Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20179-x

Exploiting multi-transformer encoder with multiple-hypothesis aggregation via diffusion model for 3D human pose estimation

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

Received: 11 June 2024 / Revised: 31 July 2024 / Accepted: 28 August 2024

Abstract
The transformer architecture has consistently achieved cutting-edge performance in the task of 2D to 3D lifting human pose estimation. Despite advances in transformer-based methods they still suffer from issues related to sequential data processing, addressing depth ambiguity, and effective handling of sensitive noisy data. As a result, transformer encoders encounter difficulties in precisely estimating human positions. To solve this problem, a novel multi-transformer encoder with a multiple-hypothesis aggregation (MHAFormer) module is proposed in this study. To do this, a diffusion module is first introduced that generates multiple 3D pose hypotheses and gradually distributes Gaussian noise to ground truth 3D poses. Subsequently, the denoiser is employed within the diffusion module to restore the feasible 3D poses by leveraging the information from the 2D keypoints. Moreover, we propose the multiple-hypothesis aggregation with a join-level reprojection (MHAJR) approach that redesigns the 3D hypotheses into the 2D position and selects the optimal hypothesis by considering reprojection errors. In particular, the multiple-hypothesis aggregation approach tackles depth ambiguity and sequential data processing by considering various possible poses and combining their strengths for a more accurate final estimation. Next, we present the improved spatial-temporal transformers encoder that can help to improve the accuracy and reduce the ambiguity of 3D pose estimation by explicitly modeling the spatial and temporal relationships between different body joints. Specifically, the temporal-transformer encoder introduces the temporal constriction & proliferation (TCP) attention mechanism and the feature aggregation refinement module (FAR) into the refined temporal constriction & proliferation (RTCP) transformer, which enhances intra-block temporal modeling and further refines inter-block feature interaction. Finally, the superiority of the proposed approach is demonstrated through comparison with existing methods using the Human3.6M and MPI-INF-3DHP benchmark datasets.

✉ 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: 10 September 2024

Keywords Multiple-hypothesis aggregation · Temporal constriction and proliferation transformer · Spatio-temporal transformer · Diffusion model · 3D Human pose estimation

1 Introduction

The field of 3D human pose estimation is a computer vision task that aims to focus on inferring the 3D positions of human body joints from images or videos [1–5]. It has achieved considerable attention in recent decades due to its pivotal role in diverse applications, including virtual reality, action recognition, and human-robot interaction. This task can be categorized into two primary approaches: the end-to-end approach and the lifting-based approach. The end-to-end approach performs a direct regression of 3D coordinates from RGB images. 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. Despite leveraging the robust efficiency of 2D pose detectors, the 2D-to-3D lifting approach remains a complex challenge due to the intrinsic lack of depth ambiguity, occlusion, deformation, and sensitive noisy data in 2D representations. To tackle these challenges, the field of human pose estimation has witnessed the development of techniques, with convolutional neural networks and transformer-based approaches playing a pivotal role.

Recently, convolutional neural networks (CNN) have gained significant attention among researchers due to their ability to automatically learn features, and capture spatial relationships, which makes them well-suited for 3D human pose estimation. In recent times, various CNN-based approaches [6–10] have emerged to address the challenges associated with occlusion and depth ambiguity in pose estimation. In this regard, the authors in [6] have presented a graph convolutional networks (GCN) based spatio-temporal approach, which effectively tackles the problem of the 3D human body and 3D hand pose estimation from a short sequence of 2D joint detections. In addition, the authors in [7] have proposed an attentional mechanism, which enhances accuracy and overcomes the occlusion problem efficiently by emphasizing informative regions in the input data. In reference [8], the authors have provided graph attention with spatio-temporal CNN architecture, which utilized the graph attention mechanisms to model intricate spatio-temporal relationships among body joints. This approach plays a vital role in improving the understanding of complex dependencies to achieve accurate pose estimation. Additionally, the authors in [9] have embedded the hierarchical poselet-guided graph convolutional network (HP-GCN), which utilizes graph convolutional networks to capture spatial and structural connections between body joints, which leads to enhanced pose estimation, especially when dealing with occluded scenarios. Moreover, the authors in [10] have introduced a higher-order regular splitting graph network (RS-Net) for 2D-to-3D human pose estimation, which aims to capture long-range dependencies between body joints using a multi-hop neighborhoods approach. Despite CNN-based methods improving accuracy, they still face challenges in effectively handling self-occlusion and long-range dependencies. In response to these concerns, transformer-based methods excel by proficiently modeling the long-range dependencies among body joints.

On the other hand, transformer-based methods outperform CNN approaches in 3D human pose estimation due to their ability to capture long-range dependencies and complex relationships among body joints. They demonstrate proficiency in modeling global dependencies, efficiently managing self-occlusion, and providing a strong framework for temporal modeling. Many researchers have recently focused on transformer-based approaches [11–15]. In this context, the authors of [11] have proposed the PoseFormer technique, which develops an efficient spatio-temporal transformer technique to decode complex local relationships between joints in each frame and capture global dependencies across the entire sequence. Nevertheless, the PoseFormer approach encounters challenges in efficiently managing intricate temporal dependencies across long sequences. To tackle this, the authors in [12] have introduced a novel MixSTE (Mix Spatio-Temporal Encoder) module, which significantly improves the computational efficiency and effectiveness of capturing temporal dependencies in 3D human pose estimation. Following this, the MixSTE method has been effectively designed with an innovative seq2seq technique to grasp the overall consistency among sequences, which improves accuracy in the reconstruction of poses. However, the MixSTE's inability to model long-range dependencies and capture spatial context leads to inaccurate pose estimation, especially for complex poses and noisy environments. To address these challenges, MHFormer [13] leverages multi-head attention and a hierarchical structure. The multi-head attention allows MHFormer to attend to different parts of the input sequence simultaneously, which helps it model long-range dependencies. Subsequently, the hierarchical structure allows MHFormer to learn different representations of the spatial relationships between tokens at different levels of granularity, which helps it capture spatial context. Consequently, the MHFormer can produce more realistic 3D poses by modeling long-range dependencies and capturing spatial context, especially for challenging postures and obstructed limbs. However, the MHFormer encounters a notable challenge by conceptualizing the human body as a unified structure without explicitly describing the connections between different body parts. This approach limits its capability to represent relationships specific to individual body parts. To leverage this problem, the authors of [14] have introduced the part-aware attention module, which employs a part-aware embedding to encode the association of each joint with its corresponding body part and utilizes a part-aware attention mechanism to learn distinct representations of spatial relationships between joints located in different body parts. Part-aware attention primarily focuses on capturing dependencies between joints belonging to the same part, making it challenging to effectively capture long-range dependencies between joints located in different body parts. To address this limitation, a strided transformer architecture [15] has been proposed, which enhances the receptive field by allowing joints from various body parts to attend to each other. This is achieved through the use of multiple strided attention layers, each having a different stride, enabling the model to capture long-range dependencies between joints located in different parts of the body. In this study, we leverage a transformer-based architecture as our baseline model for 3D human pose estimation due to the strong capability of modeling sequential data.

The 3D human pose estimation is typically divided into 2 categories: deterministic and probabilistic methods. The deterministic methods [11, 12] aim to provide a specific output, which is used to directly predict the keypoint locations in the 3D space. Moreover, the deterministic methods are known for their speed and efficiency, making them suitable for real-time applications. However, they may struggle with depth ambiguities or uncertainties in the image, leading to less accurate results in complex scenarios. In contrast, the probabilistic methods [13, 16] denote the 2D to 3D lifting approaches, which generate multiple potential outcomes for each image. This approach facilitates handling uncertainties and ambiguities inherent in the lifting technique. Furthermore, existing probabilistic approaches use

generative adversarial networks [17] to predict the multiple 3D hypotheses. Even though these methods yield multiple 3D hypotheses, real-time applications still require individual 3D poses. Despite significant advancements, these approaches still face challenges in accurately capturing intricate spatial relationships and dependencies during the pose estimation process, resulting in suboptimal outcomes in specific scenarios. To address these issues, we proposed the diffusion model and multiple hypothesis aggregation with a join-level reproject approach. This study focuses on the probabilistic method, offering customizable parameters for the number of hypotheses and iterations. This flexibility enables the generation of the final feasible 3D pose.

Specifically, the diffusion model with a multiple hypothesis aggregation approach struggles to fully exploit temporal information, which lacks short-range and long-range temporal dependencies in the motion sequences. To solve this problem, we integrated the temporal constriction and proliferation (TCP) transformer with the feature aggregation refinement (FAR) module, which effectively handles the short-range and long-range temporal dependencies problems. Additionally, the TCP and FAR modules reduce computational complexity by streamlining temporal feature processing. The TCP mechanism prioritizes crucial temporal aspects, reducing unnecessary computations by focusing on relevant interactions. The FAR module enhances feature aggregation by emphasizing significant features and minimizing redundancies. These modules enhance efficiency and maintain accuracy in temporal modeling while significantly reducing the computational load. To the best of the author's knowledge, this is the first approach to utilize the TCP and FAR modules in the diffusion method via the temporal transformer encoder (TTE) mechanism. Motivated by the above discussions, this study aims to present a novel multi-transformer encoder with multiple-hypothesis aggregation (MHAFormer) via diffusion model for 3D human pose estimation. The main contribution of the proposed approach is described as follows:

1. A diffusion model is presented to produce multiple 3D hypotheses, which gradually distributes Gaussian noise to ground truth 3D poses. Following this, the denoiser is employed within the diffusion module to effectively restore the feasible 3D poses by leveraging the information from the 2D keypoints.
2. A multiple-hypothesis aggregation with a joint-level reprojection (MHAJR) approach is proposed, which redesigns the 3D hypotheses into the 2D position and selects the optimal hypothesis by considering reprojection errors. Specifically, the selected optimal hypothesis is integrated into the final 3D position to enhance the accuracy of the prediction.
3. The proposed RTCP transformer integrates the spatial-temporal encoders with a temporal constriction and proliferation (TCP) structure to enhance intra-block temporal modeling and expose multi-scale attention information. This approach effectively captures both short-range and long-range temporal dependencies in motion sequences, leading to more accurate 3D pose predictions.
4. The feature aggregation refinement (FAR) module is introduced in the RTCP transformer to optimize feature fusion. This is achieved by employing two TCP attention blocks that facilitate interaction among queries, keys, and values. Specifically, the TCP attention mechanism plays a crucial role in enhancing feature fusion within the transformer network.
5. Extensive experiments are conducted on the benchmark datasets Human3.6M and MPI-INF-3DHP, demonstrating the superiority of our proposed method over other state-of-the-art approaches.

2 Related works

In this division, we present a brief overview of human pose estimation. First, we analyze the 3D human pose estimation in Section 2.1. Next, we describe the graph convolutional networks-based approaches in Section 2.2. Finally, we explore transformer-based architectures in Section 2.3.

