints, which enables the estimation of 3D pose keypoints from a sequence of corresponding 2D inputs. Particularly, a proficient alternating spatio-temporal feature extraction method was devised to merge the spatio-temporal features of human joints, leading to significant enhancement in the network's performance. In the task of estimating 3D human pose keypoints from video frames, the initial step involves segmenting the entire video sequence into shorter segments for separate processing. Within each segment, temporal information is obtained from the frame positioned at its midpoint. Nonetheless, the frames at the beginning and end of each short sequence lack comprehensive temporal information, resulting in diminished execution of the seq2seq method. To address the above issue, we propose a feature frame padding network with a multi-level transformer, which effectively captures the temporal information at the start and end of the sequence. Moreover, we integrate the Temporal Constriction and Proliferation Transformer (TCP) and Feature Aggregation Refinement (FAR) module to expose multi-scale attention information and effectively obtain the long-range temporal dependencies in the motion sequences. The main contribution of the proposed method is described as follows:

1. We present a novel feature frame-padding approach designed to enhance the efficiency of the seq2seq method, especially in managing edge frames within the 2D-to-3D lifting framework. This innovative technique addresses challenges related to incomplete temporal context by extending the input sequence to include additional frames, thereby capturing more comprehensive motion information. By integrating this method, we ensure that the model can better understand and predict human motion, leading to more precise and reliable pose predictions.
2. A multi-level spatio-temporal transformer approach is presented, which captures both fine-grained and high-level spatial information and temporal dynamics. By integrating features from various levels of the spatial and temporal hierarchy, our method provides a comprehensive understanding of human motion. This ensures more accurate and robust 3D pose estimation by combining detailed local movements with broader temporal patterns, leading to significantly enhanced performance in capturing complex human poses.
3. The Refined Temporal Constriction and Proliferation Transformer (RTCPT) introduces a unique architecture that incorporates spatio-temporal encoders with a Temporal Constriction and Proliferation (TCP) structure to reveal multi-scale attention information and effectively address the depth ambiguity problem. Specifically, this approach effectively captures both short-range and long-range temporal dependencies in motion sequences, leading to more accurate 3D pose predictions.
4. We propose the Feature Aggregation Refinement (FAR) module, which is integrated into the TCP block to enhance semantic representation through the persistent interaction of queries, keys, and values. By facilitating a more nuanced and refined aggregation of features, the FAR module contributes to a better understanding and representation of the human pose.
5. We conduct comprehensive experiments on benchmark datasets such as Human3.6M and MPI-INF-3DHP, showcasing the superior performance of our proposed method compared to other state-of-the-art approaches (Fig. 1).

2 Related works

In this section, we present a brief overview of human pose estimation. First, we examine the 3D human pose estimation in Sect. 2.1. Next, we describe the seq2frame and seq2seq approaches in Sect. 2.2. Finally, we explore transformer-based architectures in Sect. 2.3.

2.1 3D human pose estimation

In recent years, 3D human pose estimation has become increasingly essential in computer vision. This task involves analyzing images from single or multiple perspectives to determine the 3D positions of human joints or body parts. To accomplish this goal various techniques have been developed including one-stage detectors and two-stage detectors. These methods facilitate the classification of human poses by accurately identifying the positions of key joints and body parts in 3D space. The one-stage detectors [20, 21] predict 3D body joint coordinates directly from raw image data, whereas the two-stage detectors involve two intermediary steps. In the initial step, we calculate 2D keypoints based on the input image. In the subsequent phase, we exploit the relationship between 2D and 3D human pose to transform the initially estimated 2D keypoints into corresponding 3D positions. Due to advancements in robust 2D detection, recent 2D-to-3D lifting methods [7–9, 14] have exhibited remarkable performance than end-to-end approaches. Moreover, the authors in [22, 23] have presented the unified multi-modal unsupervised representation

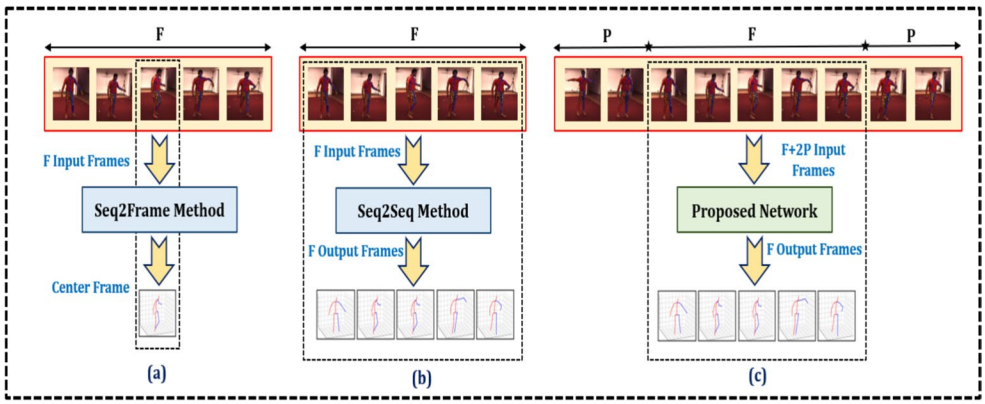

Fig. 1 The general pipeline of 3D pose estimation. **a** block denotes the Seq2Frame approach. It takes F frames as input to estimate the center frame of human position. **b** block represents the Seq2Seq approach, with F input frames and F output frames. **c** block shows that our proposed method has $F + 2P$ input frames and F output frames

learning aimed at skeleton-based action understanding and dense object grounding for 3D pose estimation, which significantly improves the accuracy of the 3D pose in complex environments. Inspired by the above discussion, we adopt a two-stage pipeline for 3D pose estimation as it proves to be a widely employed and successful method, consistently surpassing single-stage approaches.

2.2 Seq2Frame and Seq2Seq methods

3D human pose estimation is traditionally focused on estimating the 3D location of body joints in a single image. However, this static approach doesn't capture the essence of human movement. Seq2frame methods [24] emerge as a powerful approach that leverages sequence information to estimate 3D poses across multiple video frames, providing a more dynamic understanding of human motion. Specifically, the authors in [24] have proposed a novel approach for 3D human pose estimation in video. It leverages a fully convolutional model with dilated temporal convolutions to capture temporal information from 2D keypoint trajectories. Moreover, transformer-based approaches [7, 9] have employed the seq2frame technique to improve the performance of the model. However, the Seq2Frame technique in 3D human pose estimation might encounter challenges in handling long-range dependencies across frames, limiting its ability to capture complex temporal dynamics effectively. The seq2seq techniques address long-range dependencies by employing recurrent neural networks or transformer architectures [8, 25], which can capture information from distant time steps through their sequential processing mechanisms or self-attention mechanisms, respectively. In particular, the MixSTE [8] seq2seq model offers enhanced robustness to occlusions and ambiguities in pose estimation by leveraging a mixture of spatiotemporal encodings, enabling more accurate and reliable pose predictions. Following that, the Diff3DHPE [25] utilizes a diffusion process but avoids the limitations of seq2frame methods by employing a separate backbone model to handle the seq2seq aspect of pose estimation across a video sequence. These advancements demonstrate significant strides in both computational efficiency and accuracy within the realm of 3D human pose estimation. Nevertheless, the seq2seq model faces issues related to feature loss. To tackle these constraints, our approach merges the strengths of the seq2seq and seq2frame methods, which enhance edge feature extraction (Fig. 2).

2.3 Transformer-based architecture

The application of transformer architectures in 3D human pose estimation has emerged as a groundbreaking approach [26–30], revolutionizing the field by offering novel perspectives and enhanced capabilities. Traditionally, human pose estimation from images or video sequences has predominantly relied on convolutional neural networks (CNNs) or recurrent neural networks (RNNs), which excel in capturing spatial or temporal dependencies, respectively. However, these conventional methods often struggle with effectively modeling long-range dependencies and capturing intricate spatial-temporal relationships simultaneously. To tackle this challenge, transformer architecture integrates a self-attention mechanism. This feature enables the model to directly focus on any part of the input data, effectively capturing the long-range dependencies among body joints essential for precise 3D pose estimation. The authors in [27] have utilized End-to-end video object detection with spatial-temporal transformers, which efficiently capture and process temporal information across frames, enhancing object detection accuracy by understanding motion patterns and contextual relationships. In particular, the authors in [28] have proposed the Strided Transformer

