Multimedia Systems (2026) 32:149
https://doi.org/10.1007/s00530-025-02204-7

REGULAR PAPER

Exploring multi-transformer with fine-grained prompt-driven coupled with diffusion model for 3D human pose estimation

Sathiyamoorthi Arthanari[1,2] · Sathishkumar Moorthy[1,2] · Jae Hoo Jeong[3]

Received: 17 July 2025 / Accepted: 29 December 2025

Abstract

3D human pose estimation (HPE) predicts 3D joint coordinates from 2D images or videos. Despite advances via deep learning, current methods often overlook textual information and human knowledge, which can provide valuable implicit supervision. To address these limitations, we introduce a novel approach called the improved Fine-Grained Prompt-Driven Denoiser (FGPD) with Temporal Constriction and Proliferation (TCP) transformer, built on a diffusion model. FGPD includes three key elements: (1) the Fine-grained Part-aware Prompt (FPP) module, which creates detailed part-aware prompts by integrating accessible textual data and domain-specific knowledge about body parts with learnable prompts to provide implicit guidance; (2) the Fine-grained Prompt-Pose Communication (FPC) module enables detailed interaction between part-aware prompts and pose data, improving denoising; and (3) the Prompt-driven Timestamp Stylization (PTS) module, which combines learned prompt embeddings with temporal noise level information to enable adaptive adjustments during each step of the denoising process. The Refined Temporal Constriction and Proliferation Transformer (RTCPT) combines spatio-temporal encoders with a TCP framework to capture multi-scale attention and address depth ambiguity. It includes a Feature Aggregation Refinement (FAR) module using a cross-layer strategy within the TCP block. Substantial experiments on the Human3.6M and MPI-INF-3DHP datasets demonstrate that our approach achieves superior performance.

Keywords Diffusion model · Spatiotemporal transformer · Deep learning · 3D human pose estimation

1 Introduction

Deep learning has emerged as a powerful approach to pattern recognition and machine learning. It has significantly advanced solutions to various vision-based challenges, outperforming traditional hand-crafted methods in tasks such as image classification, object detection, semantic segmentation, 3D human pose estimation, and visual object tracking. Recent studies on object tracking focus on detecting and continuously monitoring the movement of specific objects across video frames. This process is essential for applications such as autonomous driving, robotics, and video surveillance [1–5]. On the other hand, 3D HPE aims to determine the 3D coordinates of human body joints from images or videos [6–11]. This field has garnered significant attention due to its vast number of potential applications in areas such as video surveillance, autonomous driving, fashion, and human-robot interaction. These pose estimation problems can be solved by two main methods: the end-to-end approach and the lifting-based approach. The

Communicated by Junyu Gao.

✉ Jae Hoo Jeong
jh7129@kunsan.ac.kr

Sathiyamoorthi Arthanari
sathyainfotech005@gmail.com

Sathishkumar Moorthy
msathishkumar.moorthy@gmail.com

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

[2] Department of Artificial Intelligence and Data Science, Sejong University, 209 Neungdong-ro, Gwangjin-gu 05006, Seoul-si, South Korea

[3] College of Computer and Software, Kunsan National University, 558 Daehak-ro, Gunsan-si 54150, Jeollabuk-do, South Korea

Published online: 09 February 2026

end-to-end approach directly predicts 3D joint positions from RGB images, whereas the lifting-based approach uses a two-step process. Initially, 2D keypoints are detected and subsequently projected into a 3D space. Recent advancements have leaned towards lifting-based techniques due to their ability to leverage accurate 2D pose detectors. Inspired by these advancements, we tackle the 3D HPE task by adopting the lifting-based approach, as it leverages accurate 2D keypoint annotations, simplifying the process of estimating 3D human poses. Although 2D pose detectors provide a strong foundation for converting 2D data into 3D pose estimations, this approach still faces major challenges including complex dynamic movements, body deformation, and noise in 2D data. To overcome these issues, researchers have made significant progress by using convolutional neural networks (CNNs) and transformer-based models. These advanced methods are essential for improving the accuracy and reliability of 3D pose estimation.

Convolutional Neural Networks (CNNs) have garnered substantial research interest due to their ability to capture spatial features and patterns making them particularly effective for VOT and 3D HPE [12–14]. Numerous CNN-based methods [15–19] have been developed to tackle critical challenges such as occlusion and dynamic environment pose analysis. Specifically, the approach developed for 3D pose estimation [15] leverages graph convolutional networks (GCNs) to exploit spatial-temporal relationships between body joints. It enhances pose accuracy by capturing both spatial dependencies and temporal dynamics from consecutive frames. Furthermore, the authors in [16] have proposed a new approach that combines graph attention mechanisms with spatio-temporal convolutions to improve 3D human pose estimation in videos. By effectively acqurring wide-area spatial dependencies and temporal dynamics, the method enhances performance in challenging conditions like occlusions and fast movements. In reference [17], the authors presented a HPGCN for 3D HPE, leveraging poselets to capture fine-grained body part relationships. The model improves estimation accuracy by effectively learning both local and global spatial dependencies. In addition, the authors in [18] have provided the Global–Local Adaptive GCN (GLA-GCN), which dynamically adjusts to different scales of pose structures to improve 3D HPE from monocular video. This approach integrates global and local information, which effectively handles complex and diverse poses. Subsequently, the Regular Splitting Graph Network (RSGN) [19] utilizes a regular splitting strategy to enhance feature representation and capture complex joint interactions, which improves 3D HPE by effectively modeling both local and global pose relationships. Although significant progress has been gained in GCN-based methods, the variability of human actions remains a considerable challenge, especially in complex environments such as self-occlusion and long-range dependencies.

The emergence of transformers has revolutionized numerous visual recognition tasks by effectively capturing long-range dependencies between input tokens. As pose estimation remains a fundamental challenge in computer vision, various Transformer-based architectures have been explored to enhance its accuracy and robustness. Transformers have exhibited remarkable performance across multiple domains, including object detection, image classification, and semantic segmentation, solidifying their role as a transformative technology in the field. They handle sequential data and multi-view inputs more efficiently, making them ideal for time-series tasks. Transformers also scale better and are more robust to occlusions, while CNNs struggle with these challenges due to their local receptive fields and limited expressiveness. Researchers across various fields have increasingly turned their attention to transformer-based methodologies [8, 10, 20–23], recognizing their potential to transform and innovate a wide range of domains. Technically, the PoseFormer [20] method captures both local and global temporal dependencies using transformers without convolutional or recurrent layers, achieving state-of-the-art results on temporal pose prediction tasks. It provides an efficient, lightweight architecture for modeling time-series data, improving accuracy and reducing computational complexity. In addition, the MixSTE [8] method introduces a spatio-temporal transformer that jointly learns spatial relationships between joints and temporal dynamics across frames. It achieves higher performance in multi-frame 3D human pose estimation by effectively modeling space and time with a unified architecture.