2.1 3D human pose estimation

In recent years, 3D human pose estimation has emerged as a pivotal role in computer vision, which involves analyzing images from either a single or multiple perspectives. In the context of 3D human pose estimation, the main objective is to accurately determine the positions of human body joints in images or videos. To achieve this objective, numerous techniques such as one-stage detectors and two-stage detectors have been developed to enable the classification of human poses. The one-stage detectors perform end-to-end learning by directly predicting 3D body joint coordinates from raw image data [18] and the two-stage detectors involve two intermediate steps. During the initial stage, we calculate 2D keypoints based on the input image. In the subsequent stage, we leverage the relationships between 2D and 3D human pose to transform the initially estimated 2D keypoints into corresponding 3D positions. Due to the progress in dependable 2D detection, recent 2D-to-3D lifting methods [19–22] have demonstrated superior performance compared to end-to-end approaches. For example, the authors in [19] introduced the concept of pose grammar to encode human body configuration for 3D pose estimation, leading to more accurate and robust pose estimation. Further, the authors proposed a novel multi-view and temporal fusing transformer [20] (MTF-Transformer), which effectively integrates multiple viewpoints and temporal information through a novel attention mechanism and a temporal fusion transformer to estimate accurate 3D poses under diverse capture conditions. Additionally, the authors utilized the CV-UGCNs [21] approach to effectively capture the spatial arrangements and cross-view relationships within 3D human poses, which enabled the network to learn and enforce structural consistency across multiple viewpoints, leading to more accurate and realistic pose estimation. Moreover, the authors designed TP-LSTM [22], a deep learning architecture that effectively utilizes temporal information, which captures the dynamic nature of human motion by incorporating long-range temporal dependencies, enabling them to generate more accurate and consistent pose estimates over time. Taking inspiration from the aforementioned discussion, we exploit a two-stage pipeline for 3D pose estimation because it is a widely used and effective approach that consistently outperforms single-stage methods.

2.2 Graph convolutional networks-based approaches

Graph convolutional networks (GCNs) have become a dominant paradigm in the field of 3D human pose estimation. This innovative method employs the inherent skeletal structure of the human body, which is represented as a graph to enhance the accuracy and robustness of pose predictions. In recent times, several methods [23–25] have employed the graph convolutional neural network architecture to demonstrate their efficiency in 3D pose estimation. The authors in [23] have presented the modulated-graph convolutional network (MGCN), which employs the affinity modulation technique to enhance the model's ability to capture complex relationships between body joints. In addition, the relation-balanced graph convolutional network (RBGC-Net) [24] method has introduced the local-global feature fusion technique,

which extracts local relationships between neighboring joints while balancing them with global relationships, improving interactions between joints and leading to more accurate estimations. Specifically, the authors in [25] have proposed the Graph-MLP approach, which integrates graph convolutional networks within the multi-layer perceptron (MLP) framework. This design aims to capture both local spatial interactions between joints and global connectivity information encoded in the human skeleton. The incorporation of GCNs enriches the feature representation, leading to more accurate pose estimation. Even though GCN-based approaches have demonstrated superior performance, it is difficult to capture long-range dependencies across the entire body. This limitation may lead to challenges in understanding complex body postures and the global interactions within the body. To address these problems, the transformer-based approach leverages the self-attention technique to obtain long-range relationships and a global context more effectively.

2.3 Transformer-based architectures

The computer vision community has increasingly focused on the transformer model [26–29], which stands out for its robust global self-attention mechanism. The self-attention technique has ignited impactful research endeavors within the expansive domain of computer vision [11–13, 30–32]. Initially, the authors in [30] have introduced the ViT method, which employs a pure transformer architecture directly on sequences of image patches, showcasing its ability to achieve state-of-the-art performance in image classification tasks. In addition, the authors in [32] have proposed the P-STMO approach that utilizes the MLP block as the spatial feature extractor, which proves more effective than transformer or fully connected layers in terms of capturing spatial relationships between joints. Subsequently, the temporal downsampling strategy reduces data redundancy and increases the temporal receptive field, allowing the model to learn long-range temporal dependencies with less computational cost. In [31], the authors have presented the PoseFormerV2 method, which effectively integrates the Time & Frequency domain features using the hybrid-attention mechanism to capture detailed information within each frame. Specifically, the authors in [13] have proposed the MHFormer that can effectively learn spatio-temporal representations of multiple pose hypotheses in an end-to-end manner. This approach utilizes the cross-hypothesis communication strategy, which combines features from multiple hypotheses to generate an accurate 3D pose. Inspired by the above analysis, we presented the multiple-hypothesis aggregation (MHA) method to accurately predict the 3D position.

3 Proposed method and implementation

3.1 Multiple hypothesis aggregation with multi-transformer encoder module

Initially, a diffusion model designed to generate multiple 3D hypotheses from a single 2D observation is presented. Following this, the refined temporal constriction and proliferation transformer is proposed, which combines the spatial-transformer encoder (STE) and temporal-transformer encoder (TTE). In the context of the TTE method, we specifically introduced two modules: the temporal constriction & proliferation, and feature aggregation refinement module. The TCP module employs key and value refinement to enhance the transformer's capabilities and expose multi-scale attention information. Subsequently, the FAR module is specifically designed to integrate spatial and semantic information across neigh-

boring temporal encoders using a cross-attention mechanism, which enhances the model's capacity and learns inter-block temporal dynamics (Fig. 2). Finally, we present the MHAJR approach in the diffusion model, which aims to aggregate multiple hypotheses to produce an accurate 3D position. The overall architecture of our proposed method is illustrated in Fig. 1. Generally, each frame of the input video is treated as a token in the transformer and represented as $\mathcal{X} \in \mathbb{R}^{B\times J\times F\times 2}$, where B denotes the batch size, J denotes the number of joints, F signifies the frames and 2 represents the 2D coordinates of the input sequence, respectively. In addition, the noisy 3D pose is defined as $\mathcal{Y}_t \in \mathbb{R}^{B\times J\times F\times 3}$. Specifically, the transformer's input combines both the 2D keypoints $\mathcal{X}$ and the noisy 3D poses $\mathcal{Y}_t$ affected by Gaussian noise. Moreover, the spatial and temporal transformers are utilized to iteratively capture both spatio-temporal information within the input sequence. The input token size reshaped as $Z_s \in \mathbb{R}^{(B\times F)\times J\times C}$ and $Z_t \in \mathbb{R}^{(B\times J)\times F\times C}$, where C denotes the channel size. During the model training phase, the transformer encoder processes input data denoted as Z_s and Z_t. Subsequently, the denoise model utilizes the acquired knowledge from the transformer network to generate precise 3D poses, which are represented as $\widetilde{\mathcal{Y}}_0 \in \mathbb{R}^{B\times J\times F\times 3}$.

3.1.1 3D human pose evaluation via diffusion model

In this section, we initially examine the diffusion model [33–35], comprising both the diffusion module and the reverse module. In the beginning stage, the diffusion module gradually diffuses the Gaussian noise to the ground truth 3D poses and the reverse module employs the denoiser to recover the unstructured 3D poses. The detailed architecture of the diffusion model is illustrated in Fig. 3.

Diffusion module The diffusion module $\mathcal{Q}$ produces the contaminated samples $\mathcal{Y}_1, \mathcal{Y}_2, ..., \mathcal{Y}_T$ by applying multiple levels of noise $\Psi \sim \mathcal{N}(0, I)$ to the ground truth 3D pose $\mathcal{Y}_0$ at each timestep $t \in [0, \mathcal{T}]$, where, $\mathcal{T}$ signifies the maximum timestep. Specifically, the diffusion module $\mathcal{Q}$ is defined as follows:

$$\mathcal{Q}(\mathcal{Y}_{1:\mathcal{T}} \mid \mathcal{Y}_0) := \prod_{t=1}^{\mathcal{T}} \mathcal{Q}(\mathcal{Y}_t \mid \mathcal{Y}_{t-1}), \tag{1}$$

$$\mathcal{Q}(\mathcal{Y}_t \mid \mathcal{Y}_{t-1}) := \mathcal{N}(\mathcal{Y}_t; \sqrt{1-\Upsilon_t}\mathcal{Y}_{t-1}, \Upsilon_t I), \tag{2}$$

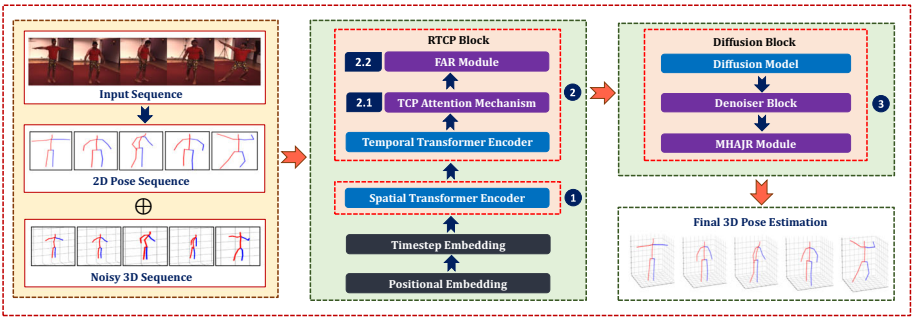

Fig. 1 The overall architecture of the proposed MHAFormer method. Initially, the input is formed by combining the 2D and 3D noisy sequences. Subsequently, the input sequences undergo processing in the proposed method section. The presented method contains of three main components: spatial transformer encoder, temporal transformer encoder, and denoiser. The comprehensive diagram of the spatio-temporal transformer encoders and diffusion model is illustrated in Figs. 2 and 3. Finally, we predicted the 3D pose sequences

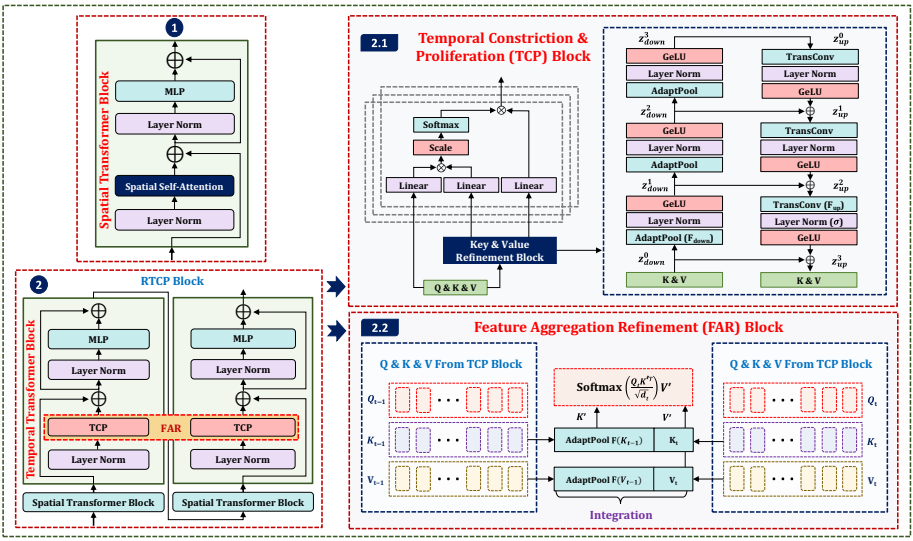

Fig. 2 The schematic diagram of the spatial and refined temporal constriction and proliferation (RTCP) transformers. The spatial transformer block is exhibited in Section 1. The temporal transformer block is illustrated in Section 2. Subsequently, we presented the temporal constriction and proliferation (TCP) transformer block and feature aggregation refinement (FAR) module in Sections 2.1 and 2.2. The network is constructed by stacking TCP blocks to facilitate the extraction of multi-scale information through attention. Additionally, the FAR module is introduced to fuse inter-block information and combine the keys and values for the feature aggregation to boost the transformer's ability

where Υ_t denotes the cosine noise variance and I indicates the identity matrix. Following the DDPM method [36], we can reformulate (1) to achieve $\mathcal{Y}_t$ from $\mathcal{Y}_0$ without the need for