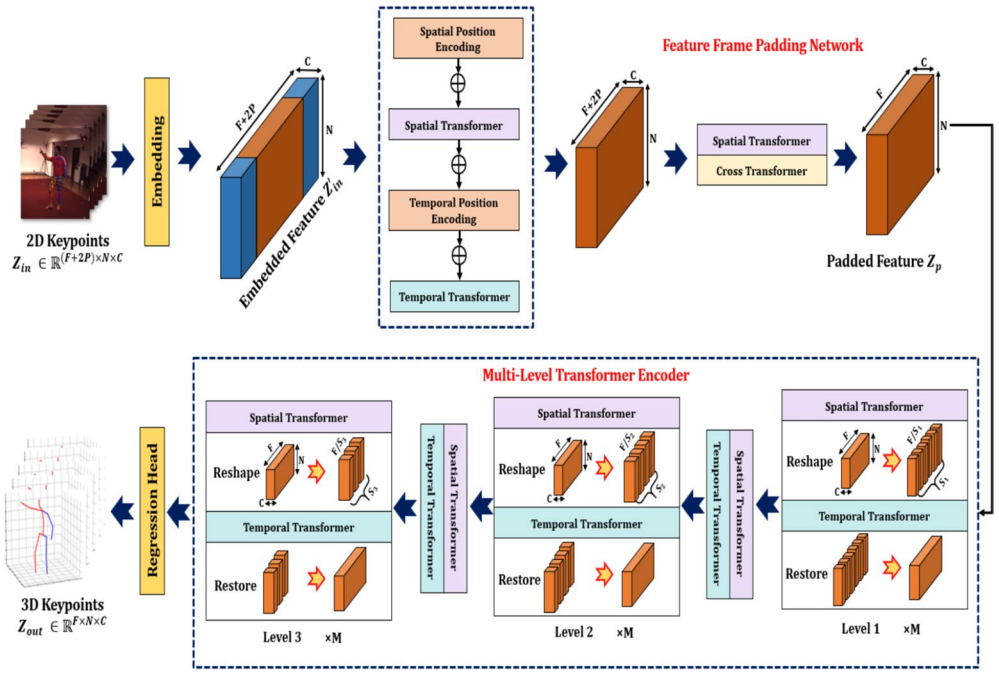

Fig. 2 The schematic diagram of proposed approach includes a feature frame padding network and multi-level spatio-temporal encoders. The detailed block diagram of the spatial and temporal transformer is illustrated in Fig. 4

approach, which effectively integrates temporal contexts into the pose estimation task. By leveraging the strided transformer architecture, it captures long-range temporal dependencies in the input sequence. Specifically, the authors in [29] introduced the HDFormer method, which incorporates multi-scale fusion mechanisms to integrate information from different levels of spatial and temporal granularity. By combining features extracted at multiple resolutions, the model gains a holistic understanding of the input pose sequence, effectively capturing both fine-grained details and global context. Moreover, the D3DP [30] method introduces a multi-hypothesis aggregation scheme, which allows the model to consider multiple plausible hypotheses during pose estimation. By incorporating diverse hypotheses and their corresponding confidence scores, the model can explore a wider range of pose configurations and make more informed decisions. In this study, we propose a multi-level spatio-temporal constriction and proliferation transformer to expose multi-scale attention information and improve the short-term correlations of human poses.

3 Proposed method and implementation

3.1 Feature frame padding network with multi-level temporal transformer approach

Initially, we utilized the 2D human pose keypoints as input $Z_{in} \in \mathbb{R}^{(\mathcal{F}+2\mathcal{P})\times\mathcal{N}\times 2}$, which contains $\mathcal{F}+2\mathcal{P}$ input frames, $\mathcal{N}$ denotes the number of joints, and 2 represents the coordinates per joint. To begin, we transform the input 2D coordinates into high-dimensional features represented as $Z_{in} \in \mathbb{R}^{(\mathcal{F}+2\mathcal{P})\times\mathcal{N}\times 2}$. Specifically, we utilize a specialized structure called the spatio-temporal transformer, where both the spatial encoder and temporal encoder are integrated into a single layer within our network architecture. Furthermore, we employed a cross-attention transformer method to obtain temporal dependencies, decreasing the input sequence to the original length, resulting in $Z_p \in \mathbb{R}^{\mathcal{F}\times\mathcal{N}\times\mathcal{C}}$, which matches the length prior to padding. Following that, the Z_p is forwarded to the multi-level spatio-temporal transformer encoder to obtain the short-range

and long-range temporal dependencies. To capture the final output $Z_{out} \in \mathbb{R}^{\mathcal{F}\times\mathcal{N}\times 3}$, we employ the regression head.

3.1.1 Feature frame padding network with cross-attention method

Due to the lack of temporal information on the edge frame in the seq2seq pipeline, the transformer encoder struggles to predict the accurate 3D pose in the sequence. To address this problem, we present the feature frame padding network model with a cross-attention transformer manner. The architecture of the cross-attention transformer is exhibited in Fig. 3b. Initially, we expand the input sequence $Z_o \in \mathbb{R}^{\mathcal{F}\times\mathcal{N}\times 2}$ by adding two frames on both ends. This augmentation yields the input $Z \in \mathbb{R}^{(\mathcal{F}+2\mathcal{P})\times\mathcal{N}\times 2}$. Moreover, we introduce cross-attention mechanism to reduce extended sequence while retaining its original length after padding. Moreover, the input of the cross-attention transformer is determined as follows:

$$Input' = Input[P : -P] \tag{1}$$

$$Q = Input'\mathcal{W}_Q, K = Input\mathcal{W}_K, V = Input\mathcal{W}_V \tag{2}$$

$$Output = MSA + Input'. \tag{3}$$

where $Input \in \mathbb{R}^{(\mathcal{F}+2\mathcal{P})\times\mathcal{N}\times\mathcal{C}}$ denotes the global feature with padded frames, $Input' \in \mathbb{R}^{\mathcal{F}\times\mathcal{N}\times\mathcal{C}}$ matches to *input*, which is not padded in the input sequence. Then, the $\mathcal{W}_Q \in \mathbb{R}^{d\times d}$, $\mathcal{W}_K \in \mathbb{R}^{d\times d}$, and $\mathcal{W}_V \in \mathbb{R}^{d\times d}$ denotes the linear layer. Like the decoder in a transformer model for natural language processing [31], we represent the Keys and Values as *Input* and the Queries as *Input*′. This facilitates the mapping of features with a length of $\mathcal{F}+2\mathcal{P}$ to features with a length of $\mathcal{F}$ during the attention computation. By utilizing matrix operations within the Multi-Head Self-Attention (MSA), we generate an output shape that matches the shape of the Queries. Consequently, the frames at the edges capture the temporal features embedded within the padded frames.

3.1.2 Self-attention transformer

Self-attention allows the model to consider the relationships between all pairs of keypoints simultaneously, capturing global context information. This is crucial for understanding the spatial dependencies among different parts of the body in 3D pose estimation. The self-attention transformer is illustrated in Fig. 3a. Specifically, the adaptive self-attention mechanisms can dynamically adjust the attention weights based on the input data. In the context of 3D human pose estimation, this adaptability enables the model to handle varying poses, body shapes, and environmental conditions effectively. The attention mechanism is defined as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where $\{Q, K, V\} \in \mathbb{R}^{n\times d}$, n represents number of joints and d indicates the feature channels. In prior studies [7, 9], all joints are treated as a token in the temporal transformer. This

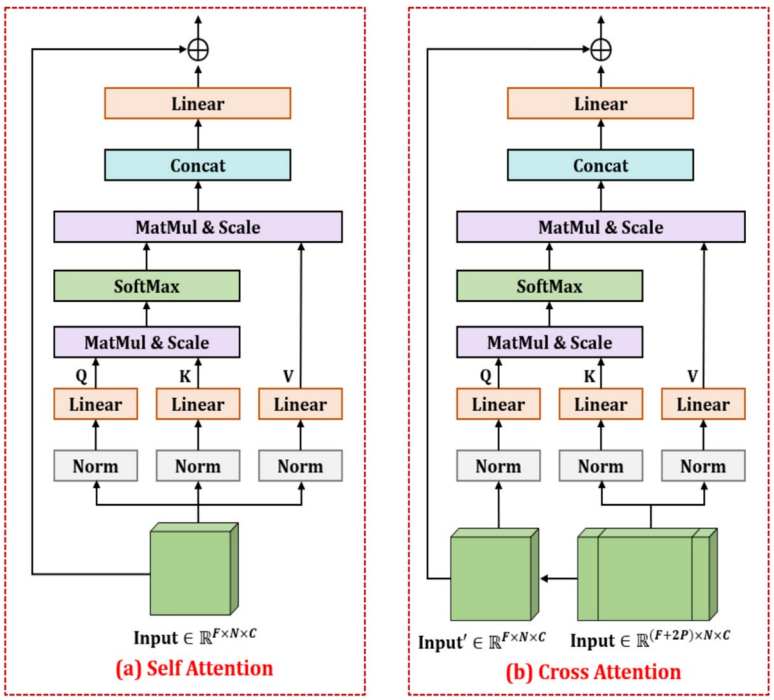

Fig. 3 Comparison of the self-attention and cross-attention transformers

approach extracts the spatial features, leading to increases in the complexity of temporal feature extraction. In this work, we employ the spatial transformer to obtain dependencies among distinct joints. Moreover, we utilize the multihead self-attention approach, which allows the model to capture complex relationships between different joints and body parts simultaneously. This is crucial in understanding the spatial dependencies and interactions within the human body, leading to more accurate pose estimations. The self-attention mechanism is defined as follows:

$$MSA = Concat(head_1, \dots, head_n)W^P \tag{5}$$

$$head_i = Attention(Q_i, K_i, V_i) \tag{6}$$

where, $W^P \in \mathbb{R}^{d\times d}$ represents the linear projection matrix. Then, we used the multi-layer perception, which is defined as follows:

$$MLP(x) = \sigma(xW_1 + b_1)W_2 + b_2 \tag{7}$$

where σ indicates the activation function, $W_1 \in \mathbb{R}^{d\times d'_m}$ and $W_2 \in \mathbb{R}^{d'_m\times d}$; $b_1 \in \mathbb{R}^{d'_m}$ and $b_2 \in \mathbb{R}^d$ indicate the fully connected layers.