Furthermore, MHFormer [10] proposes a multi-hypothesis transformer model that predicts multiple possible 3D poses from 2D keypoints, enhancing robustness under occlusions. The method aggregates diverse hypotheses, improving accuracy in ambiguous and complex pose scenarios. Following this, the authors in [21] have introduced the HOGFormer, which integrates a hierarchical structure into the transformer architecture, efficiently capturing fine-grained and global joint dependencies for 3D human pose estimation. Moreover, DGFormer [22] introduces a dynamic graph-based transformer that dynamically updates the relationships between joints, improving spatial representation for 3D pose estimation. Moreover, the model continuously adjusts to variations in human posture in real-time, leading to greater pose accuracy and improved handling of non-rigid movements. Therefore, we employ a transformer-based architecture in the proposed 3D HPE, leveraging its strong capability to effectively capture sequential data.

Monocular 3D human pose estimation (HPE) methods often face three primary challenges: 1) Depth Ambiguity:

Accurately estimating depth from 2D skeletons remains difficult due to the inherent ambiguity in the one-to-many mapping process; 2) Structural Complexity: The human body's intricate structure, including complex inter-joint relationships and high degrees of freedom in limb movement, often results in self-occlusions and difficult-to-predict poses; 3) Limited Generalization: Existing 3D HPE datasets cover a narrow range of actions, causing models trained on these datasets to overfit and struggle to generalize across a wider variety of human activities. To address the challenge, inspired by this paper FinePOSE [24], we present a fine-grained prompt-driven denoiser with temporal constriction & proliferation transformer, utilizing diffusion models for 3D human pose estimation. This method comprises three key components: the FPP, FPC, and PTS blocks. The FPP block integrates action class, detailed human body portions (such as "person, head, body, arms, legs"), along with the kinematic records (e.g., "speed") with pose features to enrich the input for subsequent processing. The FPC block then embeds these fine-grained prompts into noisy 3D poses, fostering detailed interactions between the prompts and poses to enhance denoising. Moreover, the PTS block applies timestamps alongside the fine-grained prompts during denoising, boosting the model's adaptability and improving pose predictions across different noise levels. Specifically, the TCP is integrated with the FAR module to constructively address either close-range and far-range temporal dependency issues. To the best of our knowledge this is the pioneering work on integrating TCP and FAR modules into a fine-grained, prompt-driven denoising framework, employing the temporal transformer encoder mechanism to improve accuracy and efficiency. The key contributions of the proposed method are summarized as follows:

1. A novel fine-grained, part-aware prompt learning technique is introduced, integrated with diffusion models. This approach enables precise control over individual body parts and generates high-quality outputs, offering substantial improvements for 3D human pose estimation.
2. The RTCPT combines spatiotemporal encoders with a TCP structure to effectively capture both close-range and far-range temporal dependencies, mitigating depth ambiguity and improving the prediction accuracy.
3. The RTCPT transformer incorporates a Feature Aggregation Refinement (FAR) module that improves feature fusion through two TCP attention blocks, enhancing interactions among queries, keys, and values to optimize the feature refinement process.
4. Extensive evaluations with the pose estimation datasets including Human3.6M and MPI-INF-3DHP demonstrate the effectiveness of the proposed method significantly outperforms existing SOTA methods.

2 Related works

2.1 3D human pose estimation

Recently, 3D HPE is a vital research domain in computer vision, dedicated to reconstructing human posture in three-dimensional space using 2D images or video sequences. [25, 26]. Consequently, the HPE methodology has broad applications across various domains, including motion capture, virtual reality, ergonomics, and medical rehabilitation. This task is typically addressed using two primary methodologies: one-stage and two-stage detectors. One-stage approaches directly infer 3D joint coordinates from raw image data through an end-to-end learning framework [27]. In contrast, two-stage methods first estimate 2D keypoints from the input image and subsequently lift them into 3D space by harnessing spatial correlations between the 2D and 3D representations. Recent progress in 2D-to-3D lifting techniques [28–31] has achieved higher accuracy than direct end-to-end approaches, leveraging the robustness of reliable 2D keypoint detection.

For instance, [28] introduced a pose grammar model that encodes human body structures to enhance 3D pose estimation accuracy and robustness. Similarly, MTF-Transformer [29] was proposed to integrate information across multiple viewpoints and time frames using an attention-based mechanism, improving pose predictions under varying conditions. Another approach, CV-UGCNs [30], effectively captures spatial and cross-view dependencies in 3D poses, ensuring structural consistency and enhancing accuracy across different perspectives. Additionally, GraphMLP [31] introduces a novel MLP-based architecture for pose estimation, exploiting graph-based representations to enhance spatial relationships. It achieves competitive accuracy while reducing computational complexity compared to traditional GCN-based methods. Building on these advancements, a Sequential processing framework is adopted for 3D HPE, which consistently outperforms single-stage approaches by leveraging the reliability of well-established 2D pose detection techniques.

2.2 Diffusion model

Recent advancements in 3D human pose estimation have leveraged diffusion models to enhance accuracy and robustness. Specifically, DPoser [32] introduces a diffusion-based human pose prior that improves realism and generalization, benefiting tasks like human mesh recovery and pose

completion. Additionally, FinePOSE [24] integrates fine-grained part-aware prompts and temporal information to refine 3D pose estimation. Furthermore, DDHPose [33] disentangles 3D poses into bone lengths and directions, employing hierarchical spatial and temporal denoisers for better modeling. Moreover, DTCPose [34] introduces a framework that generates multiple 3D pose candidates while enforcing temporal constraints to enhance stability in output sequences. Finally, our recent work MHAFormer [9] leverages a multi-transformer encoder for 3D HPE that significantly improves the performance. It improves accuracy by refining pose predictions through iterative denoising and hypothesis fusion. These studies highlight the potential of diffusion models in addressing ambiguity, occlusion, and temporal consistency in pose estimation.