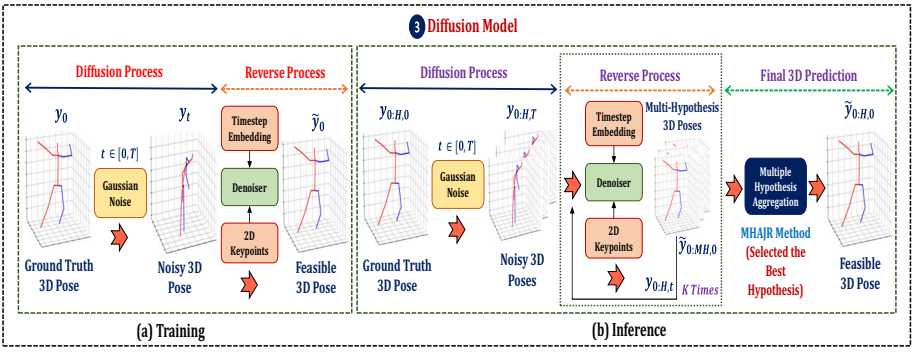

Fig. 3 The overall diagram of the denoiser model, which consists of two main sections: (a) Training and (b) Inference. (a) Training: we gradually distribute Gaussian noise to the ground truth $\mathcal{Y}_0$ to obtain the noisy 3D pose $\mathcal{Y}_t$. Next, the denoiser method utilizes the 2D keypoints $\mathcal{X}$ and timestep embedding t to restore the feasible 3D pose $\widetilde{\mathcal{Y}}_0$. (b) Inference: we process the Gaussian noise to the ground truth 3D pose $\mathcal{Y}_{0:H,0}$ to obtain the noisy 3D poses $\mathcal{Y}_{0:H,\mathcal{T}}$. Then, we produce the multiple 3D hypotheses $\widetilde{\mathcal{Y}}_{0:\mathcal{MH},0}$ by using the denoiser model. Subsequently, the iterative process is employed K times to enhance the final pose. This is achieved by passing the generated hypotheses with varying levels of noise $\mathcal{Y}_{0:H,t}$ into the denoiser model during each iteration. In the end, the final 3D pose $\widetilde{\mathcal{Y}}_{0:H,0}$ is selected by the utilization of a multiple-hypothesis aggregation technique

iteration.

$$\begin{aligned} \mathcal{Q}(\mathcal{Y}_t \mid \mathcal{Y}_0) &:= \mathcal{N}(\mathcal{Y}_t; \sqrt{\bar{\Theta}_t}\mathcal{Y}_0, (1-\bar{\Theta}_t)I), \\ &:= \sqrt{\bar{\Theta}_t \mathcal{Y}_0} + \Psi\sqrt{1-\bar{\Theta}_t}, \Psi \sim \mathcal{N}(0, I), \end{aligned} \tag{3}$$

where $\bar{\Theta}_t = \prod_{s=0}^{t} \Theta_s$, $\Theta_t = 1 - \Upsilon_t$ and Ψ indicates the Gaussian noise.

Reverse module As illustrated in Fig. 3, the reverse module is designed to reconstruct a plausible 3D posture $\widetilde{\mathcal{Y}}_0$ from the noisy 3D posture $\mathcal{Y}_t$. In this scenario, we directly predict the feasible 3D pose through a learned transformer network $\mathcal{D}$, which is formulated as follows:

$$\widetilde{\mathcal{Y}}_0 = \mathcal{D}(\mathcal{Y}_t, \mathcal{X}, t), \tag{4}$$

$$\mathcal{L} = \mathbb{E}_{\mathcal{Y}_t, t}[\left\| \mathcal{Y}_0 - \widetilde{\mathcal{Y}}_0 \right\|_2^2]. \tag{5}$$

Transformer network Unlike the existing approaches [16, 37], our works leverage the novel transformer-based network as our backbone denoted as $\mathcal{D}$, which consists of the STE and TTE methods. Specifically, the TTE employs the TCP and FAR modules to obtain multi-scale information. This additional information is added into the denoiser model $\mathcal{D}$ (Transformer network), which helps to reduce uncertainty and produce more accurate 3D poses. Additionally, the comprehensive information on STE, TTE, TCP, and FAR modules is exhibited in Sections 3.1.2, 3.1.3, 3.1.4, and 3.1.6. Moreover, we leverage the MHAJR approach in the diffusion model, which aggregates the multiple hypotheses and selects the optimal hypotheses by considering the shortest distance in terms of reprojection error. In Subsection 3.1.7, we explained the more information about the MHAJR method. Following this, the diffusion model is primarily divided into two key segments: Training and Inference.

Training As illustrated in Fig. 3 (Left), the Gaussian noise $\Psi \sim \mathcal{N}(0, I)$ is diffused to the ground truth 3D posture $\mathcal{Y}_0$ to obtain the corrupted posture $\mathcal{Y}_t$.

$$\mathcal{Q}(\mathcal{Y}_t \mid \mathcal{Y}_0) = \sqrt{\bar{\Theta}_t \mathcal{Y}_0} + \Psi\sqrt{1-\bar{\Theta}_t}. \tag{6}$$

Afterward, $\mathcal{Y}_t$ is forwarded to a denoiser $\mathcal{D}$, which aims to reconstruct the 3D pose $\widetilde{\mathcal{Y}}_0$ without the presence of noise.

$$\widetilde{\mathcal{Y}}_0 = \mathcal{D}(\mathcal{Y}_t, \mathcal{X}, t), \tag{7}$$

where $\mathcal{X}$ denotes the 2D keypoints and t indicates the timestep. The overall framework is carried out through a mean squared error (MSE) loss.

$$\mathcal{L} = \left\| \mathcal{Y}_0 - \widetilde{\mathcal{Y}}_0 \right\|_2^2. \tag{8}$$

Inference: The noisy 3D hypothesis $\mathcal{Y}_{0:H,\mathcal{T}}$ is obtained through sampling Gaussian noise. As illustrated in Fig. 3 (Right), the multiple 3D hypothesis is predicted by processing noisy 3D hypotheses $\mathcal{Y}_{0:H,\mathcal{T}}$ to the denoiser model. Subsequently, the iterative process is employed K times to refine the final pose, which is achieved by passing the generated hypotheses $\mathcal{Y}_{0:\mathcal{MH},0}$ with varying levels of noise $\mathcal{Y}_{0:H,t}$ into the denoiser model during each iteration using DDIM [38] technique. The process is defined as follows:

$$\mathcal{Y}_{0:H,t'} = \sqrt{\bar{\Theta}_{t'}}.\mathcal{Y}_{0:\mathcal{MH},0} + \sqrt{1-\bar{\Theta}_{t'} - \sigma_t^2}.\Psi_t + \sigma_t \Psi, \tag{9}$$

where t denotes the present timestep and t' represents the next timestep. Moreover, the timestep is expressed as follows: $t = \mathcal{T}.(\frac{1-k}{K})$, where $k \in [0, K)$. Following this, Ψ_t and σ_t

are defined as follows:

$$\Psi_t = (\mathcal{Y}_{0:H,t} - \sqrt{\bar{\Theta}_t}.\mathcal{Y}_{0:\mathcal{MH},0})/\sqrt{1-\bar{\Theta}_t}, \tag{10}$$

$$\sigma_t = \sqrt{(1-\bar{\Theta}_{t'})/(1-\bar{\Theta}_t)}.\sqrt{1-\bar{\Theta}_t/\bar{\Theta}_{t'}}, \tag{11}$$

where Ψ_t indicates the Gaussian noise at the t_{th} timestep, which is obtained from (6) and σ_t is a stochastic diffusion process. Specifically, the feasible 3D pose $\widetilde{\mathcal{Y}}_{0:H,0}$ is estimated by passing the noisy 3D pose $\mathcal{Y}_{0:H.t'}$ to the denoiser model with 2D keypoints $\mathcal{X}$. The denoising process is defined as follows:

$$\widetilde{\mathcal{Y}}_{0:H,0} = \mathcal{D}(\mathcal{Y}_{0:H.t'}, \mathcal{X}, t). \tag{12}$$

Conditional 2D keypoints Our main objective is to predict 3D pose hypotheses based on 2D keypoints, utilizing the noisy 3D poses represented as $\mathcal{Y}_{0:H:t}$ in the denoising process. Nevertheless, the input of the noisy 3D poses is insufficient for predicting plausible 3D poses. Therefore, we employed the 2D keypoints $\mathcal{X}$ as an additional input to the denoiser, which enhances the denoising process. Moreover, we deploy various strategies to fuse 2D keypoints detectors & noisy 3D poses, including concatenation, cross-attention, and many other methods. Notably, we directly utilize the concatenation technique to integrate the 2D keypoints with noisy 3D poses and forward this combination as input to the denoiser. Furthermore, the schematic diagram of the denoiser model is exhibited in Fig. 3.

Timestep embedding technique To adequately handle different levels of noise in 3D poses, it is crucial to incorporate details regarding the current timestep denoted as t. This timestep parameter signifies the frequency at which Gaussian noise is introduced, allowing the denoiser to adapt and handle diverse noise levels across different timesteps. Following the DDPM paradigm, we employ a sinusoidal functional to convert the t into a dedicated embedding for specific timestep. This embedding is then incorporated into the input embedding, following a similar approach to the integration of positional embeddings in transformer models.

Hence, the proposed approach enables the flexibility for users to customize the number of hypotheses (H) and iterations (K) according to their preferences. As a result, we can generate multiple hypotheses, which gradually improve the final predictions throughout the inference phase. Specifically, this approach resolves the issue of a fixed number of hypotheses encountered in earlier methods [13, 39, 40].