3.1.3 Spatial transformer encoder

The spatial transformer encoder adeptly captures the inter-joint relationships within the human skeleton for every frame. The joint matrix for each frame is considered as a spatial attention token denoted by $Z_s \in \mathbb{R}^{F\times N\times C}$. The tokens are subsequently integrated with a spatial position matrix $\mathcal{E}_s \in \mathbb{R}^{F\times N\times C}$ and introduced into the key components of the transformer model such as multi-layer perceptron and multi-head self-attention, as explained in [31]. The dimensions of the tokens remain constant after the feature extraction by the spatial transformer encoder. This process is defined as follows:

$$\begin{aligned}\widetilde{\mathrm{ST}}(Z_s) &= Z_s + \mathrm{MSA}(Z_s),\\ \mathrm{ST}(Z_s) &= \widetilde{\mathrm{ST}}(Z_s) + \mathrm{MLP}\Big(\widetilde{\mathrm{ST}}(Z_s)\Big).\end{aligned} \tag{8}$$

The spatial position encoding $\mathcal{E}_s$ is embedded into the Z_s.$MSA(\cdot)$ and $MLP(\cdot)$.

3.1.4 Temporal transformer encoder

The temporal transformer encoder utilizes the feature aggregation refinement module, which analyzes the trajectory of each joint across input frames and leverages the attention mechanisms to capture multi-scale information, enhancing the understanding of joint movements. The joints are segmented into separate tokens $Z_t \in \mathbb{R}^{F\times N\times C}$ in the temporal encoder. Afterward, a temporal positional encoding $\mathcal{E}_t \in \mathbb{R}^{F\times N\times C}$ is embedded to the input token before being passed into TTE. The process is formulated as follows:

$$\begin{aligned}\widetilde{\mathrm{TT}}(Z_t) &= Z_t + \mathrm{FAR}(Z_{t-1}, Z_t),\\ \mathrm{TT}(Z_t) &= \widetilde{\mathrm{TT}}(Z_t) + \mathrm{MLP}\Big(\widetilde{\mathrm{TT}}(Z_t)\Big).\end{aligned} \tag{9}$$

In Eq. (9), FAR$(\cdot)$ denotes the feature aggregation refinement module, which receives the input tokens from both the current temporal block Z_t and the previous temporal block Z_{t-1}. Specifically, the FAR module is employed to aggregate features within the attention block, facilitating the learning of more comprehensive information.

3.1.5 Multi-level temporal constriction & proliferation transformer

The 3D pose estimation performance is influenced by the length of input frames, affecting both seq2seq and seq2frame pipelines. Despite the exceptional ability of transformers to capture global dependencies, many prior works overlook the short-term correlation among video frames. To enhance comprehensive extraction of information from the self-attention layer, we introduce a multi-level temporal constriction and proliferation attention block, which aims to investigate the multi-scale information inherent in keys $K \in \mathbb{R}^{n\times d}$ and values $V \in \mathbb{R}^{n\times d}$. The dimensionality of queries $Q \in \mathbb{R}^{n\times d}$ remains constant, while processing is applied to keys and values across multiple stages. This approach allows the attention matrix to learn multi-scale information and enhance the short-term dependencies while maintaining consistent temporal resolution.

The TCP attention module has shown its effectiveness in various tasks by reducing redundancy and capturing high-level semantic information while preserving low-level details. To achieve more refined representations for keys and values, we utilize the temporal constriction and proliferation network to augment intra-block exploration. The overall architecture of the TCP block illustrated in Fig. 4(2.1). Given an input feature vector $z \in \mathbb{R}^{n\times d}$ with a sequence length n and channel dimension d, the TCP attention block generates an output feature vector with the same dimensions. The U-shaped temporal attention operation is denoted as TCP$(\cdot)$. It can be expressed as follows:

$$\begin{aligned}z^0_{\text{down}} &= z, \quad z^{l+1}_{\text{down}} = \sigma\big(\mathrm{LN}\big(\mathcal{F}_{\text{down}}\big(z^l_{\text{down}}\big)\big)\big),\\ z^0_{\text{up}} &= z^m_{\text{down}}, \quad z^{l+1}_{\text{up}} = \sigma\Big(\mathrm{LN}\Big(\mathcal{F}_{\text{up}}\Big(z^l_{\text{up}}\Big)\Big)\Big) + z^{m-1-l}_{\text{down}},\end{aligned} \tag{10}$$

where $\sigma(\cdot)$ denotes the activation function, LN$(\cdot)$ represents the LayerNorm layer, and $\mathcal{F}_{down}$ and $\mathcal{F}_{up}$ denote the constriction and proliferation functions, respectively. $z^l_{\text{down}} \in \mathbb{R}^{\frac{n}{l'}\times d}$

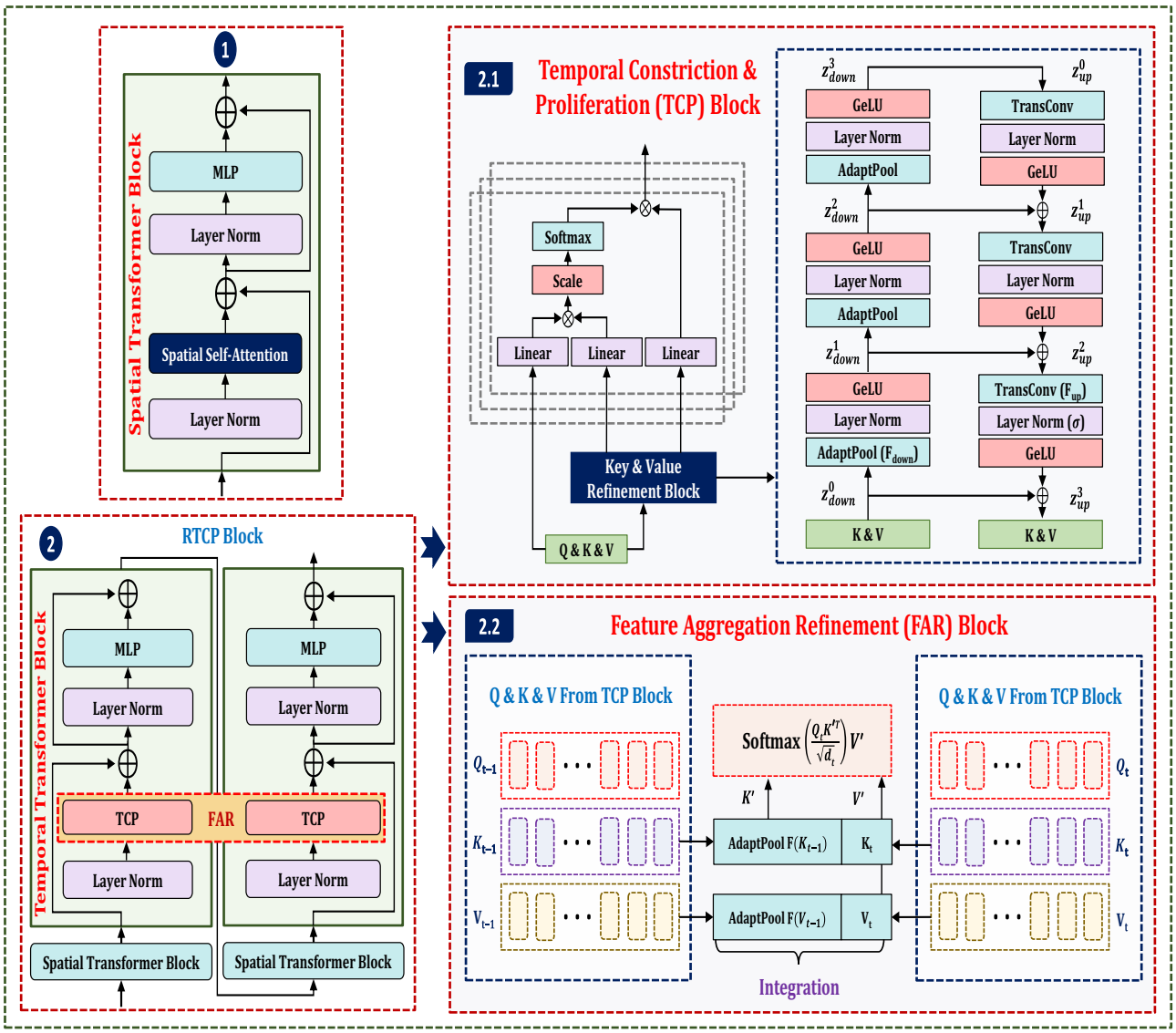

Fig. 4 The schematic diagram displays the spatial and refined temporal constriction and proliferation (RTCP) transformers. Section 1 exhibits the spatial transformer block, and Sect. 2 illustrates the temporal transformer block. Furthermore, Sects. 2.1 and 2.2 present the temporal constriction and proliferation (TCP) transformer block and the feature aggregation refinement (FAR) module, respectively

indicates the constrictions process and $z_{up}^{l} \in \mathbb{R}^{\frac{n}{r^{m-l}} \times d}$ signifies the proliferation process, $l \in [0, 1, \ldots, m-1]$ is the index of the sampling stage, and $\mathrm{TCP}(z) = z_{up}^{m} \in \mathbb{R}^{n \times d}$ denotes the final output.