2.3 Prompt learning

Recent 3D HPE advancements have explored prompt learning techniques to enhance model performance. Regarding this, the author in [35] has introduced an ActionPrompt that effectively mines action-related clues from pose sequences, improving estimation accuracy by incorporating action labels and patterns. However, it primarily focuses on single-person scenarios. To address multi-person contexts, LAMP [36] leveraged language prompts to enhance crowd pose estimation, exploiting text representations to understand poses at both instance and joint levels, thereby improving performance in crowded scenes. Nonetheless, it may struggle with fine-grained body part details. Building upon this, FinePOSE [24] proposed a fine-grained prompt-driven approach, integrating part-aware prompts and temporal information to refine 3D pose estimation. While effective, it requires substantial computational resources due to the diffusion model's complexity. To enhance efficiency, HDFormer [37] introduced a High-order Directed Transformer that models complex joint interactions, reducing parameters and computational costs while maintaining accuracy. These studies collectively demonstrate the evolution of prompt learning in 3D human pose estimation, progressively addressing challenges related to multi-person scenarios, fine-grained details, computational efficiency, and generalization.

2.4 Transformer-based methods

Recent advances in HPE have utilized transformer-based methods to address challenges in modeling global joint dependencies, handling occlusions, and improving temporal consistency in dynamic sequences [38]. In article [39], HSTFormer introduced hierarchical spatial-temporal transformers, effectively capturing multi-level joint correlations, thereby improving pose estimation accuracy. However, it may face challenges in handling complex, occlusion-heavy scenarios. To address this, HDFormer [37] proposed high-order bone and joint relationships through a multi-order attention module, enhancing performance in complex situations. Despite this, it may not fully leverage external knowledge sources. To mitigate this limitation, EVOPOSE [40] integrated human body priors using a structural priors representation module, effectively incorporating kinematic structure priors into the transformer framework. However, it may require substantial computational resources due to the complexity of integrating these priors. To enhance efficiency, ConvFormer [41] introduced a dynamic multi-headed convolutional attention mechanism, reducing parameters and computational costs while maintaining accuracy. Nonetheless, it may not fully capture long-range dependencies inherent in human pose sequences. To address this, the MLTFFPN [11] method addresses the challenge of capturing long-range dependencies by incorporating a multi-level transformer architecture combined with a feature frame padding network. This allows it to effectively maintain spatial and temporal consistency across varying input sequence lengths, ensuring better handling of long-range dependencies in 3D HPE. Building on the insights from the previous analysis, we leveraged the Transformer method to enhance the modeling of complex dependencies and improve the accuracy of pose estimation.

3 Proposed method

3.1 Diffusion pipeline for 3D pose estimation

The diffusion model [42, 43] consists of two primary components: the diffusion process and the reverse process. During the forward diffusion phase, Gaussian noise is incrementally added to the ground-truth 3D poses, transforming them into noisy representations. The reverse process utilizes a denoiser to reconstruct structured 3D poses from noisy samples. The overall diagram is illustrated in Fig. 1

Diffusion Process: The diffusion process, denoted by $\mathcal{Q}$, generates noisy samples $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_T$ by progressively adding Gaussian noise $\Psi \sim \mathcal{N}(0, I)$ to the original ground-truth 3D pose $\mathcal{Y}_0$ over a series of time steps $t \in [0, \mathcal{T}]$, where $\mathcal{T}$ indicates the Upper bound of time step.. The mathematical calculation of diffusion process $\mathcal{Q}$ is defined as:

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

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

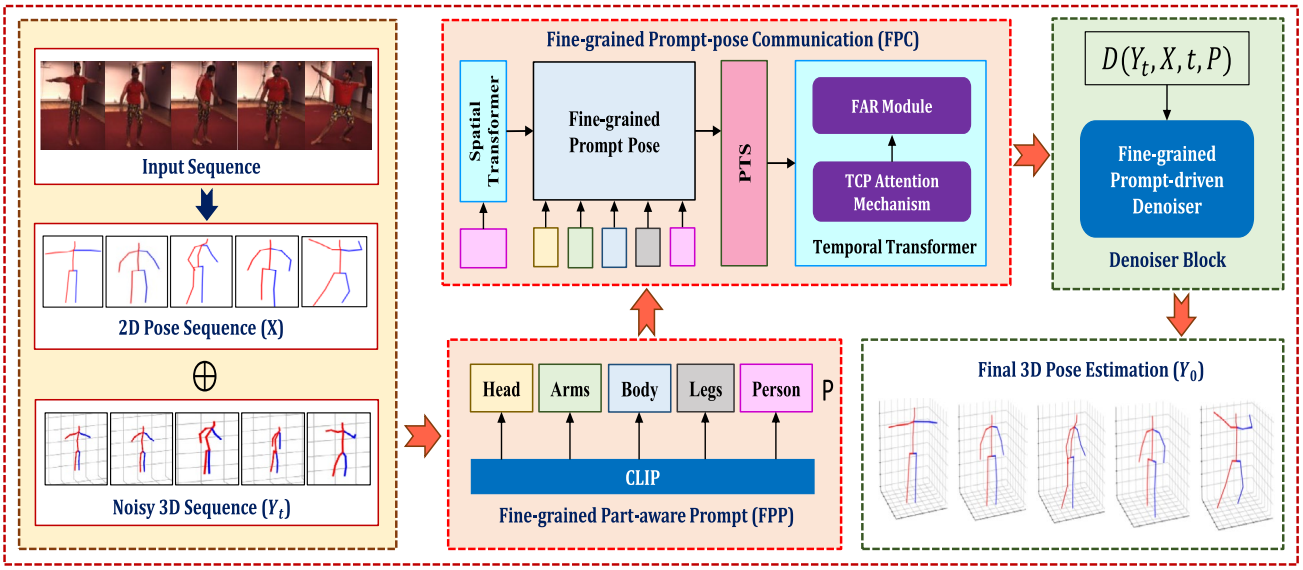

Fig. 1 Overview of the FGPDFormer architecture

Here, Υ_t represents the cosine noise variance, and I is the identity matrix. Using the principles of DDPM [44], the noisy sample $\mathcal{Y}_t$ can be directly obtained from $\mathcal{Y}_0$ without performing iterative steps. This is expressed as:

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

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

Reverse Process: The reverse process reconstructs the denoised 3D pose $\widetilde{\mathcal{Y}}_0$ from the noisy input $\mathcal{Y}_t$. This is achieved through a learned transformer-based denoiser network $\mathcal{D}$, which predicts the refined 3D pose using the following formulation:

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

$$\mathcal{L} = \mathbb{E}_{\mathcal{Y}_t, t}\left[\mathcal{Y}_0 - \widetilde{\mathcal{Y}}_{0_2}^{\,2}\right], \tag{5}$$

Here, $\mathcal{X}$ represents auxiliary inputs (e.g., 2D keypoints or other contextual information), while $\mathcal{L}$ denotes the mean squared error loss function used to minimize the difference between the reconstructed 3D pose $\widetilde{\mathcal{Y}}_0$ and the ground-truth pose $\mathcal{Y}_0$.

3.2 Fine-grained part-aware prompt learning (FPP)

Our proposed method improve the denosing by incorporating 2D keypoints $\mathcal{X}$, timestamp t, and fine-grained part-aware prompt embeddings P. The FPP block generates P by encoding action class data, along with the kinematic records into the prompt embedding space. The learnable prompt embedding $P = \{p_k\}_{k=1}^{K}$ has shape $K \times L \times D$, where K is the number of prompts, L is the token count per prompt, and D is the embedding dimension. Each prompt p_k is formulated as:

$$p_k = \mathrm{Concat}(r_k, \tilde{p}_k), \quad \tilde{p}_k = E_{\mathrm{tx}}(\mathrm{text}_k)[:4], \tag{6}$$

where $r_k \in \mathbb{R}^{(L_k-4)\times D}$ is a learnable vector initialized with $\mathcal{N}(0, 0.02)$, and E_{tx} is the text encoder. The FPP block ensures precise guidance for denoising by generating multi-granularity prompts.

3.3 Fine-grained prompt-pose communication (FPC)

The FPC structure develops a detailed communication within noisy 3D poses $\mathcal{Y}_t$ and the learned prompts P. This is achieved by integrating $\mathcal{Y}_t$, $\mathcal{X}$, t, and P into Z_t:

$$Z_t = \mathrm{Concat}(\mathcal{Y}_t, \mathcal{X}) + P[L] + F(t), \tag{7}$$

where $F(t)$ is the timestamp. This embedding ensures that timestamp-specific noise characteristics are appropriately modeled. The integrated features Z_t are processed by a spatial transformer to extract intra-frame relationships among the joints, yielding Z_t^s:

$$Z_t^s = \mathrm{SpatialTransformer}(Z_t). \tag{8}$$

Next, the fine-grained relationship between the learned prompt embeddings P and the spatially transformed pose features Z_t^s is captured using a multi-head cross-attention mechanism. The weight matrices for query, key, and value are represented by:

$$Q = W_Q Z_t^s, \quad K = W_K P, \quad V = W_V P, \tag{9}$$

where W_Q, W_K, and W_V are learnable projection matrices. The cross-attention mechanism computes:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{10}$$

$$Z_t^{sp} = A \otimes V, \tag{11}$$

$$\widetilde{Z}_t^{sp} = \mathcal{P}(Z_t^{sp}), \tag{12}$$

where d is the dimensionality of each attention head. This process integrates the fine-grained prompt information into the pose features. $\mathcal{P}$ represents the PTS block, which incorporates the timestamp t and produces the timestamp-stylized output $\widetilde{Z}_t^{sp}$. The resulting features $\widetilde{Z}_t^{sp}$ are further refined by temporal transformers to model inter-frame dependencies, producing $\widetilde{Z}_t^{spf}$:

$$\widetilde{Z}_t^{spf} = \text{TemporalTransformer}(\widetilde{Z}_t^{sp}). \tag{13}$$

Finally, a spatio-temporal transformer produces the fine-grained prompt-driven feasible attributes, which are decoded to predict the denoised 3D pose sequence $\hat{\mathcal{Y}}_0$.

3.4 Prompt-driven timestamp stylization (PTS)

The Prompt-driven Timestamp Stylization (PTS) block is crucial for adapting the denoising process to varying noise levels across timestamps. It explicitly incorporates timestamp embeddings into the refinement of pose features, enabling effective denoising at each step.

First, the timestamp embedding $F(t)$ is calculated using a sinusoidal function. This embedding captures the temporal information corresponding to the current noise level. The PTS block [45] integrates this embedding with the learnable prompt embeddings P generated by the FPP block. The combined representation is defined as:

$$v = P[L] + F(t), \tag{14}$$

where $P[L]$ represents the learnable prompt embedding corresponding to the timestamp.

Given the intermediate pose features Z_t^{sp} from the FPC block, the PTS block refines these features by introducing a timestamp-aware modulation. This is achieved using a linear transformation of v:

$$\tilde{Z}_t^{sp} = Z_t^{sp} \odot \psi_w(\phi(v)) + \psi_b(\phi(v)), \tag{15}$$

where ϕ, ψ_w, and ψ_b are linear projection functions, and $\odot$ denotes the Hadamard (element-wise) product. Here, $\phi(v)$ transforms v into a latent representation, while ψ_w and ψ_b scale and shift the intermediate features Z_t^{sp}, respectively.

This stylization mechanism ensures that the denoising process adapts to the noise characteristics at each timestamp, enhancing the quality of the reconstructed 3D poses $\hat{\mathcal{Y}}_0$. By incorporating both temporal and prompt-driven guidance, the PTS block enables FGPDFormer to effectively handle diverse noise conditions during the denoising process.