3.1.2 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 $Z_s \in \mathbb{R}^{(B\times F)\times J\times C}$. The tokens are subsequently integrated with a spatial position matrix $\mathcal{E}_s \in \mathbb{R}^{(B\times F)\times J\times C}$ and introduced into the key components of the transformer model such as multi-layer perceptron and multi-head self-attention, as explained in [26]. The dimensions of the tokens remain constant after the feature extraction by the spatial transformer encoder. This process is defined as follows:

$$\begin{aligned} \widetilde{\mathrm{ST}}(Z_s) &= Z_s + \mathrm{MSA}(Z_s), \\ \mathrm{ST}(Z_s) &= \widetilde{\mathrm{ST}}(Z_s) + \mathrm{MLP}\left(\widetilde{\mathrm{ST}}(Z_s)\right). \end{aligned} \tag{13}$$

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

3.1.3 Temporal transformer encoder

The temporal transformer encoder utilizes the feature aggregation refinement (FAR) 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 $Z_t \in \mathbb{R}^{(B\times J)\times F\times C}$ in the temporal encoder. Afterward, a temporal positional encoding $\mathcal{E}_t \in \mathbb{R}^{(B\times J)\times F\times C}$ is embedded to the input token before being passed into TTE. The process is formulated as follows:

$$\begin{aligned} \widetilde{\mathrm{TT}}\left(Z_t\right) &= Z_t + \mathrm{FAR}\left(Z_{t-1}, Z_t\right), \\ \mathrm{TT}\left(Z_t\right) &= \widetilde{\mathrm{TT}}\left(Z_t\right) + \mathrm{MLP}\left(\widetilde{\mathrm{TT}}\left(Z_t\right)\right). \end{aligned} \tag{14}$$

In (14), 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.

3.1.4 Temporal constriction & proliferation attention module

To enhance the comprehensive extraction of information from the self-attention layer, we introduce a 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 $Q \in \mathbb{R}^{n\times d}$ is maintained, whereas the keys and values undergo processing across multiple stages. This approach enables the attention matrix to learn multi-scale information while preserving consistent temporal resolution. As shown in Fig. 2, the TCP block progressively compresses both K and V. The sequence length of K and V is systematically reduced by sampling ratio (r) at each stage.

3.1.5 TCP attention block with keys & values

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. 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:

$$\begin{aligned} z_{\text{down}}^{0} &= z, \quad z_{\text{down}}^{l+1} = \sigma\left(\mathrm{LN}\left(\mathcal{F}_{\text{down}}\left(z_{\text{down}}^{l}\right)\right)\right), \\ z_{\text{up}}^{0} &= z_{\text{down}}^{m}, \quad z_{\text{up}}^{l+1} = \sigma\left(\mathrm{LN}\left(\mathcal{F}_{\text{up}}\left(z_{\text{up}}^{l}\right)\right)\right) + z_{\text{down}}^{m-1-l}, \end{aligned} \tag{15}$$

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. $z_{\text{down}}^{l} \in \mathbb{R}^{\frac{n}{r^{l}}\times d}$ 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. As shown in Fig. 2 (2.1), the TCP block constricts and amplifies the feature z through m stages.

The integration of constriction and proliferation attention block for keys and values enables the extraction of more refined representations from the data. The constriction process facilitates information distillation, reducing noise levels, while the proliferation process restores the lost information and integrates it with higher-level features. The direct utilization of keys and values may expose the model to vulnerability to data noise, which leads to degradation of model performance. Therefore, the TCP block serves as a robust and efficient method for refining key and value representations, leading to an enhancement in the overall performance of the model.

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 [41] 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 exhibited in Fig. 2 (2.2). Each module is built with two neighboring 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$(\cdot)$. This process is formulated as follows:

$$\begin{aligned} \mathrm{FAR} &= \mathrm{Attn}\left(Z_{t-1}, Z_t\right), \\ &= \mathrm{Attn}\left(Q_t, K', V'\right), \\ &= \mathrm{Softmax}\left(\frac{Q_t K'^{\top}}{\sqrt{d}}\right) V', \\ K' &= \mathrm{Concat}\left(K_t, \mathcal{F}\left(K_{t-1}\right)\right), \\ V' &= \mathrm{Concat}\left(V_t, \mathcal{F}\left(V_{t-1}\right)\right). \end{aligned} \tag{16}$$

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 (16), 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.

3.1.7 Multiple hypothesis aggregation via join level reprojection approach

Initially, the previous probabilistic 3D pose estimation approach [39] mainly focused on generating the 3D hypotheses, which aggregate multiple hypotheses to obtain the optimal

final 3D pose. These probabilistic methods typically emphasize superiority by reporting the error corresponding with the best 3D hypothesis (nearest to the ground truth 3D pose). However, the probabilistic method is not appropriate for real-time scenarios when there is a lack of accessible ground truth values. Consequently, the final prediction is obtained from the averaged 3D pose, showcasing a performance that is notably less impressive than the best result. A couple of studies [40, 42] explore various techniques for aggregating multiple hypotheses, which demonstrate their superior effectiveness over averaging when tested on videos recorded in real-world settings. Moreover, these investigations fail to demonstrate a relation between 3D predictions & 2D observations. Therefore, additional enhancements are required to obtain more precise outcomes.

To address these concerns, we initially verified the maximum efficiency of aggregation approaches at two tiers by selecting the optimal hypothesis that demonstrates the nearest proximity to the ground truth. During the initial phase, the pose-level selection method is employed to choose the most favorable pose for the final result. In the second stage, the joint-level selection process identifies the optimal joints, which are subsequently integrated into the final prediction. Hence, the joint-level selection outperforms pose-level selection in terms of accuracy in pose estimation. Specifically, this study employed joint-level selection, which allows the model to adapt more effectively to variations in body movements and enhance the overall performance of pose estimation. More precisely, the model leverages input 2D keypoints to guide the selection of the most probable 3D hypothesis $\widetilde{\mathcal{Y}}_0$. Subsequently, the final prediction is determined by the maximum posterior probability $\mathcal{Q}(\widetilde{\mathcal{Y}}_0|\mathcal{X}, \widetilde{\mathcal{Y}}_{0:H,0})$ during the inference process. Despite the absence of depth information for 3D poses, 2D keypoints adeptly indicate potential locations of human joints in 3D space. Therefore, 2D keypoints remain pivotal in the context of aggregating multiple hypotheses.

Inspired by the aforementioned discussion, we propose multiple hypothesis aggregation via the join-level reprojection (MHAJR) approach, which is exhibited in Fig. 4. We employ the estimated intrinsic camera parameters to project the 3D hypotheses $\widetilde{\mathcal{Y}}_{0:H:0}$ into the 2D pose. Following that, we compute the distance between the input 2D keypoints and multiple hypotheses and select the optimal hypothesis by considering the shortest distance in terms of reprojection error. This approach enables the selection of specific hypotheses for different

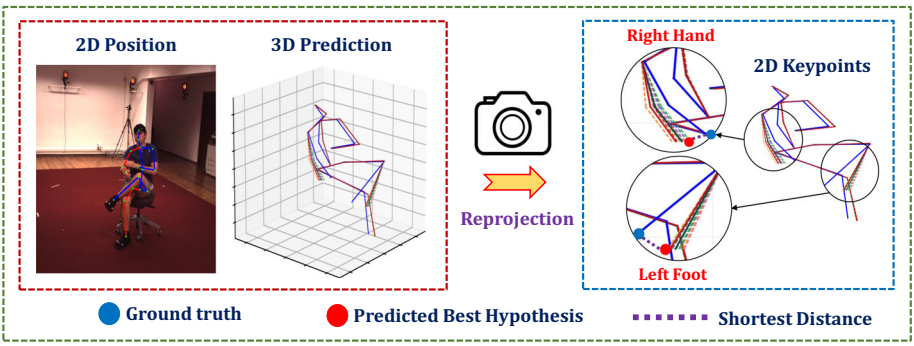

Fig. 4 The overall architecture of the multiple hypotheses aggregation via joint-level reprojection (MHAJR) method. The MHAJR approach reprojected the 3D pose hypotheses into the 2D poses and we compare each joint with corresponding input 2D keypoints. Subsequently, we choose the optimal hypothesis with the shortest distance to ground truth 3D poses. The dotted line denotes the predicted 3D hypotheses, we selected the best hypothesis that achieved the most joints closest to the ground truth, and the solid blue line indicates ground truth 3D postures

joints, which combines all the selected joints to formulate the final prediction $\widetilde{\mathcal{Y}}_0$. This process is defined as follows:

$$\widetilde{\mathcal{Y}}_0^{(i)} = \widetilde{\mathcal{Y}}_{h',0}^{(i)}, \quad h' = \underset{h \in [0,H]}{\arg\min} \left\| \mathcal{P}\left(\widetilde{\mathcal{Y}}_{h,0}^{(i)}\right) - \mathcal{X}^{(i)} \right\|_2, \tag{17}$$

where i denotes the joint index and $\mathcal{P}(\cdot)$ represents the reprojection method.

4 Experimental results and analysis

4.1 Dataset and evaluation metrics

The presented approach is estimated on two challenging benchmark datasets Human3.6M [43] and MPI-INF-3DHP [44]. 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 [10, 11, 13] the training set consists of five subjects (S1, S5, S6, S7, S8), while the evaluation set involves two subjects (S9 and S11). The MPIINF-3DHP contains both indoor and outdoor environmental datasets, involving more diverse motions than Human3.6M.

The experiments employ two standard evaluation protocols. Initially, the MPJPE quantifies the average Euclidean distance between the predicted joint pose and their corresponding ground truth pose. This protocol is commonly represented as protocol-1 in various studies [10, 11]. Subsequently, the P-MPJPE is a metric used to evaluate the accuracy of 3D pose estimation models. It provides a measure of how well the estimated poses align with the ground truth, considering both translation and rotation. This specific protocol is commonly denoted as protocol-2 in the relevant literature [12, 13].

4.2 Implementation details

In this work, we implemented the proposed MHAFormer 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 [45] to evaluate the effectiveness of our method. Moreover, the model was trained using the Adam optimizer, and we established the batch size, dropout rate, and activation function at 1024, 0.1, and GELU, respectively. Following the approach demonstrated by [12], we defined the input sequence length as 243 for the Human3.6M dataset and 27 for the MPI-INF-3DHP dataset. In the training phase, both the hypotheses (H) and iterations (K) start with an initial value of 1. In the inference phase, these values are adjusted to 20 for hypotheses and 10 for iterations. The maximum timestep value $\mathcal{T}$ is defined as 1000.