3.1.6 Feature aggregation refinement module

The feature aggregation refinement module plays a crucial role in enhancing the accuracy and robustness of human pose estimation models, which are designed to refine and integrate information across different layers of the neural network architecture. In a recent study, the Deformable ConvNets [32] analysis has demonstrated the remarkable effectiveness of aggregating features from neighboring layers in merging spatial information and semantics. However, the exploration of this feature aggregation method in transformer architectures has not been fully realized. In this research, we proposed a novel transformer network that utilizes feature aggregation modules to improve spatial-temporal semantics. This is attained by stacking numerous cross-layer feature modules in the FAR approach. The overall diagram of the FAR module is shown in Fig. 4(2.2). Each module consists of two adjacent spatial-temporal encoders that aggregate features from two temporal transformer blocks.

The proposed FAR module employs the interaction between queries, keys, and values across two temporal

constriction and proliferation attention blocks within neighboring STEs. This design seamlessly extends feature fusion to the transformer network by leveraging the attention scheme effectively. Additionally, the cross-layer attention operation is denoted by FAR(·). This process is formulated as follows:

$$\begin{aligned} \mathrm{FAR} &= \mathrm{Attn}\big(Z_{t-1}, Z_t\big), \\ &= \mathrm{Attn}\big(Q_t, K', V'\big), \\ &= \mathrm{Softmax}\left(\frac{Q_t K'^{\top}}{\sqrt{d}}\right) V', \\ K' &= \mathrm{Concat}\big(K_t, \mathcal{F}\big(K_{t-1}\big)\big), \\ V' &= \mathrm{Concat}\big(V_t, \mathcal{F}\big(V_{t-1}\big)\big). \end{aligned} \tag{11}$$

The latent features of the previous block are denoted as Z_{t-1}, and the current block represents Z_t, respectively. Z_t can be transformed into query, key, and value representations using distinct weight matrices. Q_t, K_t, and V_t represent the query, key, and value from the second TTE. K_{t-1} and V_{t-1} denote the keys and values from the first TTE and K' and V' are derived through cross-layer attention using the Eq. (11), where, *Concat* represents the concatenation operation, $\mathcal{F}$ denotes the adaptive pooling applied to K_{t-1} and V_{t-1} from the first TTE. Both K_t, V_t, and K_{t-1}, V_{t-1} are processed after the temporal constriction and proliferation attention block. Specifically, the concatenation of the refined key and value using the TCP attention block from the second TTE with the pooled refined key and value from the first TTE allows for the extraction of multi-scale information and more comprehensive details.

3.1.7 Regression and loss

Following feature extraction, we apply a linear layer to obtain the 3D pose keypoints from high-dimensional features, which represented as $Z \in \mathbb{R}^{\mathcal{F} \times \mathcal{N} \times 3}$. Moreover, we employ the loss function as follows:

$$\mathcal{L} = \lambda_w \mathcal{L}_w + \lambda_t \mathcal{L}_t + \lambda_m \mathcal{L}_m. \tag{12}$$

where $\mathcal{L}_w$ denotes the WMPJPE, $\mathcal{L}_t$ denotes the TCLoss, and $\mathcal{L}_m$ indicates the MPJVE. Moreover, λ_w, λ_t, and λ_m, denote the hyperparameters.

4 Experimental results and analysis

4.1 Dataset and evaluation metrics

We estimated our proposed work on two challenging benchmark datasets Human3.6M [33] and MPI-INF-3DHP [34]. The Human3.6M dataset, the largest publicly available resource for 3D human pose estimation, encompasses 3.6 million images obtained from a setup of 4 synchronized cameras operating at a frequency of 50 Hz. Specifically, seven professional subjects are involved in performing 15 daily activities, including main tasks such as "*Waiting*," "*Smoking*," and "*Posing*." Following the established protocol from prior studies [7, 9] the training set consists of five subjects (S1, S5, S6, S7, S8), while the evaluation set involves two subjects (S9 and S11). The assessment standards for this benchmark included MPJPE and P-MPJPE metrics. The MPJPE measures the average Euclidean distance between predicted and ground truth joint positions, indicating overall accuracy. P-MPJPE enhances MPJPE by considering pose similarity through alignment. Lower values for both metrics signify higher accuracy and pose similarity.

MPI-INF-3DHP is a well-known dataset for 3D human pose estimation, which encompasses both indoor and outdoor environmental datasets. It includes videos of 8 actors engaged in 8 different activities, recorded from 14 camera angles, totaling over 1.3 million frames. As in previous studies [8, 9], we used 8 subjects for the training set and 6 subjects for the test set. To compare with prior methodologies, we used a 14-joint skeleton for estimation, including the head, neck, shoulders, elbows, wrists, hips, knees, and ankles. Moreover, we conducted experiments utilizing a 17-joint skeleton to validate the efficacy of our method. Alongside the MPJPE, we employed PCK and AUC as additional metrics to evaluate our model's performance on this dataset.

4.2 Implementation details

In this study, we implemented the proposed MLTFFPN using the PyTorch framework and our experiments were performed on GeForce RTX 4090 GPU. In addition, we employed the 2D keypoints obtained from a 2D pose detector [35] to evaluate the effectiveness of our method. Specifically, we employed varying input and padding lengths for various benchmarks. In the Human3.6M dataset, we used the setup F = 243 and P = 81. Following that, in the MPI-INF-3DHP dataset, we used the setup F = 81 and P = 27 as well as F = 27 and P = 9. Additionally, the model was trained for 120 epochs using the Adam optimizer. Then, we set the learning rate = 1e−4, and the loss functions weights λ_w, λ_t and λ_m are set to 1, 0.5, and 2, respectively.

4.3 Human3.6M dataset evaluation

We compare the proposed approach with different state-of-the-art methods on the Human3.6M dataset as exhibited in Table 1. As illustrated in Table 1, the most favorable results were reported by different approaches, showcasing their performance in terms of MPJPE and P-MPJPE metrics with cascaded pyramid network (CPN) input, respectively. Moreover,

the qualitative evaluation results of our proposed method on testing sequence S9 (Posing) of the Human3.6M dataset are exhibited in Fig. 6. Additionally, we use the 2D pose detectors provided by the widely-used CPN [35] and ground truth data as inputs for training. Our proposed MLTFFPN method achieves superior performance, with scores of 41.9 and 32.5% under the MPJPE and P-MPJPE metrics, respectively. When compared to transformer-based methods such as PoseFormer, HDFormer, MHFormer, MHFormer++ and HSTFormer, our proposed approach shows improvements in MPJPE values by (2.4, 0.7, 1.1, 0.6, 0.8%) under protocol-1 and in P-MPJPE values by (2.1, 0.6, 1.9, 1.7, 1.2%) under protocol-2. Moreover, we obtain better MPJVE outcomes for all actions and the highest average score. This demonstrates the effectiveness of our proposed temporal feature extraction method for human pose estimation in motion. Specifically, we performed the MPJPE error analysis on the Human3.6M dataset, as illustrated in Fig. 5. As shown in Fig. 5, our proposed method demonstrates superior performance in reducing the MPJPE value across all joints in the Human 3.6M dataset when compared to conventional methods such as PoseFormer and MHFormer++. In particular, our approach achieves consistently lower MPJPE values, with notable improvements in critical joints such as the right knee, right wrist, and neck. For example, while PoseFormer and MHFormer++ record higher MPJPEs, our method significantly reduces these errors, showcasing enhanced accuracy in joint position estimation. This reduction in MPJPE underscores the efficacy of our method in delivering more precise and reliable human pose estimations across diverse joint categories, thereby advancing the state-of-the-art in this domain.

4.4 MPI-INF-3DHP dataset evaluation

To evaluate our proposed approach against state-of-the-art methods, we use the MPI-INF-3DHP benchmark dataset and measure performance using the PCK, AUC, and MPJPE metrics. Additionally, the MPI-INF-3DHP dataset is a large collection of 3D human pose data, comprising 1.3 million frames captured in a multi-camera studio, with ground truth obtained through commercial markerless motion capture. The dataset includes various motions performed by eight actors in both indoor and outdoor environments. The evaluation results in Table 2 demonstrate that our proposed approach achieves the best performance with a PCK of 97.7%, an AUC of 76.9%, and an MPJPE of 31.5 mm. When compared to the Transformer-based methods such as PoseFormer, MixSTE, STRFormer, and HSTFormer, our method obtains the best improvements in terms of PCK, AUC, and MPJPE by (10.2%, 20.8% & 45.9 mm), (1.9%, 1.4% & 4.2 mm), (4.0%, 10.1% & 23.2 mm), and (1.5%, 5.7% & 10.2 mm), respectively. In the end, the proposed method achieves superior performance when compared to the conventional methods under the PCK, AUC, and MPJPE metrics.