3.5 Training and inference

3.5.1 Training

During the training phase, our FGPDFormer is optimized to reconstruct clean 3D pose sequences from noisy inputs. For each input, the contaminated 3D pose sequence $\mathcal{Y}_t$ is processed by the denoiser $\mathcal{D}$, which predicts the reconstructed pose sequence $\hat{\mathcal{Y}}_0$:

$$\hat{\mathcal{Y}}_0 = \mathcal{D}(\mathcal{Y}_t, \mathcal{X}, t, P), \tag{16}$$

where $\mathcal{X}$ represents the corresponding 2D keypoints, t is the timestamp, and P denotes the fine-grained prompt embeddings generated by the FPP block. The denoising model $\mathcal{D}$ is trained by minimizing the mean squared error (MSE) loss between the predicted 3D poses $\hat{\mathcal{Y}}_0$ and the ground truth 3D poses $\mathcal{Y}_0$:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{Y}_0^{(i)} - \hat{\mathcal{Y}}_0^{(i)}\|_2^2, \tag{17}$$

where N is the number of samples in the training batch. This objective ensures that the denoiser learns to progressively refine the noisy poses toward the ground truth.

3.5.2 Inference

At inference, the denoising process begins by sampling an initial noisy 3D pose sequence $\mathcal{Y}_T$ from a standard Gaussian distribution $\mathcal{N}(0, I)$. This noisy sequence is iteratively refined over M timesteps using the denoiser $\mathcal{D}$:

$$\mathcal{Y}_{t-1} = \mathcal{D}(\mathcal{Y}_t, \mathcal{X}, t, P), \quad t = T, T-1, \ldots, 1. \tag{18}$$

To improve prediction accuracy, FGPDFormer employs a multi-hypothesis strategy. A set of H initial noisy pose sequences $\{\mathcal{Y}_T^h\}_{h=1}^H$ is generated, and the denoising process is applied independently to each hypothesis. After M steps, the final denoised hypotheses $\{\hat{\mathcal{Y}}_0^h\}_{h=1}^H$ are obtained.

The final 3D pose prediction $\hat{\mathcal{Y}}_0$ is determined by Joint-Wise Reprojection-Based Multi-Hypothesis Aggregation (JPMA). Each joint in the final prediction is selected from the hypothesis that minimizes the reprojection error against the input 2D keypoints $\mathcal{X}$:

$$\hat{h} = \arg\min_h \|\mathrm{P}_R(\hat{\mathcal{Y}}_0^h)[j] - \mathcal{X}[j]\|_2^2, \quad \hat{\mathcal{Y}}_0[j] = \hat{\mathcal{Y}}_0^{\hat{h}}[j], \tag{19}$$

where P_R is the reprojection function, and j indexes the joints. This aggregation strategy ensures that the final prediction leverages the best components of multiple hypotheses, resulting in a highly accurate 3D pose estimate.

3.6 Multi-human 3D pose estimation extension

Our FGPDFormer is extended to multi-human 3D pose estimation by incorporating a post-processing integration step that enables handling multiple individuals in a given scene. Given a multi-person 2D keypoints sequence $\mathcal{X}^{\mathrm{mul}} \in \mathbb{R}^{C\times N\times J\times 2}$, where C is the number of detected individuals, FGPDFormer processes each person separately to predict their respective 3D poses $\hat{\mathcal{Y}}_0^c$ for $c \in \{1, 2, ..., C\}$. These individual predictions are temporally synchronized and spatially integrated to ensure coherent representation of multi-human interactions. To maintain temporal consistency, FGPDFormer employs a frame-wise association strategy, ensuring that individuals are correctly tracked across consecutive frames. If a person temporarily leaves the camera's field of view, their respective pose is set to zero, preserving the sequence alignment. Finally, the estimated 3D poses for all individuals are stacked along the character dimension to generate the final multi-human pose representation $\hat{\mathcal{Y}}_0^C \in \mathbb{R}^{C\times N\times J\times 3}$, effectively capturing interactions in complex multi-person scenarios.

3.7 TCP attention module

To capture a more comprehensive representation of temporal features, we propose the TCP Attention Module. This module enhances multi-scale information extraction by dynamically adjusting the temporal granularity of keys $K \in \mathbb{R}^{n\times d}$ and values $V \in \mathbb{R}^{n\times d}$, while maintaining the query dimensionality $Q \in \mathbb{R}^{n\times d}$. The TCP block progressively refines the length of the time series of K and V by iteratively constricting and then proliferating temporal features across multiple resolution levels.

As depicted in Fig. 2 the TCP module initially shrinks the temporal span by downsampling K and V with a sampling ratio r, followed by a proliferation stage that reconstructs fine-grained details. This dual-stage transformation enables the model to retain both high-level semantic information and fine-grained temporal dependencies, ultimately

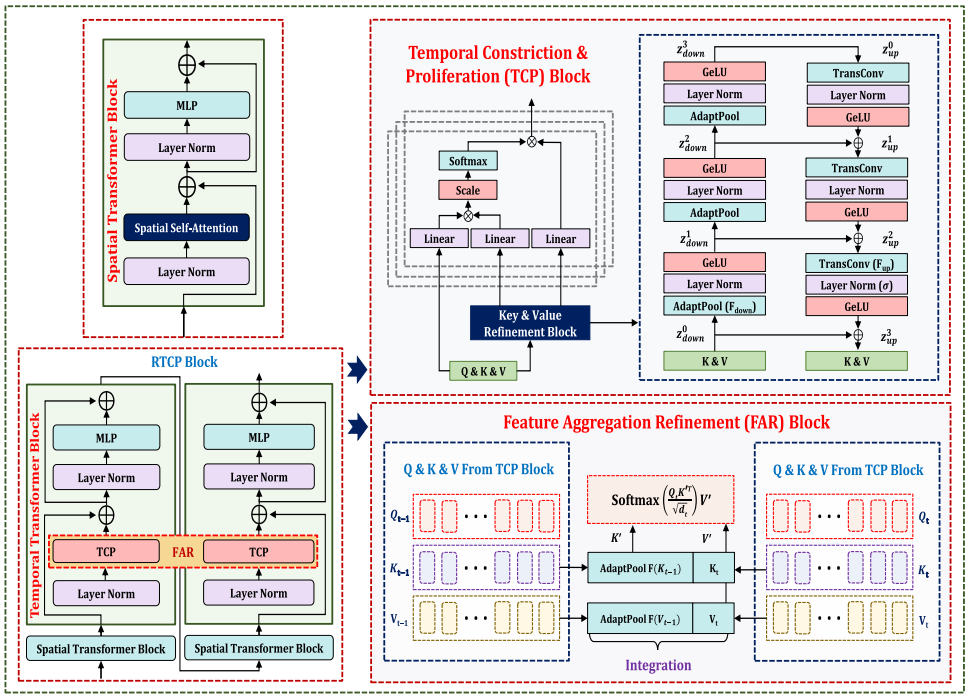

Fig. 2 The overall structure of spatio-temporal transformers with feature aggregation module

improving the model's capacity to capture motion dynamics accurately.