4.3 Human3.6M dataset evaluation

We compare the proposed approach with different cutting-edge deterministic approaches on the Human3.6M dataset as exhibited in Table 1 (Top). As illustrated in Tables 1 and 2, the most favorable results were reported by different approaches, showcasing their efficiency under the MPJPE and P-MPJPE metrics with cascaded pyramid network (CPN)

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

Deterministic methods Protocal #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
TCN [46] (N=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 [47] (N=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 [11] (N=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 [48] (N=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 [49] (N=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 [50] (N=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 [51] (N=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 [52] (N=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 [53] (N=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 [32] (N=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
MixSTE [12] (N=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**
MLP-JCG [54]	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 [10] (N=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
Uplift & Upsample [55] (N=351)	WACV'23	38.6	41.0	37.6	39.7	44.2	**47.9**	40.9	39.8	**51.7**	60.3	43.1	41.1	41.6	28.4	29.2	**40.9**
STRFormer [56] (N=300)	IMAVIS'23	38.3	**40.2**	38.2	39.8	44.1	51.6	**39.1**	**38.8**	53.0	**54.7**	43.0	39.6	41.3	**27.5**	28.2	41.0
HDFormer [57] (N=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 [58] (N=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
JoyPose [59] (-)	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
STCFormer [60] (N=81)	CVPR'23	40.6	43.0	38.3	40.2	43.5	52.6	40.3	40.1	51.8	57.7	42.8	**39.8**	42.3	28.0	29.5	42.0
PoseFormerV2 [31] (N=27)	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	45.2
DAF-DG [61]	CVPR'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	44.4
GLA-GCN [62] (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
MHAFormer (N=243, K=1, H=1)		**37.9**	**40.1**	**37.0**	**38.9**	**41.5**	**47.4**	**39.6**	**39.2**	**50.7**	**54.5**	**41.9**	**40.1**	**40.2**	**27.4**	**27.8**	**40.2**

Table 1 continued

Deterministic methods Protocal #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Probabilistic methods																	
CVAE [42] (N=1, H=200, P-Agg)	ICCV'19	48.6	54.5	54.2	55.7	62.6	72.0	50.5	54.3	70.0	78.3	58.1	55.4	61.4	45.2	49.7	58.0
GAN [63] (N=1, H=10, P-Agg)	BMVC'20	67.9	75.5	71.8	81.8	81.4	93.7	75.2	81.3	88.8	114.1	75.9	79.1	83.3	74.3	79.0	81.1
GraphMDN [40] (N=1, H=5, P-Agg)	IJCNN'21	51.9	56.1	55.3	58.0	63.5	75.1	53.3	56.5	69.4	92.7	60.1	58.0	65.5	49.8	53.6	61.3
NF [64] (N=1, H=1, P-Agg)	ICCV'21	52.4	60.2	57.8	57.4	65.7	74.1	56.2	59.1	69.3	78.0	61.2	63.7	67.0	50.0	54.9	61.8
DiffuPose [33] (N=1, H=10, P-Agg)	IEEE/RSJ'23	43.4	50.7	45.4	50.2	49.6	53.4	48.6	45.0	56.9	70.7	47.8	48.2	51.3	43.1	43.4	49.4
DRPose [34] (N=1, H=10, P-Agg)	ICASSP'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	41.8
MHFormer [13] (N=351, H=3, P-Agg)	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++ [65] (N=351, H=1, P-Agg)	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
EMHIFormer [66] (N=81, H=1, P-Agg)	JVCIR'23	41.5	44.8	40.1	42.1	46.4	52.4	41.3	41.7	54.7	61.5	45.3	41.7	45.7	30.9	32.4	44.1
EMHIFormer [66] (N=351, H=1, P-Agg)	JVCIR'23	39.9	42.3	39.9	40.9	44.7	50.7	41.3	40.9	52.6	59.3	43.5	40.3	44.4	30.4	30.7	42.8
MHAFormer (N=243, K=1, H=1, P-Agg)		38.8	40.7	37.0	**38.9**	**41.5**	47.4	39.6	39.2	50.7	54.5	41.9	40.1	40.2	27.7	**27.8**	40.4
MHAFormer (N=243, K=10, H=20, P-Agg)		**38.7**	**40.6**	**36.9**	**38.9**	**41.5**	**47.2**	**39.5**	39.1	**50.6**	**54.4**	**41.8**	**40.0**	**40.1**	**27.6**	**27.8**	**40.3**
MHAFormer (N=243, K=10, H=20, J-Agg)		**38.6**	**40.4**	**36.7**	**38.6**	**41.1**	**46.8**	**39.4**	**38.6**	**50.2**	**54.1**	**41.5**	**39.7**	**39.8**	**27.4**	**27.6**	**40.0**
CVAE [42] (N=1, H=200, P-Best)	ICCV'19	43.8	48.6	49.1	49.8	57.6	64.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
MDN [39] (N=1, H=5, P-Best)	CVPR'19	**37.8**	43.2	43.0	44.3	51.1	57.0	39.7	43.0	56.3	64.0	48.1	45.4	50.4	37.9	39.9	46.8
GAN [63] (N=1, H=10, P-Best)	BMVC'20	62.0	69.7	64.3	73.6	75.1	84.8	68.7	75.0	81.2	104.3	70.2	72.0	75.0	67.0	69.0	73.9
GraphMDN [40] (N=1, H=200, P-Best)	IJCNN'21	40.0	43.2	41.0	43.4	50.0	53.6	40.1	41.4	52.6	67.3	48.1	44.2	49.0	39.5	40.2	46.2
NF [64] (N=1, H=200, P-Best)	ICCV'21	38.5	42.5	39.9	41.7	46.5	51.6	39.9	40.8	**49.5**	56.8	45.3	46.4	46.8	37.8	40.4	44.3
MHAFormer (N=243, K=1, H=1, P-Best)		38.9	40.7	37.0	38.9	41.5	47.4	39.6	39.3	50.7	54.5	41.9	40.1	40.2	27.7	27.9	40.4
MHAFormer (N=243, K=10, H=20, P-Best)		38.4	**40.4**	**36.7**	**38.7**	**41.2**	**46.9**	**39.4**	**39.0**	50.2	**54.0**	**41.6**	**39.8**	**39.9**	**27.5**	**27.7**	**40.1**
MHAFormer (N=243, K=10, H=20, J-Best)		**35.6**	**37.0**	**34.0**	**35.4**	**38.4**	**43.6**	**36.4**	**35.8**	**47.1**	**50.3**	**38.6**	**36.8**	**36.6**	**25.3**	**25.4**	**37.1**

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

Table 2 Quantitative evaluation on the Human3.6M dataset under protocol-2 (P-MPJPE)

Deterministic methods Protocal #2		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
TCN [46] (N=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 [11] (N=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 [48] (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 [49] (N=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 [50] (N=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 [52] (N=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 [53] (N=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 [32] (N=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
MixSTE [12] (N=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
MLP-JCG [54]	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 [10] (N=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 & Upsample [55] (N=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
STRFormer [56] (N=300)	IMAVIS'23	30.7	**32.9**	30.5	31.4	**32.7**	38.9	**30.2**	30.4	41.8	45.5	33.8	30.6	**32.6**	**21.6**	**22.7**	**32.4**
HDFormer [57] (N=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 [58] (N=81)	arXiv'23	31.1	33.7	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
JoyPose [59] (−)	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
STCFormer [60] (N=81)	CVPR'23	30.4	33.8	31.1	31.7	33.5	39.5	30.8	**30.0**	41.8	45.8	34.3	**30.1**	32.8	21.9	23.4	32.7
PoseFormerV2 [31] (N=27)	CVPR'23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	35.6
DAF-DG [61]	CVPR'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	34.6
GLA-GCN [62] (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
MHAFormer (N=243, K=1, H=1)		**29.5**	**33.0**	**30.4**	**31.2**	**32.2**	**37.1**	**30.1**	**30.3**	**41.3**	**44.7**	**34.2**	**30.5**	**32.0**	**21.5**	**22.6**	**32.0**
Probabilistic methods																	
CVAE [42] (N=1, H=200, P-Agg)	ICCV'19	35.3	35.9	45.8	42.0	40.9	52.6	36.9	35.8	43.5	51.9	44.3	38.8	45.5	29.4	34.3	40.9
GAN [63] (N=1, H=10, P-Agg)	BMVC'20	42.1	44.7	45.4	51.0	49.3	51.5	41.2	46.2	57.5	70.8	48.7	44.1	50.8	42.1	43.7	48.7

Table 2 continued

Deterministic methods Protocal #2		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
GraphMDN [40] (N=1, H=5, P-Agg)	IJCNN'21	39.7	43.4	44.0	46.2	48.8	54.5	39.4	41.1	55.0	69.0	48.0	43.7	49.6	38.4	42.4	46.9
NF [64] (N=1, H=1, P-Agg)	ICCV'21	37.8	41.7	42.1	41.8	46.5	50.2	38.0	39.2	51.7	61.8	45.4	42.6	45.7	33.7	38.5	43.8
DiffuPose [33] (N=1, H=10, P-Agg)	IEEE/RSJ'23	35.9	40.3	36.7	41.4	39.8	43.4	37.1	35.5	46.2	59.7	39.9	38.0	41.9	32.9	34.2	39.9
DRPose [34] (N=1, H=10, P-Agg)	ICASSP'24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	33.7
MHFormer [13] (N=351, H=3, P-Agg)	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 [13] (N=351, H=3, P-Agg)	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++ [65] (N=351, H=1, P-Agg)	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
EMHIFormer [65] (N=81, H=1, P-Agg)	JVCIR'23	31.4	35.1	32.6	34.1	35.5	40.0	31.9	32.4	43.8	49.2	36.6	32.3	35.6	24.2	26.2	34.7
EMHIFormer [65] (N=351, H=1, P-Agg)	JVCIR'23	31.6	34.2	32.8	33.5	34.9	39.4	32.4	32.0	42.2	47.7	36.0	31.8	35.0	24.3	25.0	34.2
MHAFormer (N=243, K=1, H=1, P-Agg)		30.9	33.3	30.5	31.4	32.2	37.1	30.7	**30.3**	41.3	**44.7**	**34.2**	**30.7**	32.0	21.6	22.6	32.2
MHAFormer (N=243, K=10, H=20, P-Agg)		**30.8**	**33.1**	**30.4**	**31.3**	**32.1**	**37.0**	**30.6**	**30.2**	**41.1**	**44.6**	**34.1**	30.9	**31.9**	**21.5**	**22.5**	**32.1**
MHAFormer (N=243, K=10, H=20, J-Agg)		**30.7**	**33.2**	**30.3**	**31.2**	**32.0**	**36.9**	**30.5**	**30.2**	**41.0**	**44.6**	**34.1**	**30.8**	**31.8**	**21.4**	**22.4**	**32.0**
MDN [39] (N=1, H=5, P-Best)	CVPR'19	35.5	39.8	41.3	42.3	46.0	48.9	36.9	37.3	51.0	60.6	44.9	40.2	44.1	33.1	36.9	42.6
CVAE [42] (N=1, H=200, P-Best)	ICCV'19	**27.6**	**27.5**	34.9	32.3	33.3	42.7	28.7	**28.0**	**36.1**	**42.7**	36.0	30.7	37.6	24.3	27.1	32.7
GAN [63] (N=1, H=10, P-Best)	BMVC'20	38.5	41.7	39.6	45.2	45.8	46.5	37.8	42.7	52.4	62.9	45.3	40.9	45.3	38.6	38.4	44.3
GraphMDN [40] (N=1, H=200, P-Best)	IJCNN'21	30.8	34.7	33.6	34.2	39.6	42.2	31.0	31.9	42.9	53.5	38.1	34.1	38.0	29.6	31.1	36.3
NF [64] (N=1, H=200, P-Best)	ICCV'21	27.9	31.4	29.7	30.2	34.9	37.1	**27.3**	28.2	39.0	46.1	34.2	32.3	33.6	26.1	27.5	32.4
MHAFormer (N=243, K=1, H=1, P-Best)		30.9	33.3	30.5	31.4	32.2	37.1	30.7	30.3	41.3	44.7	34.2	**30.6**	32.0	21.6	22.6	32.2
MHAFormer (N=243, K=10, H=20, P-Best)		30.6	33.0	**30.2**	**31.1**	**32.0**	**36.7**	30.4	30.0	40.7	44.3	**33.8**	30.7	**31.7**	**21.5**	**22.4**	**32.0**
MHAFormer (N=243, K=10, H=20, J-Best)		**27.8**	**30.1**	**27.9**	**28.3**	**29.6**	**34.0**	**28.0**	**27.2**	**38.0**	**40.9**	**31.3**	**28.3**	**29.0**	**19.6**	**20.5**	**29.4**

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

input, respectively. Furthermore, the qualitative evaluation results of our proposed method on testing sequences (S9 & S11) of the Human3.6M dataset are exhibited in Figs. 5 and 6. Moreover, we utilize the 2D pose detectors obtained through the widely-used CPN [45] and ground truths as inputs for training. Specifically, our proposed MHAFormer method obtains superior performance of 40.2% and 32.0% under the MPJPE and P-MPJPE metrics. When compared to transformer-based methods such as PoseFormer, MixSTE, STR-Former, HDFormer, and MSTFormer, our proposed approach shows improvements in MPJPE values by (4.1%, 0.7%, 0.8%, 2.4%, 2.5%) under protocol-1 and in P-MPJPE values by (2.6%, 0.6%, 0.4%, 1.1%, 1.7%) under protocol-2.