4.5 Ablation study

4.5.1 Component analysis

To validate the impact of each integrated element in our proposed model, we conducted a comprehensive ablation analysis using the Human3.6M dataset with the MPJPE metric. In this study, we utilize the Seq2Seq approach as

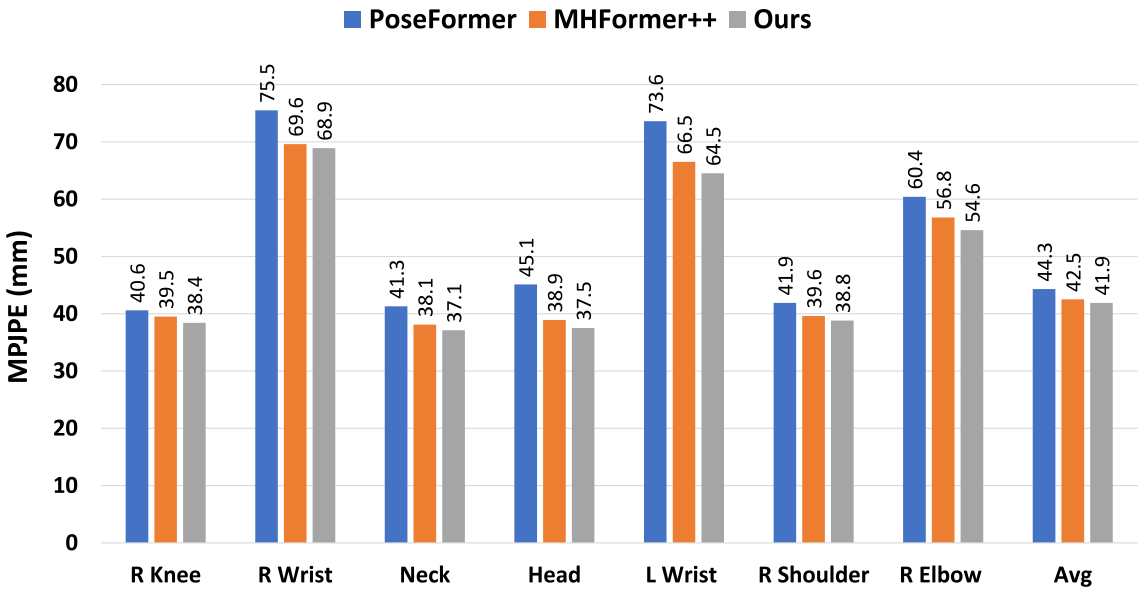

Fig. 5 The average joint error MPJPE comparison with the state-of-the-art results on Human3.6M dataset

Table 1 Quantitative evaluation on the Human3.6M dataset under protocol-1 (MPJPE)

Methods		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
MPJPE - Protocal #1																	
TCN [36] (F = 243)	CVPR'19	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
SRNet [37] (F = 243)	ECCV'20	46.6	47.1	43.9	41.6	45.8	***49.6***	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
PoseFormer [7] (F = 81)	ICCV'21	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
RIE [38] (F = 243)	MM'21	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy [39] (F = 243)	TCSVT'21	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
U-CDGCN [40] (F = 96)	MM'21	**38.0**	43.3	39.1	***39.4***	45.8	53.6	41.4	41.4	55.5	61.9	44.6	41.9	44.5	31.6	29.4	43.4
Ray3D [41] (F = 9)	CVPR'22	44.7	48.7	48.7	48.4	51.0	59.9	46.8	46.9	58.7	61.7	50.2	46.4	51.5	38.6	41.8	49.7
STE [42] (F = 351)	TMM'22	39.9	43.4	40.0	40.9	46.4	50.6	42.1	39.8	55.8	61.6	44.9	43.3	44.9	29.9	30.3	43.6
3D-HPE-PAA [43] (F = 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
P-STMO [44] (F = 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
MLP-JCG [45]	TMM'23	43.7	46.6	46.9	48.9	50.3	60.1	45.7	43.9	56.0	73.7	48.9	48.1	50.9	39.8	41.4	49.7
RS-Net [46] (F= 243)	TIP'23	44.7	48.4	44.8	49.7	49.6	58.2	47.4	44.8	55.2	59.7	49.3	46.4	51.4	38.6	40.6	48.6
HDFormer [29] (F = 96)	arXiv'23	***38.1***	43.1	39.3	39.4	44.3	**49.1**	41.3	40.8	***53.1***	62.1	***43.3***	41.8	43.1	31.0	29.7	42.6
HSTFormer [47] (F = 81)	arXiv'23	39.5	***42.0***	39.9	40.8	44.4	50.9	40.9	41.3	54.7	***58.8***	43.6	**40.7**	43.4	30.1	30.4	42.7
MHFormer [9] (F = 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++ [48] (F = 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***
JoyPose [49] (–)	PR'24	39.3	48.4	47.1	**39.1**	***43.8***	62.7	48.4	40.9	45.2	72.4	49.5	41.6	51.6	34.7	40.7	47.0
PoseFormerV2 [50] (N = 27)	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	45.2
DAF-DG [51]	CVPR'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	44.4
GLA-GCN [52] (N = 243)	ICCV'23	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
MLTFFPN (F = 243, P = 81)		40.0	**41.0**	**37.8**	41.0	**43.7**	51.2	**40.3**	41.5	**53.0**	**58.0**	**43.2**	41.7	**41.5**	**27.1**	**27.9**	**41.9**
P-MPJPE - Protocal #2																	
TCN [36] (F = 243)	CVPR'19	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
PoseFormer [7] (F = 81)	ICCV'21	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	***34.8***	32.4	35.3	24.5	36.0	34.6
RIE [38] (N = 243)	MM'21	32.5	36.2	33.2	35.3	35.6	42.1	32.6	31.9	42.6	47.9	36.6	32.1	34.8	24.2	25.8	35.0
Anatomy [39] (F = 243)	TCSVT'21	32.6	35.1	32.8	35.4	36.3	40.4	32.4	32.3	42.7	49.0	36.8	32.4	36.0	24.9	26.5	35.0
U-CDGCN [40] (F = 96)	MM'21	29.8	34.4	31.9	31.5	35.1	40.0	30.3	30.8	42.6	49.0	35.9	31.8	35.0	25.7	23.6	33.8
STE [42] (F = 351)	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
3D-HPE-PAA [43] (F = 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
P-STMO [44] (F = 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
MLP-JCG [45]	TMM'23	33.6	37.4	37.3	39.6	39.8	47.1	33.7	33.7	45.7	60.4	39.7	37.7	40.0	30.0	33.8	39.3
RS-Net [46] (F = 243)	TIP'23	35.5	38.3	36.1	40.5	39.2	44.8	37.1	34.9	45.0	49.1	40.2	35.4	41.5	31.0	34.3	38.9
Uplift and Upsample [53] (F = 351)	WACV'23	31.6	33.7	***31.8***	33.3	34.7	38.7	32.2	31.2	***41.9***	48.9	35.5	32.6	33.7	23.4	24.0	33.8
HDFormer [29] (F = 96)	arXiv'23	***29.6***	33.8	**31.7**	***31.3***	33.7	**37.7**	***30.6***	31.0	**41.4**	47.6	35.0	30.9	33.7	25.3	***23.6***	***33.1***
HSTFormer [47] (F = 81)	arXiv'23	31.1	***33.6***	33.0	33.2	33.6	38.8	31.9	31.5	43.7	***46.3***	35.7	31.5	***33.1***	24.2	24.5	33.7

Table 1 (continued)

Methods		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
MHFormer [9] (F = 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++ [48] (F = 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
JoyPose [49] (-)	PR'24	**28.8**	40.9	35.8	32.4	***33.3***	52.1	34.5	33.3	34.4	54.9	40.2	**29.5**	43.7	27.5	33.4	37.0
PoseFormerV2 [50] (N = 27)	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	35.6
DAF-DG [51]	CVPR'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	34.6
GLA-GCN [52] (N = 243)	ICCV'23	32.4	35.3	32.6	34.2	35.0	42.1	32.1	31.9	45.5	49.5	36.1	32.4	35.6	23.5	24.7	34.8
MLTFFPN (F = 243, P = 81)		30.8	**33.1**	37.8	**31.2**	**33.2**	***38.2***	**30.2**	***30.6***	42.0	**46.1**	**34.1**	***30.8***	**31.8**	**21.2**	**22.6**	**32.5**
MPJVE																	
PoseFormer [7] (F = 81)	ICCV'21	2.7	2.8	2.0	3.1	2.0	2.4	2.4	2.8	1.8	2.4	2.0	2.1	3.4	2.7	2.4	2.5
MixSTE [8] (F = 243)	CVPR'22	3.2	3.4	2.6	3.6	2.6	3.0	2.9	3.2	2.6	3.3	2.7	2.7	3.8	3.2	2.9	3.1
Anatomy [39] (F = 243)	TCSVT'21	***2.5***	***2.7***	***1.9***	***2.8***	***1.9***	***2.2***	***2.3***	***2.6***	***1.6***	***2.2***	***1.9***	***2.0***	***3.1***	***2.6***	***2.2***	***2.3***
MLTFFPN (F = 243, P = 81)		**2.4**	**2.5**	**1.8**	**2.7**	**1.8**	**2.1**	**2.2**	**2.5**	**1.5**	**2.1**	**1.8**	**1.9**	**3.0**	**2.5**	**2.2**	**2.2**