3.8 TCP attention block for key and value refinement

The TCP attention block is essential for minimizing redundant information while retaining both local and global semantic features. By refining keys and values through a multi-scale attention mechanism, the model is able to improve representation learning for sequential data.

For an input feature sequence $z \in \mathbb{R}^{n \times d}$ where n is the sequence length and feature dimension d is the feature dimension, and the TCP attention mechanism iteratively processes the features in a hierarchical manner. The transformation can be formulated as:

$$\begin{aligned} z^0_{\text{down}} &= z, \quad z^{l+1}_{\text{down}} = \sigma\left(\text{LN}\left(\mathcal{F}_{\text{down}}\left(z^l_{\text{down}}\right)\right)\right), \\ z^0_{\text{up}} &= z^m_{\text{down}}, \quad z^{l+1}_{\text{up}} = \sigma\left(\text{LN}\left(\mathcal{F}_{\text{up}}\left(z^l_{\text{up}}\right)\right)\right) + z^{m-1-l}_{\text{down}}, \end{aligned} \tag{20}$$

where $\sigma(\cdot)$ represents an activation function, LN denotes Layer Normalization, and F_{down} and F_{up} correspond to the constriction and proliferation functions. The downsampling process at each stage reduces the sequence length by a factor of r^l, while the upsampling process reconstructs detailed temporal features with a restoration factor r^{m-l}. The final output is given by:

$$\text{TCP}(z) = z^{\text{up}}_m \in \mathbb{R}^{n \times d}. \tag{21}$$

This process ensures that the model not only distills essential motion patterns by removing noise during constriction but also enhances the expressive power by selectively restoring critical temporal cues during proliferation. By refining the keys and values, the TCP block effectively mitigates noise sensitivity and strengthens the model's competence in preserving extended-range interactions.

3.9 FAR module

The FAR module is crafted to refine and seamlessly integrate features across multiple temporal layers, enhancing the robustness of 3D human pose estimation. While convolutional networks have demonstrated the benefits of hierarchical feature fusion, the transformer-based approach lacks explicit aggregation mechanisms across different layers. To address this, we propose a cross-layer feature aggregation mechanism that improves temporal consistency and representation quality. The FAR module operates by integrating information from neighboring transformer layers, leveraging interactions between queries, keys, and values. The aggregation function is defined as follows:

$$\begin{aligned} \text{FAR} &= \text{Attn}\left(Z_{t-1}, Z_t\right), \\ &= \text{Attn}\left(Q_t, K', V'\right), \\ &= \text{Softmax}\left(\frac{Q_t K'^{\top}}{\sqrt{d}}\right) V', \\ K' &= \text{Concat}\left(K_t, \mathcal{F}\left(K_{t-1}\right)\right), \\ V' &= \text{Concat}\left(V_t, \mathcal{F}\left(V_{t-1}\right)\right). \end{aligned} \tag{22}$$

Here, Z_{t-1} and Z_t represent features from adjacent layers, and the attention mechanism effectively fuses temporal information to improve continuity. The function $F(\cdot)$ represents an adaptive pooling operation that extracts relevant details from previous layers. The use of cross-layer attention ensures that each frame benefits from prior contextual information, leading to more stable and accurate pose predictions. By introducing FAR, the model can dynamically aggregate multi-scale information and reinforce temporal coherence, making it more resilient to variations in motion patterns and occlusions. This enhancement significantly contributes to improved performance in 3D HPE tasks.

4 Experimental results and analysis

4.1 Datasets and evaluation criteria

To rigorously assess the performance of the proposed method, we utilize two widely recognized benchmark datasets: Human3.6M and MPI-INF-3DHP. These datasets provide comprehensive test scenarios for evaluating 3D human pose estimation algorithms.

Human3.6M Dataset: This is the most extensive dataset for 3D human pose estimation, comprising roughly 3.6 million images. The dataset was collected using four synchronized cameras operating at 50 Hz and features professional subjects performing 15 daily activities, including "Waiting," "Smoking," and "Posing." We follow the standard experimental setup used in previous studies, where subjects S1, S5, S6, S7, and S8 are used for training, while subjects S9 and S11 are designated for evaluation.

MPI-INF-3DHP Dataset: This dataset captures diverse human motions in either constrained or unconstrained environments, providing a more challenging setting for 3D human pose estimation compared to Human3.6M. It contains 1.3 million frames collected from multiple camera viewpoints, ensuring a robust evaluation framework.

To evaluate the accuracy of our approach, we employ the standard Mean Per Joint Position Error (MPJPE), commonly known as Protocol-1. MPJPE calculates the average Euclidean distance between predicted joint positions and ground truth, offering a direct measure of 3D pose

estimation accuracy. This metric ensures a comprehensive and fair assessment of our proposed method.

4.2 Implementation setup

Our proposed method, FGPDFormer is implemented using the PyTorch deep learning framework, with all experiments conducted on an NVIDIA GeForce RTX 4090 GPU. The model processes 2D keypoints extracted from a state-of-the-art 2D pose detector as input. We optimize training using the Adam optimizer, set a batch size of 1024, a dropout rate of 0.1, and employ the GELU activation function. The sequence length for Human3.6M and MPI-INF-3DHP is set to 243 and 27 respectively. At the training stage, the number of hypotheses (H) and iterations (K) start at 1 and increase to H=20 and K=10 in the inference phase to enhance accuracy. The model leverages a diffusion-based inference mechanism with a maximum timestep value of 1000, ensuring robustness in pose prediction. These configurations balance computational efficiency and predictive accuracy, improving generalization across datasets and scenarios.

4.3 Results on Human3.6M

We evaluate the effectiveness of our proposed approach on the Human3.6M dataset by comparing it with various state-of-the-art methods, as presented in Table 1. The results highlight the performance of different approaches using MPJPE and P-MPJPE metrics. Additionally, Fig. 3 illustrates the qualitative evaluation of the proposed algorithm on the Human3.6M dataset where the test sequence S9 (Posing) is considered. For training, we utilize 2D pose detections from the widely adopted CPN [76] as well as ground truth data. Our FGPDFormer model demonstrates superior accuracy, achieving MPJPE score of 41.1%, respectively. Simultaneously, we evaluate the presented approach with PoseFormer, HDFormer, MixSTE, STRFormer, and HSTFormer, the developed methodology exhibits MPJPE enhancements of (4.2%, 2.5%, 0.8%, 0.9%, and 2.6%) under protocol-1, showcasing its effectiveness in 3D human pose estimation.