As illustrated in Table 1 (Bottom), the proposed approach is compared with other probabilistic methods. The results are presented according to four specified experimental conditions:

P-Agg: The pose-level aggregation method represents the most straightforward approach, where the 3D coordinates of each joint are averaged across all pose hypotheses.

J-Agg (Proposed): We utilize the joint-level aggregation approach (MHAJR) to obtain the final prediction, which individually aggregates results for each joint and improves the 3D pose accuracy.

P-Best : This approach entails selecting the single-pose hypothesis with the highest overall score, representing the 3D posture that closely aligns with the ground truth.

J-Best (Proposed): We adopt the optimal joint-level approach, selecting the hypothesis nearest to the ground truth and integrating the selected joints to construct the final 3D pose.

Moreover, compared with transformer-based methods under P-Agg ($H = 20$) evaluation settings, our approach has a significant improvement over MHFormer ($2.7mm$), MHFormer++ ($2.2mm$), EMHIFormer (N=81) ($3.8mm$), and EMHIFormer (N=351) ($2.5mm$), respectively. Specifically. when increasing the hypothesis $H(1 \rightarrow 20)$ under the P-Agg and P-Best levels, the proposed method doesn't significantly improve the performance ($40.2mm \rightarrow 40.3mm$) and ($40.2mm \rightarrow 40.1mm$). The outcome encourages us to propose

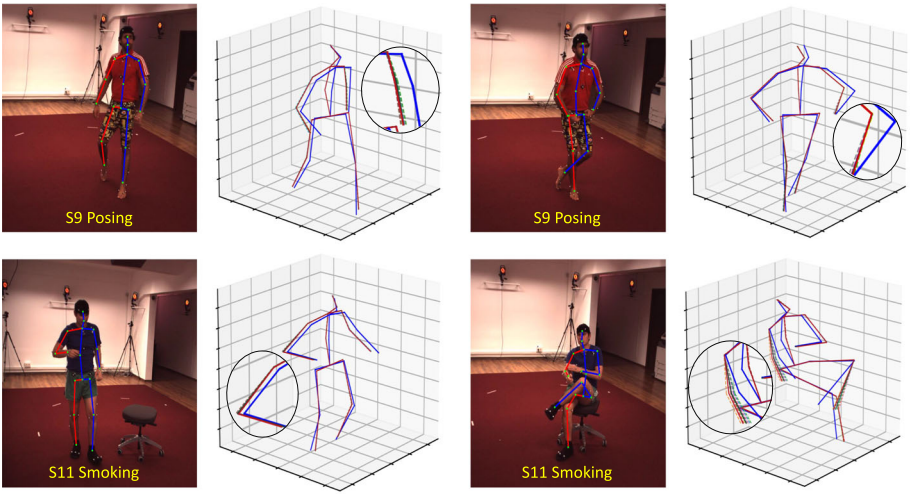

Fig. 5 Qualitative comparison results on Human3.6M dataset. We evaluated our proposed method using the test set ($S9$ & $S11$) sequences such as posing and smoking. The dashed line represents multiple hypotheses ($H = 10$), we selected the most suitable hypothesis by choosing the one with the minimum distance to the 2D input and the solid blue line indicates the ground-truth 3D pose

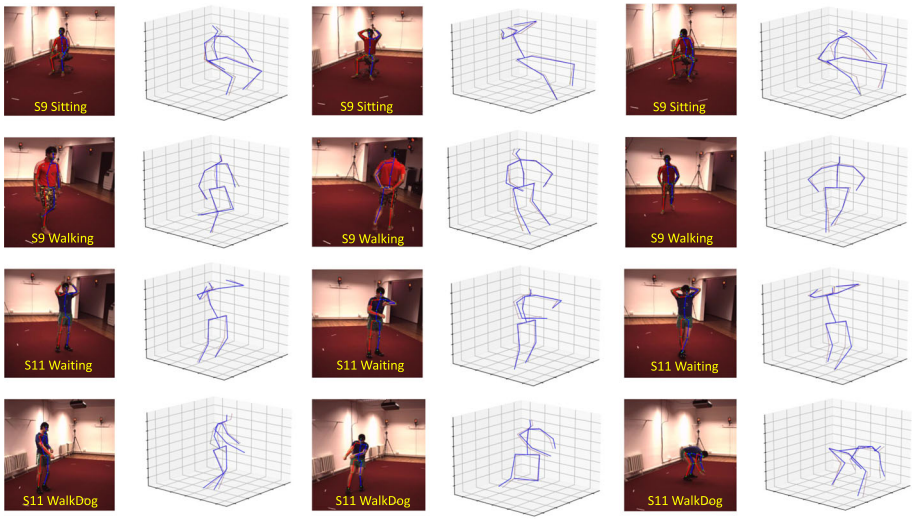

Fig. 6 Qualitative evaluation results on Human3.6M dataset. We evaluated our suggested approach under multiple hypotheses with different challenging test sets ($S9$ & $S11$) sequences. The dotted line denotes the multiple hypotheses ($H = 5$), and the solid blue line represents the ground-truth 3D pose

a J-Best setting, which demonstrates a significant improvement (40.2 $\rightarrow$ 37.1 mm). In particular, the best performance is attained when employing the J-Best setting, where specific joints within the same hypothesis are evaluated. This observation encourages us to introduce J-Agg ($H = 20$), aiming to exploit the distinctions at the joint level among hypotheses. In contrast to P-Agg, the efficiency is improved under the J-Agg setting, demonstrating an improvement from (40.3 $\rightarrow$ 40.0mm).

4.4 MPI-INF-3DHP dataset evaluation

To evaluate our proposed approach against state-of-the-art methods, we utilize the MPI-INF-3DHP benchmark dataset and measure performance using PCK, AUC, and MPJPE metrics. Also, the MPI-INF-3DHP dataset is a large collection of 3D human pose data, including 1.3 million frames captured in a multi-camera studio with ground truth obtained through commercial markerless motion capture. It presents various motions performed by eight actors in both indoor and outdoor environments. In addition, the evaluation results are illustrated in Table 3 and our proposed approach obtains the best results with PCK is 97.7%, AUC is 76.9% and MPJPE is 31.5mm, particularly in the scenario of a single hypothesis ($H = 1$). Moreover, when increasing the hypothesis $H(1 \rightarrow 20)$ under the P-Agg and P-Best levels, our approach attains the second-best outcomes with PCK, AUC, and MPJPE (98.0%, 77.0% & 31.3mm), and (98.0%, 77.2% & 30.6mm) respectively. Specifically, when evaluating our proposed method performance under the J-Agg and J-Best (Proposed) levels, we obtain the top results with PCK, AUC, and MPJPE (98.1%, 77.1% & 30.9mm), and (98.3%, 78.4% & 29.2mm) respectively. Compared with transformer-based approaches such as MHFormer, and MixSTE, our method obtains the best improvements in terms of PCK, AUC, and MPJPE by (4.3%, 15.1% & 28.8mm) and (1.4%, 2.6% & 6.2mm) under the J-

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

Methods		PCK ↑	AUC ↑	MPJPE ↓
TCN [46] (N=81)	CVPR'19	86.0	51.9	84.0
Anatomy [49] (N=81)	TCSVT'21	87.9	54.0	78.8
PoseFormer [11] (N=9, H=3, P-Agg)	ICCV'21	88.6	56.4	77.1
U-CDGCN [50] (N=96)	MM'21	97.9	69.5	42.5
MHFormer [13] (N=27)	CVPR'22	93.8	63.3	58.0
P-STMO [32] (N=81)	ECCV'22	97.9	75.8	32.2
MixSTE [12] (N=243)	CVPR'22	96.9	75.8	35.4
MHFormer++ [65] (N=9)	PR'23	94.8	65.8	54.0
Uplift & Upsample [55] (N=81)	WACV'23	97.9	75.8	32.2
EMHIFormer [66] (N=9)	JVCIR'23	97.1	74.9	33.8
STRFormer [56] (N=27)	IMAVIS'23	94.8	67.1	54.4
HDFormer [57] (N=32)	arXiv'23	96.8	64.0	51.5
HSTFormer [58] (N=81)	arXiv'23	97.3	71.5	41.4
JoyPose [59]	PR'24	94.1	–	–
DAF-DG [61]	CVPR'24	92.9	60.7	63.1
GLA-GCN [62] (N=27)	ICCV'23	**98.1**	76.5	**31.3**
MHAFormer (N=243, K=1, H=1)		97.7	76.9	31.5
MHAFormer (N=243, K=10, H=20, P-Agg)		**98.0**	**77.0**	**31.3**
MHAFormer (N=243, K=10, H=20, J-Agg)		**98.1**	**77.1**	**30.9**
MHAFormer (N=243, K=10, H=20, P-Best)		**98.0**	**77.2**	**30.6**
MHAFormer (N=243, K=10, H=20, J-Best)		**98.3**	**78.4**	**29.2**

The best and second-best results are highlighted in red and blue fonts, respectively

Best level. In the end, the proposed method achieves superior performance when compared to the conventional methods under the single and multiple-hypotheses evaluation.

4.5 Ablation study

4.5.1 Component analysis

To validate the impact of each integrated element in our suggested model, we conducted a thorough ablation analysis using the Human3.6M dataset under the MPJPE metric within the J-Agg and J-Best settings. In this work, we employ the MixSTE approach as our baseline method. As shown in Table 4, the MixSTE approach obtained the 40.9mm MPJPE result. Further, when integrating the diffusion module with our baseline method *(MixSTE + Diffusion)*, we achieve the 0.1mm improvement (40.9mm $\rightarrow$ 40.8mm) in the single-hypothesis case. By leveraging a diffusion process, it effectively propagates information across the pose sequence, ensuring smoother and more realistic transitions between frames. This leads to a more robust handling of occlusions and ambiguities in the data. In addition, we observe a 0.2mm increment (40.8mm $\rightarrow$ 40.6mm) when incorporating the TCP module into our proposed approach *(MixSTE + Diffusion + TCP)*. The TCP structure is a key innovation in our model, which dynamically adjusts the diffusion process based on temporal constraints and feature proliferation, enabling more accurate modeling of time-dependent changes in data.