The CPN is employed as the 2D keypoint detector to produce the input. Notably, the first & second-best outcomes are denoted by bold and bold italic fonts, respectively

our baseline method. As shown in Table 3, the Seq2Seq approach obtained the 42.9*mm* MPJPE result. Further, when integrating the Frame Padding Network with our baseline method *(Seq2Seq+ Frame Padding Network)*, we achieve the 0.3 mm improvement (42.9 mm → 42.6 mm). In addition, we observe a 0.4 mm increment (42.6 mm → 42.2 mm) when incorporating the Multi-Level TCP module into our proposed approach *(Seq2Seq + Frame Padding Network + Multi-Level TCP)*. By including the FAR module in our proposed technique *(Seq2Seq + Frame Padding Network + Multi-Level TCP + FAR)*, we observe a 0.3*mm* improvement (42.2 mm → 41.9 mm). Specifically, our proposed method achieves the most significant improvement, reducing the MPJPE metric by 41.9 mm. In particular, our ablation study has provided valuable insights into the effectiveness of individual components within our proposed model. The experimental outcomes highlighted the importance of different modules, including Frame Padding Network, Multi-Level TCP, and FAR module, playing crucial roles in enhancing performance within specific contexts.

4.5.2 Hyperparameter analysis

We conducted the hyperparameter analysis of our proposed method on the Human3.6M dataset, which is illustrated in Table 4. We set different parameter settings such as the number of levels=2, 3, number of layers=2, 3, frames=81, 243, 300, and padded frames=27, 54, 81, 135. Furthermore, the configurations with 2 levels and 3 layers generally perform worse, with MPJPE values up to 42.6. Using 243 frames is optimal compared to 81 or 300 frames. Padded frames set to 81 consistently perform better than other configurations. Specifically, the proposed method uses the following hyperparameters: 3 levels, 2 layers, 243 frames, and 81 padded frames, achieving an MPJPE of 41.9. This configuration performs better than all other tested configurations, indicating that this combination of hyperparameters is optimal for minimizing the MPJPE. The method stands out by carefully balancing the number of levels, layers, and frame configurations to achieve the lowest error, which demonstrates superior performance.

4.5.3 Parameter analysis

We have analyzed the parameters, FLOPs, FPS, and MPJPE for various models under protocol-1 on the Human3.6M dataset, as illustrated in Table 5. Specifically, the proposed method demonstrates a balanced performance in terms of parameters, FLOPs, FPS, and MPJPE compared to conventional methods like PoseFormer, Anatomy, and P-STMO. In addition, our proposed model MLTFFPN has 33.8M parameters, which is higher than PoseFormer (9.6M) and P-STMO (6.7M), but significantly lower than Anatomy (58.1M). It

Table 2 Quantitative evaluation on the MPI-INF-3DHP dataset under three evaluation metrics

Methods		PCK ↑	AUC ↑	MPJPE ↓
TCN [36] (N = 81)	CVPR'19	86.0	51.9	84.0
Anatomy [39] (N = 81)	TCSVT'21	87.9	54.0	78.8
PoseFormer [7] (N = 9, H = 3, P-Agg)	ICCV'21	88.6	56.4	77.1
U-CDGCN [40] (N = 96)	MM'21	97.9	69.5	42.5
MHFormer [9] (N = 27)	CVPR'22	93.8	63.3	58.0
P-STMO [44] (N = 81)	ECCV'22	97.9	75.8	32.2
MixSTE [8] (N = 243)	CVPR'22	96.9	75.8	35.4
MHFormer++ [48] (N = 9)	PR'23	94.8	65.8	54.0
Uplift & Upsample [53] (N = 81)	WACV'23	97.9	75.8	32.2
EMHIFormer [54] (N = 9)	JVCIR'23	97.1	74.9	33.8
STRFormer [55] (N = 27)	IMAVIS'23	94.8	67.1	54.4
HDFormer [29] (N = 32)	arXiv'23	96.8	64.0	51.5
HSTFormer [47] (N = 81)	arXiv'23	97.3	71.5	41.4
JoyPose [49]	PR'24	94.1	–	–
DAF-DG [51]	CVPR'24	92.9	60.7	63.1
GLA-GCN [52] (N = 27)	ICCV'23	***98.1***	***76.5***	***31.3***
MLTFFPN (F = 81, p = 27)		**98.8**	**77.2**	**31.2**

The best and second-best results are highlighted in bold and bold italic fonts, respectively

Table 3 Ablation analysis on different proposed elements

Seq2Seq	Frame padding network	Multi-level TCP	FAR	MPJPE
✓	✗	✗	✗	42.9
✓	✓	✗	✗	42.6
✓	✓	✓	✗	***42.2***
✓	✓	✓	✓	**41.9**

The evaluation is conducted on the Human3.6M dataset with the MPJPE metric. The results marked in bold and bold italic denote the first & second-best outcomes, respectively

achieves the lowest FLOPs at 645 M, outperforming PoseFormer (851 M) and P-STMO (1737 M), and is slightly more efficient than Anatomy (656 M). In terms of MPJPE, MLTFFPN scores the best at 41.9, surpassing PoseFormer (44.3), Anatomy (44.1), and P-STMO (42.8). This indicates that MLTFFPN provides a superior balance of efficiency and accuracy, making it a highly effective method among the evaluated approaches.

4.5.4 Qualitative comparison of wild videos

We conduct a qualitative analysis of our proposed method, comparing it against competitive approaches such as MHFormer, GraphMLP, MixSTE, and PoseFormer. This evaluation is conducted on challenging in-the-wild videos to demonstrate the effectiveness of our approach in real-world scenarios. Moreover, the comparison results are exhibited in Fig. 7. Following that, the deviated 3D pose prediction is highlighted within a dotted black circle. Notably, the green circle indicates locations where our approach yields superior outcomes. Moreover, the 2D detector CPN [35] is used to extract 2D poses, which are then fed into the models to ensure a fair comparison. Despite the complex actions and rapid motions, the proposed approach excels in producing realistic and plausible 3D predictions that surpass those of previous approaches. This demonstrates the robustness of our method in dealing with partial occlusions and its ability to tolerate depth ambiguity.

4.5.5 Qualitative comparison of real-time videos

To demonstrate the effectiveness of our proposed approach, we conducted real-time experiments using four video sequences. Each sequence showcases intricate and challenging poses, providing a comprehensive evaluation of our method's performance across a range of difficult scenarios. The 2D and 3D pose predictions are illustrated in Fig. 8. Specifically, the green circle marks signify the superior outcomes achieved by our proposed method, while the black circle marks indicate the wrongly predicted positions. Furthermore, the application of our proposed approach in real-world scenarios has exhibited improved performance.