Table 1 (Bottom) presents a comparative analysis of our proposed method against other probabilistic approaches under four distinct experimental conditions. **P-Agg** represents a straightforward pose-level aggregation technique where the 3D positions of each joint are aggregated by taking the average overall pose conjectures. In contrast, our **J-Agg** method employs a joint-level aggregation strategy to refine predictions by aggregating results at the joint level, thereby enhancing 3D pose accuracy. **P-Best** selects the maximum cumulative rating, identifying the 3D pose that most closely corresponds to the reference data. Meanwhile, our **J-Best** approach optimally selects joints from different hypotheses based on proximity to the ground truth, integrating them to generate the final 3D pose.

Furthermore, under the P-Agg evaluation setting with $H = 20$, our approach achieves notable improvements over transformer-based methods, outperforming MHFormer by 3.0mm, MHFormer++ by 2.5mm, EMHIFormer $N = 81$ by 4.1mm, and EMHIFormer $N = 351$ by 2.8mm. However, when increasing the hypothesis count from $H = 1$ to $H = 20$, the proposed method shows minimal variation in performance under both P-Agg $40.1mm \rightarrow 40.0mm$ and P-Best $40.2mm \rightarrow 39.7mm$ settings. This finding motivated us to introduce a new evaluation approach, J-Best, which significantly enhances performance $40.2mm \rightarrow 39.0mm$. The J-Best setting achieves the most accurate results by evaluating specific joints within the same hypothesis. Inspired by this improvement, we further propose J-Agg $H = 20$, which leverages joint-level variations across hypotheses. Unlike P-Agg, J-Agg enhances efficiency, leading to a measurable improvement from 40.0*mm* to 39.0*mm*. In addition, we provide qualitative feature visualizations for the right shoulder (top row) and right elbow (bottom row), comparing the ground-truth heatmaps with those produced by the FAR, TCP, and combined FAR+TCP modules as illustrated in Fig. 4. These visual results highlight how each module progressively enhances feature localization. The integrated FAR+TCP approach yields heatmaps that align more closely with the ground truth, demonstrating their complementary effectiveness.

4.4 Results on MPI-INF-3DHP

The proposed approach is compared with several state-of-the-art methods using three standard metrics: Percentage of Correct Keypoints (PCK), Area Under the Curve (AUC), and MPJPE. Table 2, shows that our developed framework, obtains the highest efficiency, recording a PCK of 98.7%, an AUC of 78.0%, and an MPJPE of 27.2 mm. In comparison to conventional approaches, FGPDFormer significantly outperforms MHFormer (PCK: 93.8%, AUC: 63.3%, MPJPE: 58.0 mm), achieving absolute improvements of 4.9%, 14.7%, and 30.8 mm, respectively. Furthermore, it demonstrates superior accuracy over MixSTE (PCK: 96.9%, AUC: 75.8%, MPJPE: 35.4 mm), surpassing it by 1.8%, 2.2%, and 8.2 mm across the three metrics. These results highlight the robustness and effectiveness of FGPDFormer in advancing 3D human pose estimation. Our method also improves upon MHFormer++ (PCK: 94.8%, AUC: 65.8%, MPJPE: 54.0 mm) by 3.9%, 12.2%, and 26.8 mm, and STRFormer (PCK: 94.8%, AUC: 67.1%, MPJPE: 54.4 mm) by 3.9%, 10.9%, and 27.2 mm, respectively. Furthermore, FGPDFormer surpasses HSTFormer (PCK: 97.3%, AUC: 71.5%, MPJPE: 41.4 mm) with improvements of 1.4%, 6.5%, and

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 [48] (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 [49] (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 [50] (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 [51] (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 [52] (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 [53] (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 [54] (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 [55] (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 [56] (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 [57]	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 [19] (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 [58] (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 [59](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 [37] (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 [39] (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 [60] (-)	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[61] (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[62] (N=27)	CVPR'23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
DAF-DG[63]	CVPR'24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44.4
GLA-GCN [64] (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
NanoHTNet [65]	TIP'25	42.5	47.7	44.3	47.4	49.3	55.9	46.0	44.2	55.0	61.8	47.8	45.0	50.6	38.1	39.2	47.6
PGFormer [66]	TMM'25	43.8	49.9	45.5	49.0	51.0	58.2	47.0	45.7	58.1	65.8	49.6	46.5	53.6	38.0	40.6	49.5
GraphMLP [67]	PR'25	43.7	49.3	45.5	47.9	50.5	56.0	46.3	44.1	55.9	59.0	48.4	45.7	51.2	37.1	39.1	48.0
DGFormer [68]	PR'25	45.8	49.6	46.2	49.6	51.4	58.7	48.9	46.2	56.6	65.1	50.9	47.2	53.2	38.8	41.5	50.0
FGPDFormer (N=243, K=1, H=1)		*37.7*	***40.0***	***37.0***	***38.5***	***41.1***	***47.4***	*39.4*	*39.2*	*50.2*	***54.2***	***41.9***	***39.5***	***40.1***	***26.9***	***27.1***	***40.1***
	Probabilistic Methods																
Protocal #1		**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
CVAE [69] (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 [70] (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 [71] (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 [72] (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 [42] (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 [43] (N=1, H=10, P-Agg)	ICASSP'24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	41.8
MHFormer [10] (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

Table 1 (continued)