Table 4 Ablation analysis on different proposed elements

MixSTE	Diffusion	TCP	FAR	MHAJR	Hypothesis	Setup	MPJPE
✓	✗	✗	✗	✗	1	Not Applicable	40.9
✓	✓	✗	✗	✗	1	Not Applicable	40.8
✓	✓	✓	✗	✗	1	Not Applicable	40.6
✓	✓	✓	✓	✗	1	Not Applicable	40.5
✓	✓	✓	✓	✓	1	P-Agg	40.4
✓	✓	✓	✓	✓	20	P-Agg	40.3
✓	✓	✓	✓	✓	20	**J-Agg**	**40.0**
✓	✓	✓	✓	✓	20	**J-Best**	**37.1**

The evaluation is conducted on the Human3.6M dataset with the MPJPE metric under J-Agg and J-Best setup. The results marked in red and blue denote the first & second-best outcomes, respectively

This approach adeptly captures both short-range and long-range temporal dependencies in motion sequences, enhancing the accuracy of 3D pose predictions. By including the FAR module in our proposed technique *(MixSTE + Diffusion + TCP + FAR)*, we observe a $0.1mm$ improvement ($40.6mm \rightarrow 40.5mm$). The FAR module improves feature representation by aggregating and refining spatial-temporal information. This process highlights essential features and minimizes noise, leading to more accurate pose estimation and reduced computational overhead. Specifically, when incorporating the MHAJR approach into our proposed method *(MixSTE + Diffusion + TCP + FAR + MHAJR)*, we observe a ($0.1mm$) improvement ($40.5mm \rightarrow 40.4mm$) under the P-Agg setting in a single-hypothesis scenario. When increasing the hypothesis $H(1 \rightarrow 20)$ into our proposed method, we can notice that $0.1mm$ improvement ($40.4mm \rightarrow 40.3mm$) under the P-Agg setting. In particular, our proposed method attains the second-best results, with a ($0.3mm$) improvement ($40.3mm \rightarrow 40.0mm$), when evaluated under the J-Agg setting with 20 hypotheses. Significantly, we achieved the most favorable results with a ($2.9mm$) improvement ($40.0mm \rightarrow 37.1mm$) when assessing the proposed method under the J-Best setting with 20 hypotheses. The MHAJR technique allows the model to evaluate and aggregate multiple potential hypotheses, improving accuracy by considering various plausible solutions and minimizing reprojection errors. This feature further refines pose estimation and enhances the robustness of the model. Moreover, 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 Diffusion, TCP, FAR, and MHAJR, playing crucial roles in enhancing performance within specific contexts. Finally, under the J-Best setting with 20 hypotheses, we achieved the most substantial enhancement of ($2.9mm$), emphasizing the effectiveness of our comprehensive approach in refining 3D pose estimation accuracy.

4.5.2 Hyperparameter analysis

We conducted a hyperparameter analysis of our proposed method, comparing it with cutting-edge approaches using the Human3.6M dataset, as illustrated in Table 5. The proposed MHAFormer method outperforms existing methods in terms of MPJPE, particularly excelling

Table 5 Ablation analysis of hyperparameter settings of our proposed method with cutting-edge approaches on the Human3.6M dataset

Methods	Hypotheses	Iterations	Params	Infer. FPS	MPJPE ↓
MixSTE	1	N/A	33.6	4547	40.9
P-STMO	1	N/A	6.7	3040	42.8
GraphMDN	5	N/A	–	–	61.3
PoseFormer	1	N/A	9.5	1952	44.3
MHFormer	3	N/A	24.7	–	43.0
MHFormer++	1	N/A	18.92	–	42.5
MHAFormer	1	1	34.9	4050	40.2
MHAFormer	5	20	34.9	1564	40.0
MHAFormer	10	20	34.9	1134	37.1

with 10 hypotheses and 20 iterations to achieve an MPJPE of 37.1. This significant reduction in error highlights the model's accuracy and effectiveness.

1. With 1 hypothesis and 1 iteration: MHAFormer achieves an MPJPE of 40.2, which outperforms most existing methods such as P-STMO, PoseFormer, MHFormer, and MHFormer++.
2. With 5 hypotheses and 20 iterations: The MPJPE improves slightly to 40.0, demonstrating that increasing the number of hypotheses and iterations refines the accuracy.
3. With 10 hypotheses and 20 iterations: The MPJPE significantly drops to 37.1, highlighting the advantage of using more hypotheses and iterations for better joint position accuracy.

Despite having a parameter count (34.9 million) comparable to MixSTE (33.6 million), MHAFormer performance in MPJPE shows its superior model efficiency and capability in 3D human pose estimation. The adaptability of MHAFormer, with varying hypotheses and iterations, allows it to handle more complex scenarios effectively.

4.6 Qualitative results

In this section, we first analyze the proposed multiple hypotheses performance on the Human3.6M dataset in Section 4.6.1. Subsequently, we conduct the comparison between the suggested approach and cutting-edge methods in Section 4.6.2. Finally, we investigate the performance of the proposed method using real-time videos in Section 4.6.3.

4.6.1 Qualitative comparison of Human3.6M dataset

We conducted a qualitative comparison of our proposed method with multiple hypotheses and ground truth 3D pose in the Human3.6M dataset. Also, we evaluated our proposed method on Human3.6M test set (S9 & S11) sequences such as *Sitting, Walking, Waiting*, and *WalkDog*. The qualitative comparison results are illustrated in Fig. 6. As shown in Fig. 6, the dotted lines denote our proposed multiple hypotheses ($H = 5$), and the solid blue line represents the ground truth 3D pose. Additionally, we conduct a qualitative evaluation of different sequences, such as *Posing* and *Smoking* (Test set), within the Human3.6M dataset. Moreover, the evaluation outcomes are exhibited in Fig. 5. In the case of qualitative comparison, we

generated the 10 hypotheses and selected the best hypothesis using the reprojection error. As shown in Fig. 5, the best-selected hypothesis indicates the solid red line that is nearest to the ground truth 3D pose. Furthermore, the dotted lines in each sequence represent the multiple hypotheses ($H = 10$) highlighted within the circled area.

4.6.2 Qualitative comparison of wild videos

We conduct a qualitative analysis of our proposed method, comparing it with competitive approaches such as MHFormer, GraphMLP, MixSTE, and PoseFormer. This evaluation is performed on challenging in-the-wild videos to demonstrate the effectiveness of our approach in real-world scenarios. Further, the comparison results are illustrated in Fig. 7. The deviated 3D pose prediction is emphasized within a dotted black circle. Notably, the green circle signifies locations where our approach demonstrates superior outcomes. Moreover, the 2D detector CPN [45] is utilized to extract 2D poses, which are then fed into the models to ensure a fair comparison. Despite the complicated actions and rapid motions, the suggested approach excels in generating realistic and plausible 3D predictions that outperform previous approaches.

4.6.3 Qualitative comparison of real-time videos

To demonstrate the effectiveness of our proposed approach, we carry out real-time experiments using four video sequences. Each sequence showcases intricate and challenging poses, providing a comprehensive evaluation of the proposed method's performance across a range of difficult scenarios. Also, the 2D and 3D pose prediction is illustrated in Fig. 8. Specifically, the presence of the green circle mark signifies the superior outcomes demonstrated

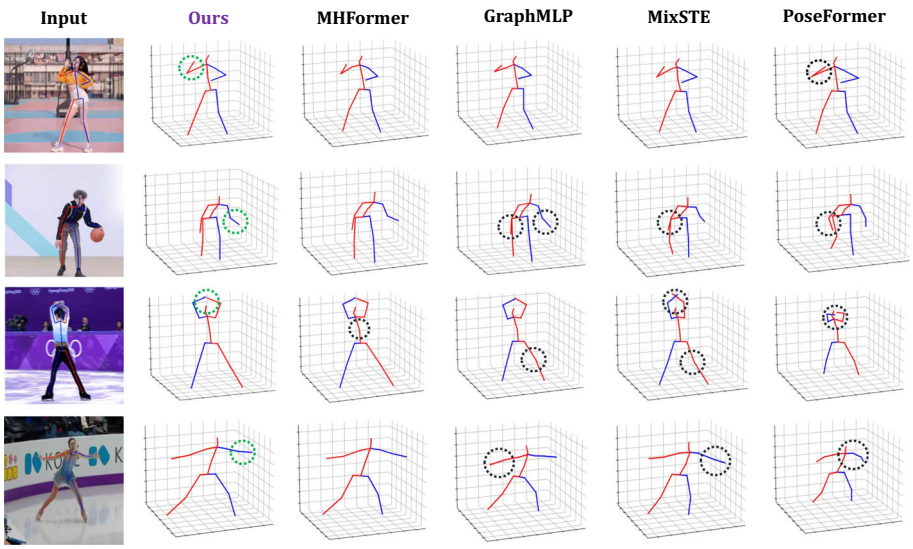

Fig. 7 Qualitative evaluation is conducted on in-the-wild videos to compare the suggested approach with cutting-edge methods such as MHFormer, GraphMLP, MixSTE, and PoseFormer. The correctly predicted pose is marked with a green circle (proposed), whereas the incorrectly predicted pose is denoted by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

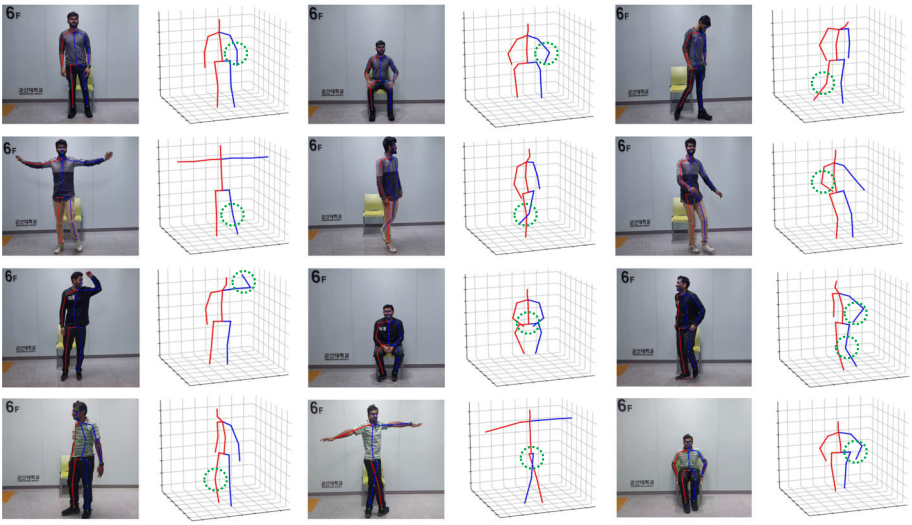

Fig. 8 We conducted a real-time experiment to evaluate our proposed method on different challenging sequences. The illustration of 2D and 3D pose estimation is showcased in the first and second columns. The presence of green color circle marks illustrates the superior performance of our proposed method

by our proposed method. Moreover, the application of our proposed approach in real-world scenarios has exhibited improved performance.