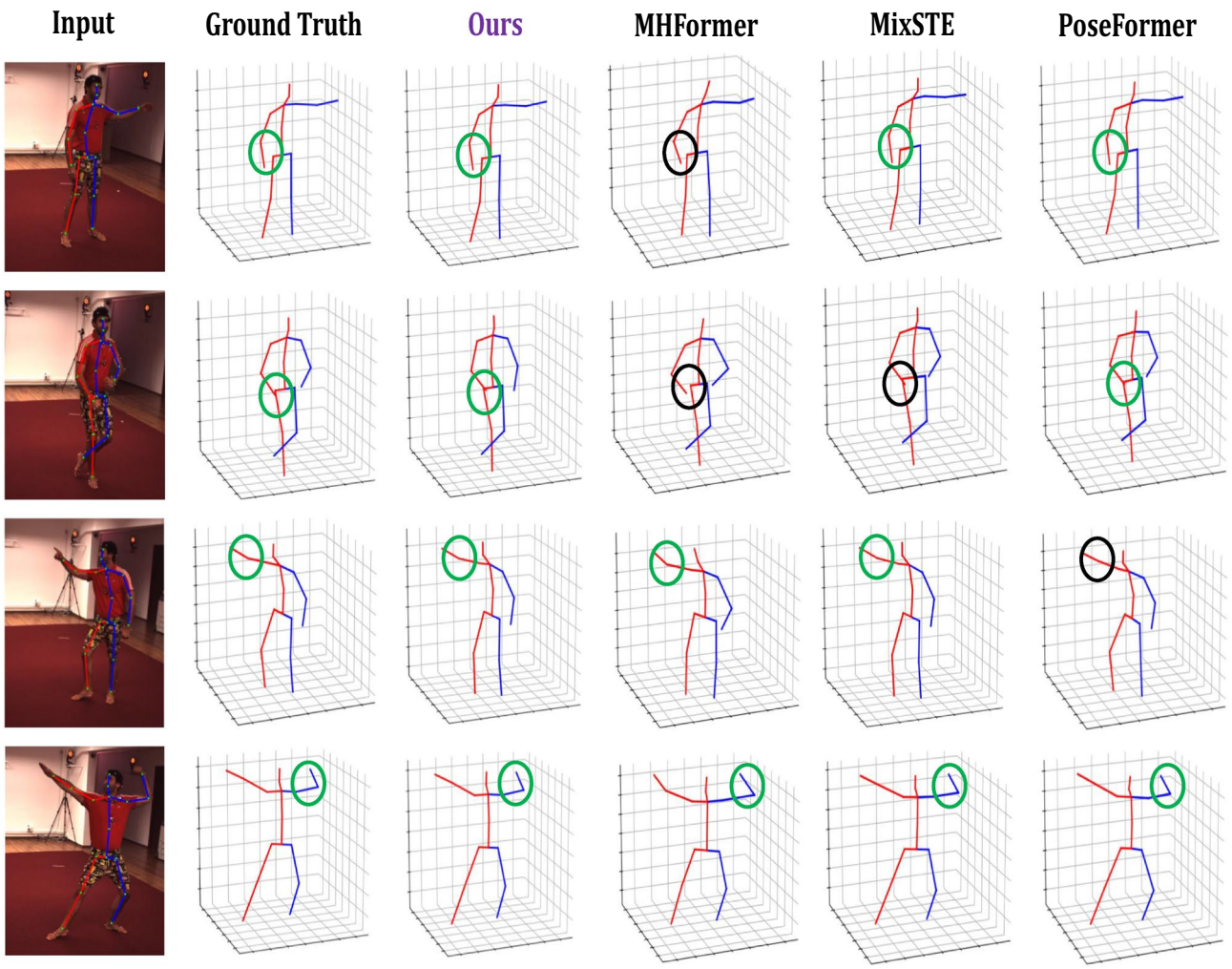

Fig. 6 Qualitative evaluation is conducted on the Human3.6M dataset to compare the proposed approach with cutting-edge methods such as MHFormer, MixSTE, and PoseFormer. The correctly predicted pose is indicated by a green circle, while the incorrectly predicted pose is marked by a black circle

Table 4 Hyperparameter settings in our proposed method on the Human3.6M dataset under the MPJPE metric

Number of levels	Layers	Frames (F)	Padded frames ($\mathcal{P}$)	MPJPE ↓
2	2	243	81	42.5
2	3	243	81	42.4
3	3	243	81	42.3
3	3	243	27	42.1
3	2	81	81	42.6
3	2	300	81	42.2
3	2	243	54	42.0
3	2	243	135	42.3
3	2	243	81	41.9

Table 5 Analysis of parameters, FLOPs, FPS, and MPJPE for various models under protocol-1 on the Human3.6M dataset

Methods	Parameter ↓	FLOPs ↓	FPS	MPJPE ↓
PoseFormer [7] (N = 81)	9.6 M	851 M	1952	44.3
Anatomy [39] (N = 243)	58.1 M	656 M	264	44.1
P-STMO [44] (N = 243)	6.7 M	1737 M	3040	42.8
MLTFFPN (Ours) (N = 243)	33.8 M	645 M	3568	41.9

5 Conclusion

In this study, a Multi-Level Transformers with a Feature Frame Padding Network (MLTFFPN) approach has been proposed. Our innovative feature frame-padding network

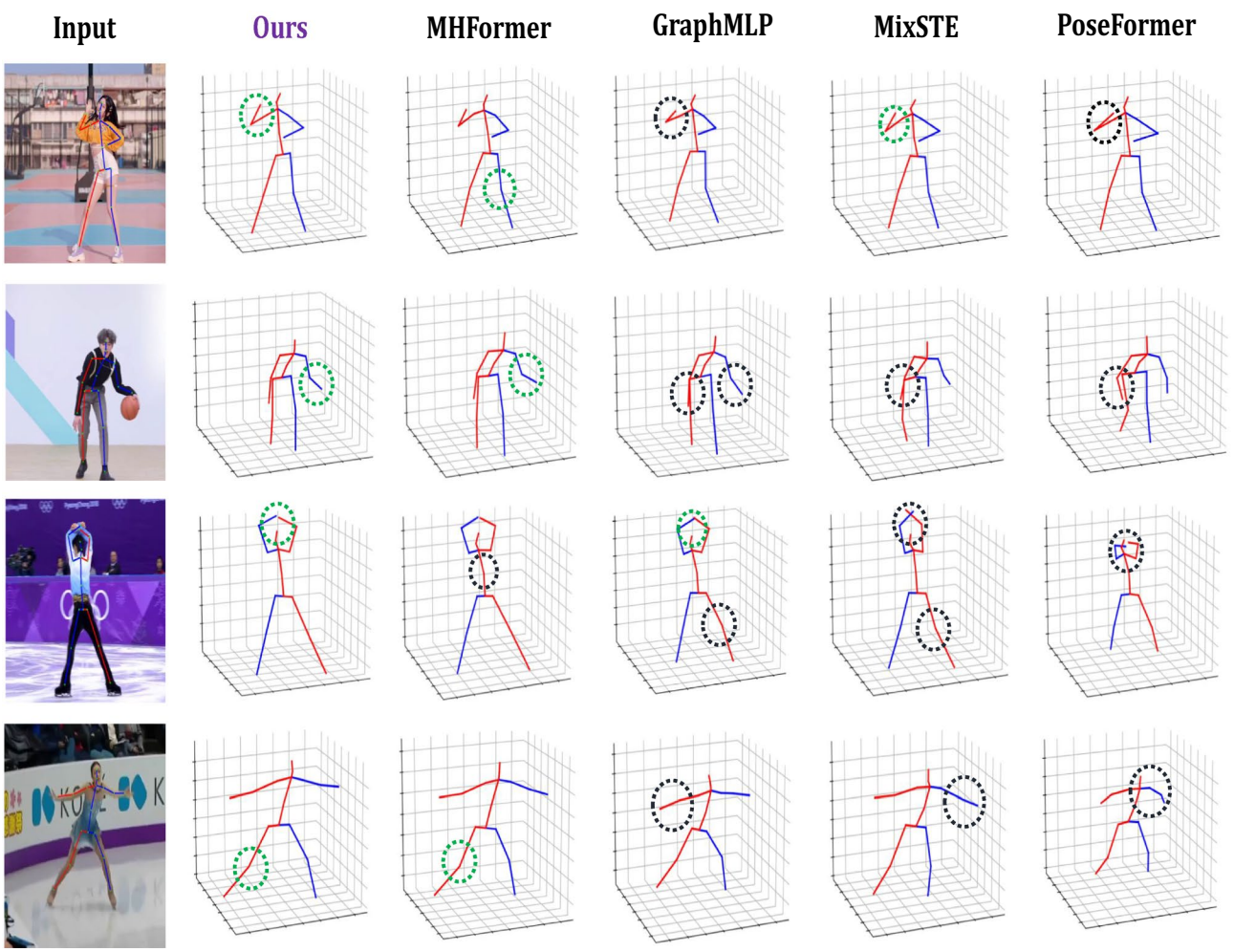

Fig. 7 Qualitative evaluation is conducted on in-the-wild videos to compare the suggested approach with cutting-edge methods such as MHFormer, GraphMLP, MixSTE, and PoseFormer. The correctly predicted pose is marked with a green circle, whereas the incorrectly predicted pose is denoted by a black circle

effectively captures longer temporal dependencies and compensates for the absence of edge frame information. This enhancement leads to a better understanding of the sequential nature of human motion and significantly improves pose estimation accuracy. Additionally, the multi-level transformer architecture has been employed, which effectively extracts temporal information from 3D human poses, strengthening the short-term correlation among keypoints of the human pose skeleton. Specifically, a key component of our approach is the refined temporal constriction and proliferation transformer, which incorporates spatio-temporal encoders and a TCP structure. This design effectively addresses the depth ambiguity problem by revealing multi-scale attention information. Moreover, the integration of the feature aggregation refinement module into the TCP block in a cross-layer manner enhances semantic representation through continuous interaction among queries, keys, and values. Our extensive experiments on the Human3.6M and MPI-INF-3DHP benchmark datasets demonstrate the effectiveness of our proposed approach. The results indicate significant improvements in the accuracy of 3D human pose estimation, showcasing the potential of MLTFFPN in advancing the state-of-the-art in this domain.

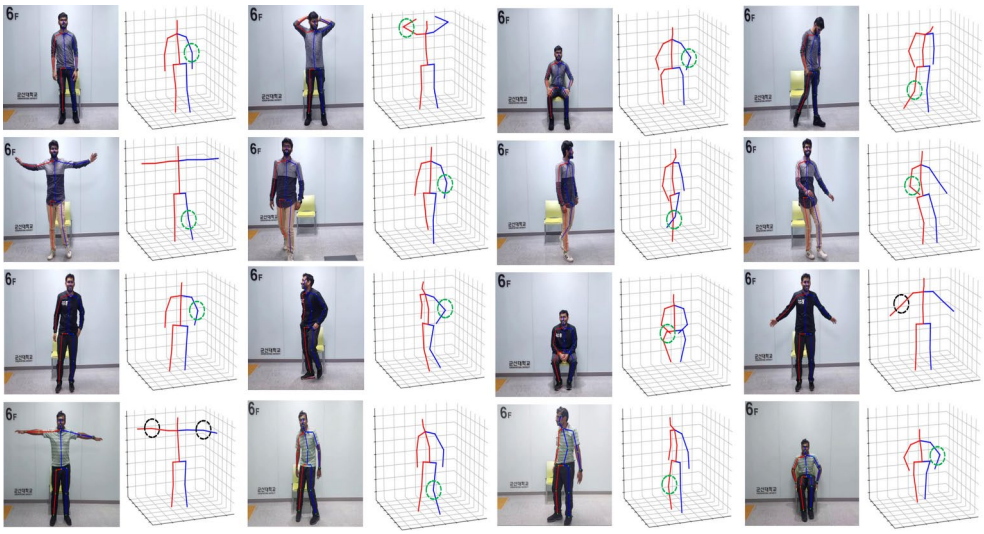

Fig. 8 We conducted real-time experiments to evaluate our proposed method on various challenging sequences. The 2D and 3D pose estimations are illustrated in the first and second columns, respectively

Acknowledgements This work was supported in part by the Basic Science Research Program under Grant NRF -2016R1A6A1A03013567 and Grant NRF-2021R1A2B5B01001484 and by the framework of the International Cooperation Program under Grant NRF-2022K2A9A2A06045121 through the National Research Foundation of Korea (NRF) funded by the Ministry of Education.

Author contributions The author confirms responsibility for data collection, analysis and interpretation of results, and manuscript preparation. S.A. wrote the main manuscript text. All authors reviewed the manuscript.

Data availability Data will be made available on request.

Declarations

Conflict of interest The author declares that they have no conflict of interest.

References

1. Moorthy, S., Joo, Y.H.: Learning dynamic spatial-temporal regularized correlation filter tracking with response deviation suppression via multi-feature fusion. Neural Netw. **167**, 360–379 (2023)
2. Sachin Sakthi, K.S., Jeong, J.H., Joo, Y.H.: A multi-level hybrid Siamese network using box adaptive and classification approach for robust tracking. Multimed. Tools Appl. (2024). https://doi.org/10.1007/s11042-024-19465-5
3. Elayaperumal, D., Joo, Y.H.: Learning spatial variance-key surrounding-aware tracking via multi-expert deep feature fusion. Inf. Sci. **629**, 502–519 (2023)
4. Moorthy, S., Joo, Y.H.: Adaptive spatial-temporal surrounding-aware correlation filter tracking via ensemble learning. Pattern Recogn. **139**, 109457 (2023)
5. Kuppusami Sakthivel, S.S., Moorthy, S., Arthanari, S., Jeong, J.H., Joo, Y.H.: Learning a context-aware environmental residual correlation filter via deep convolution features for visual object tracking. Mathematics **12**(14), 2279 (2024)
6. Elayaperumal, D., Joo, Y.H.: Robust visual object tracking using context-based spatial variation via multi-feature fusion. Inf. Sci. **577**, 467–482 (2021)
7. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11656–11665 (2021)
8. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13232–13242 (2022)
9. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13156 (2022)
10. Li, W., Liu, H., Guo, T., Ding, R., Tang, H.: Graphmlp: A graph mlp-like architecture for 3d human pose estimation. arXiv preprint arXiv:2206.06420 (2022)
11. Wu, Y., Kong, D., Wang, S., Li, J., Yin, B.: Hpgcn: Hierarchical poselet-guided graph convolutional network for 3d pose estimation. Neurocomputing **487**, 243–256 (2022)
12. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.-h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8818–8829 (2023)
13. Hassan, M.T., Hamza, A.B.: Regular splitting graph network for 3d human pose estimation. IEEE Trans. Image Process. **32**, 4212–4222 (2023)
14. Mehraban, S., Adeli, V., Taati, B.: Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6920–6930 (2024)

15. Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., Tao, D.: Transvod: end-to-end video object detection with spatial-temporal transformers. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7853–7869 (2022)
16. Tian, X., Jin, Y., Tang, X.: Local-global transformer neural network for temporal action segmentation. Multimed. Syst. **29**(2), 615–626 (2023)
17. Tian, X., Jin, Y., Tang, X.: Tsrn: two-stage refinement network for temporal action segmentation. Pattern Anal. Appl. **26**(3), 1375–1393 (2023)
18. Kim, D., Xie, J., Wang, H., Qiao, S., Yu, Q., Kim, H.-S., Adam, H., Kweon, I.S., Chen, L.-C.: Tubeformer-deeplab: Video mask transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13914–13924 (2022)
19. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Video k-net: A simple, strong, and unified baseline for video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18847–18857 (2022)
20. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131 (2018)
21. Wang, Z., Nie, X., Qu, X., Chen, Y., Liu, S.: Distribution-aware single-stage models for multi-person 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13096–13105 (2022)
22. Sun, S., Liu, D., Dong, J., Qu, X., Gao, J., Yang, X., Wang, X., Wang, M.: Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 2973–2984 (2023)
23. Huang, W., Liu, D., Hu, W.: Dense object grounding in 3d scenes. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 5017–5026 (2023)
24. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2019)
25. Zhou, J., Zhang, T., Hayder, Z., Petersson, L., Harandi, M.: Diff3dhpe: A diffusion model for 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2092–2102 (2023)
26. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8844–8854 (2022)
27. He, L., Zhou, Q., Li, X., Niu, L., Cheng, G., Li, X., Liu, W., Tong, Y., Ma, L., Zhang, L.: End-to-end video object detection with spatial-temporal transformers. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1507–1516 (2021)
28. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimed. **25**, 1282–1293 (2022)
29. Chen, H., He, J.-Y., Xiang, W., Cheng, Z.-Q., Liu, W., Liu, H., Luo, B., Geng, Y., Xie, X.: Hdformer: High-order directed transformer for 3d human pose estimation. arXiv preprint arXiv:2302.01825 (2023)
30. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14761–14771 (2023)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems 30 (2017)
32. Yu, B., Jiao, L., Liu, X., Li, L., Liu, F., Yang, S., Tang, X.: Entire deformable convnets for semantic segmentation. Knowl. Based Syst. **250**, 108871 (2022)
33. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2013)
34. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV), pp. 506–516 (2017). IEEE
35. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
36. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2019)
37. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pp. 507–523 (2020). Springer
38. Shan, W., Lu, H., Wang, S., Zhang, X., Gao, W.: Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3446–3454 (2021)
39. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Trans Circuits Syst. Video Technol. **32**(1), 198–209 (2021)
40. Hu, W., Zhang, C., Zhan, F., Zhang, L., Wong, T.-T.: Conditional directed graph convolution for 3d human pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 602–611 (2021)
41. Zhan, Y., Li, F., Weng, R., Choi, W.: Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13116–13125 (2022)
42. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimed. **25**, 1282–1293 (2022)
43. Xue, Y., Chen, J., Gu, X., Ma, H., Ma, H.: Boosting monocular 3d human pose estimation with part aware attention. IEEE Trans. Image Process. **31**, 4278–4291 (2022)
44. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision, pp. 461–478 (2022). Springer
45. Tang, Z., Li, J., Hao, Y., Hong, R.: Mlp-jcg: multi-layer perceptron with joint-coordinate gating for efficient 3d human pose estimation. IEEE Trans. Multimed. **25**, 8712–8724 (2023). https://doi.org/10.1109/TMM.2023.3240455
46. Hassan, M.T., Ben Hamza, A.: Regular splitting graph network for 3d human pose estimation. IEEE Trans. Image Process. **32**, 4212–4222 (2023). https://doi.org/10.1109/TIP.2023.3275914
47. Qian, X., Tang, Y., Zhang, N., Han, M., Xiao, J., Huang, M.-C., Lin, R.-S.: Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. arXiv preprint arXiv:2301.07322 (2023)

48. Li, W., Liu, H., Tang, H., Wang, P.: Multi-hypothesis representation learning for transformer-based 3d human pose estimation. Pattern Recogn. **141**, 109631 (2023)
49. Du, S., Yuan, Z., Lai, P., Ikenaga, T.: Joypose: Jointly learning evolutionary data augmentation and anatomy-aware global-local representation for 3d human pose estimation. Pattern Recogn. **147**, 110116 (2024)
50. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8877–8886 (2023)
51. Peng, Q., Zheng, C., Chen, C.: A dual-augmentor framework for domain generalization in 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2240–2249 (2024)
52. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.-h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8818–8829 (2023)
53. Einfalt, M., Ludwig, K., Lienhart, R.: Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2903–2913 (2023)
54. Xiang, X., Zhang, K., Qiao, Y., El Saddik, A.: Emhiformer: An enhanced multi-hypothesis interaction transformer for 3d human pose estimation in video. J. Vis. Commun. Image Represent. **95**, 103890 (2023)
55. Liu, X., Tang, H.: Strformer: Spatial-temporal-retemporal transformer for 3d human pose estimation. Image Vis. Comput. **140**, 104863 (2023)

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com