5 Conclusion

In this study, the multi-transformer encoder with a multiple-hypothesis aggregation method has been proposed for 3D pose estimation. Initially, the diffusion model has been presented to generate multiple customizable 3D hypotheses, which ensured compatibility with our specifications. The diffusion module gradually diffuses Gaussian noise to ground truth 3D poses and the reverse module employs the denoiser to recover the unstructured 3D poses. Subsequently, the MHAJR approach has been proposed to aggregate the multiple 3D hypotheses and select the optimal 3D hypothesis for the final prediction by considering reprojection errors. Specifically, the proposed RTCP transformer has integrated the spatial & temporal encoders with a temporal constriction and proliferation structure to enhance intra-block temporal modeling and extract the multi-scale information. Moreover, the FAR module has been integrated into the RTCP transformer to optimize feature fusion, which is achieved by employing two TCP attention blocks that facilitate interaction among queries, keys, and values. Finally, the extensive experimental results on the Human3.6M and MPI-INF-3DHP benchmark datasets have demonstrated the superiority of the proposed method when compared to other state-of-the-art approaches.

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.

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. Fan L, Jiang K, Zhou W, Gao Z, Luo Y (2024) 3d human pose estimation from video via multi-scale multi-level spatial temporal features. Multimed Tools Appl 1–20
2. Gu R, Jiang Z, Wang G, McQuade K, Hwang J-N (2022) Unsupervised universal hierarchical multi-person 3d pose estimation for natural scenes. Multimed Tools Appl 81(23):32883–32906
3. Liu Y, Cheng X, Ikenaga T (2024) Motion-aware and data-independent model based multi-view 3d pose refinement for volleyball spike analysis. Multimed Tools Appl 83(8):22995–23018
4. Yan L, Ma S, Wang Q, Chen Y, Zhang X, Savakis A, Liu D (2022) Video captioning using global-local representation. IEEE Transactions on Circuits and Systems for Video Technology 32(10):6642–6656
5. Yan L, Wang Q, Ma S, Wang J, Yu C (2022) Solve the puzzle of instance segmentation in videos: a weakly supervised framework with spatio-temporal collaboration. IEEE Transactions on Circuits and Systems for Video Technology 33(1):393–406
6. Cai Y, Ge L, Liu J, Cai J, Cham T-J, Yuan J, Thalmann NM (2019) Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision pp 2272–2281
7. Liu R, Shen J, Wang H, Chen C, Cheung S.-c, Asari V (2020) Attention mechanism exploits temporal contexts: real-time 3d human pose reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5064–5073
8. Liu J, Rojas J, Li Y, Liang Z, Guan Y, Xi N, Zhu H (2021) A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In: 2021 IEEE International conference on robotics and automation (ICRA), IEEE, pp 3374–3380
9. Wu Y, Kong D, Wang S, Li J, Yin B (2022) Hpgcn: hierarchical poselet-guided graph convolutional network for 3d pose estimation. Neurocomputing 487:243–256
10. Hassan MT, Ben Hamza A (2023) Regular splitting graph network for 3d human pose estimation. IEEE Trans Image Process 32:4212–4222. https://doi.org/10.1109/TIP.2023.3275914
11. Zheng C, Zhu S, Mendieta M, Yang T, Chen C, Ding Z (2021) 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 11656–11665
12. Zhang J, Tu Z, Yang J, Chen Y, Yuan J (2022) Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video, in 2022 IEEE. In: CVF Conference on computer vision and pattern recognition (CVPR), pp 13222–13232
13. Li W, Liu H, Tang H, Wang P, Van Gool L (2022) 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
14. Xue Y, Chen J, Gu X, Ma H, Ma H (2022) Boosting monocular 3d human pose estimation with part aware attention. IEEE Trans Image Process 31:4278–4291
15. Li W, Liu H, Ding R, Liu M, Wang P, Yang W (2022) Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans Multimed 25:1282–1293
16. Holmquist K, Wandt B (2023) Diffpose: Multi-hypothesis human pose estimation using diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15977–15987
17. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
18. Ma X, Su J, Wang C, Ci H, Wang Y (2021) Context modeling in 3d human pose estimation: a unified perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6238–6247
19. Fang H-S, Xu Y, Wang W, Liu X, Zhu S-C (2018) Learning pose grammar to encode human body configuration for 3d pose estimation. In: Proceedings of the AAAI conference on artificial intelligence, vol 32

20. Shuai H, Wu L, Liu Q (2023) Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. IEEE Trans Pattern Anal Mach Intell 45(4):4122–4135. https://doi.org/10.1109/TPAMI.2022.3188716
21. Hua G, Liu H, Li W, Zhang Q, Ding R, Xu X (2023) Weakly-supervised 3d human pose estimation with cross-view u-shaped graph convolutional network. IEEE Trans Multimed 25:1832–1843. https://doi.org/10.1109/TMM.2022.3171102
22. Lee K, Kim W, Lee S (2023) From human pose similarity metric to 3d human pose estimator: Temporal propagating lstm networks. IEEE Trans Pattern Anal Mach Intell 45(2):1781–1797. https://doi.org/10.1109/TPAMI.2022.3164344
23. Zou Z, Tang W (2021) Modulated graph convolutional network for 3d human pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11477–11487
24. Chen L, Liu Q (2023) Relation-balanced graph convolutional network for 3d human pose estimation. Image Vision Comput 140:104841
25. Li W, Liu H, Guo T, Ding R, Tang H (2022) Graphmlp: a graph mlp-like architecture for 3d human pose estimation. arXiv:2206.06420
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst
27. Cui Y, Yan L, Cao Z, Liu D (2021) Tf-blender: temporal feature blender for video object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8138–8147
28. Geng Z, Liang L, Ding T, Zharkov I (2022) Rstt: real-time spatial temporal transformer for space-time video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 17441–17451
29. Lu Y, Wang Q, Ma S, Geng T, Chen YV, Chen H, Liu D (2023) Transflow: transformer as flow learner. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18063–18073
30. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
31. Zhao Q, Zheng C, Liu M, Wang P, Chen C (2023) 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
32. Shan W, Liu Z, Zhang X, Wang S, Ma S, Gao W (2022) P-stmo: pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European conference on computer vision, Springer, pp 461–478
33. Choi J, Shim D, Kim HJ (2023) Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. In: 2023 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 3773–3780
34. Kang H, Wang Y, Liu M, Wu D, Liu P, Yuan X, Yang W (2024) Diffusion-based pose refinement and multi-hypothesis generation for 3d human pose estimation. In: ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 5130–5134
35. Han C, Liang JC, Wang Q, Rabbani M, Dianat S, Rao R, Wu YN, Liu D (2024) Image translation as diffusion visual programmers. arXiv:2401.09742
36. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851
37. Choi J, Shim D, Kim HJ (2023) Diffupose: monocular 3d human pose estimation via denoising diffusion probabilistic model. In: 2023 IEEE/RSJ International conference on intelligent robots and systems (IROS), pp 3773–3780. https://doi.org/10.1109/IROS55552.2023.10342204
38. Song J, Meng C, Ermon S (2020) Denoising diffusion implicit models. arXiv:2010.02502
39. Li C, Lee GH (2019) 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, pp 9887–9895
40. Oikarinen T, Hannah D, Kazerounian S (2021) Graphmdn: leveraging graph structure and deep learning to solve inverse problems. In: 2021 International joint conference on neural networks (IJCNN), IEEE, pp 1–9
41. Yu B, Jiao L, Liu X, Li L, Liu F, Yang S, Tang X (2022) Entire deformable convnets for semantic segmentation. Knowl-Based Syst 250:108871
42. Sharma S, Varigonda PT, Bindal P, Sharma A, Jain A (2019) Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2325–2334
43. Ionescu C, Papava D, Olaru V, Sminchisescu C (2013) 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

44. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C (2017) Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International conference on 3D vision (3DV), IEEE, pp 506–516
45. Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J (2018) Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7103–7112
46. Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 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
47. Zeng A, Sun X, Huang F, Liu M, Xu Q, Lin S (2020) 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, Springer, pp 507–523
48. Shan W, Lu H, Wang S, Zhang X, Gao W (2021) 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
49. Chen T, Fang C, Shen X, Zhu Y, Chen Z, Luo J (2021) Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Trans Circuits Syst Video Technol 32(1):198–209
50. Hu W, Zhang C, Zhan F, Zhang L, Wong T-T (2021) Conditional directed graph convolution for 3d human pose estimation. In: Proceedings of the 29th ACM international conference on multimedia, pp 602–611
51. Zhan Y, Li F, Weng R, Choi W (2022) 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
52. Li W, Liu H, Ding R, Liu M, Wang P, Yang W (2022) Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans Multimed 25:1282–1293
53. Xue Y, Chen J, Gu X, Ma H, Ma H (2022) Boosting monocular 3d human pose estimation with part aware attention. IEEE Trans Image Process 31:4278–4291
54. Tang Z, Li J, Hao Y, Hong R (2023) Mlp-jcg: multi-layer perceptron with joint-coordinate gating for efficient 3d human pose estimation. IEEE Trans Multimed 25:8712–8724. https://doi.org/10.1109/TMM.2023.3240455
55. Einfalt M, Ludwig K, Lienhart R (2023) 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
56. Liu X, Tang H (2023) Strformer: spatial-temporal-retemporal transformer for 3d human pose estimation. Image Vision Comput 140:104863
57. Chen H, He J-Y, Xiang W, Liu W, Cheng Z-Q, Liu H, Luo B, Geng Y, Xie X (2023) Hdformer: high-order directed transformer for 3d human pose estimation. arXiv:2302.01825
58. Qian X, Tang Y, Zhang N, Han M, Xiao J, Huang M-C, Lin R-S (2023) Hstformer: hierarchical spatial-temporal transformers for 3d human pose estimation. arXiv:2301.07322
59. Du S, Yuan Z, Lai P, Ikenaga T (2024) Joypose: jointly learning evolutionary data augmentation and anatomy-aware global-local representation for 3d human pose estimation. Pattern Recognit 147:110116
60. Tang Z, Qiu Z, Hao Y, Hong R, Yao T (2023) 3d human pose estimation with spatio-temporal criss-cross attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4790–4799
61. Peng Q, Zheng C, Chen C (2024) 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
62. Yu BX, Zhang Z, Liu Y, Zhong S-h, Liu Y, Chen CW (2023) 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
63. Li C, Lee GH (2020) Weakly supervised generative network for multiple 3d human pose hypotheses. arXiv:2008.05770
64. Wehrbein T, Rudolph M, Rosenhahn B, Wandt B (2021) Probabilistic monocular 3d human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11199–11208
65. Li W, Liu H, Tang H, Wang P (2023) Multi-hypothesis representation learning for transformer-based 3d human pose estimation. Pattern Recognit 141:109631

66. Xiang X, Zhang K, Qiao Y, El Saddik A (2023) Emhiformer: an enhanced multi-hypothesis interaction transformer for 3d human pose estimation in video. J Visual Commun Image Represent 95:103890