	Probabilistic Methods															
Protocal #1		**Dir.**	**Disc.**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Pur.**	**Sit**	**SitD.**	**Smoke**	**Wait**	**WalkD.**	**Walk**	**WalkT.**	**Avg.**
MHFormer++ [73] (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 [74] (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 [74] (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
FGPDFormer (N=243, K=1, H=1, 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	40.1
FGPDFormer (N=243, K=5, H=20, P-Agg)		*38.4*	*40.2*	*36.0*	*38.1*	*41.1*	*46.9*	*39.3*	*39.0*	*50.1*	*54.0*	*41.4*	*39.2*	***39.2***	*27.4*	*27.1*	*40.0*
FGPDFormer (N=243, K=5, H=20, J-Agg)		***38.1***	***40.1***	***35.7***	***37.6***	***41.0***	***46.3***	***39.1***	***38.1***	***50.0***	***53.0***	***40.9***	***39.0***	*39.5*	***27.3***	***27.0***	***39.7***
CVAE [69] (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 [75] (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 [70] (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 [71] (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 [72] (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
FGPDFormer (N=243, K=1, H=1, P-Best)		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
FGPDFormer (N=243, K=10, H=20, P-Best)		38.5	*38.7*	*34.6*	*37.4*	41.6	*47.1*	*40.0*	*38.4*	51.3	*53.4*	*41.8*	*39.7*	*38.5*	27.5	*26.9*	*39.7*
FGPDFormer (N=243, K=10, H=20, J-Best)		***37.6***	***37.7***	***33.9***	***36.6***	***40.9***	***46.4***	***39.4***	***37.7***	*50.6*	***52.5***	***41.1***	***39.0***	***37.7***	***26.9***	***26.3***	***39.0***

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

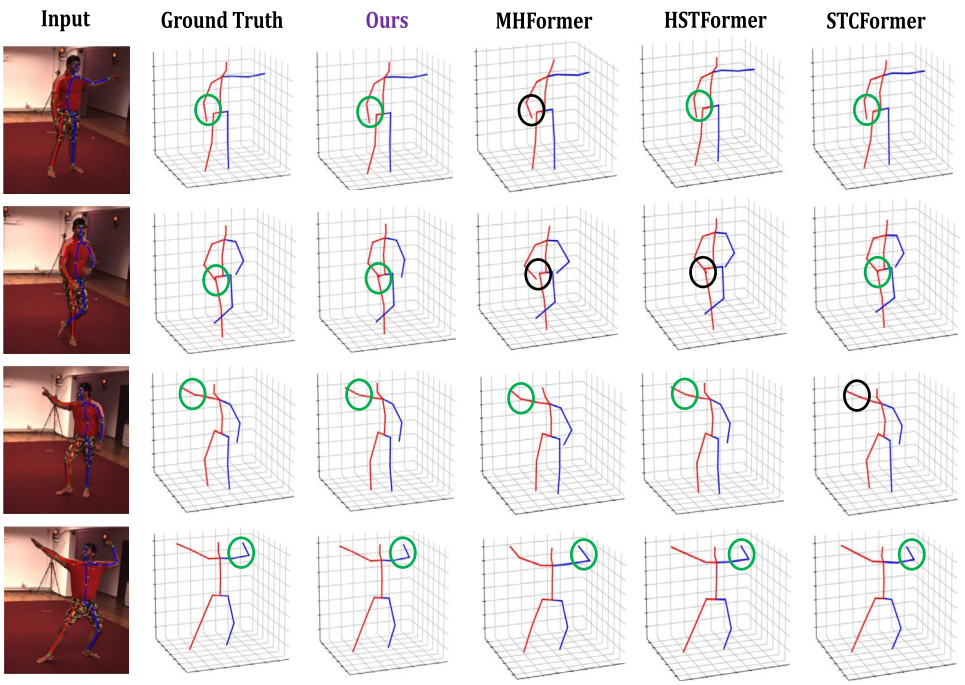

Fig. 3 A qualitative evaluation on the Human3.6M dataset. The green circle represents the correctly predicted poses, whereas the black circle denotes the incorrectly predicted poses

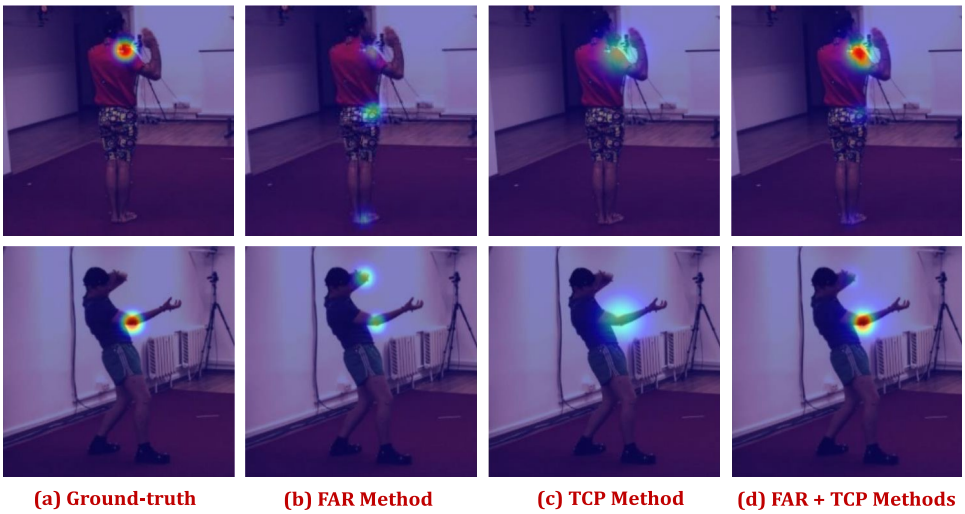

Fig. 4 Heatmap-based feature visualizations for the right shoulder (top) and right elbow (bottom) across four configurations on the Human3.6M dataset

14.2 mm. These results highlight that FGPDFormer consistently outperforms conventional transformer-based methods, achieving the highest accuracy and lowest error in 3D HPE, thereby setting a new benchmark on the MPI-INF-3DHP dataset.

4.5 Component analysis

As shown in Table 3, the stepwise integration of various components in the developed method demonstrates a clear improvement in MPJPE error reduction. Initially, using only FPP (Fine-grained Part-aware Prompt), the MPJPE remains relatively high at 42.4. Adding FPC (Fine-grained Prompt-Pose Communication) further refines pose estimation, reducing the MPJPE to 42.0. By employing only the TCP and FAR modules with 1 and 20 hypotheses, the MPJPE values gradually improved to 41.8 and 41.4, respectively. When PTS (Prompt-driven Timestamp Stylization) is incorporated alongside FPP and FPC, the MPJPE improves further to 41.2, showing the effectiveness of temporal information stylization. Introducing Multi-Level TCP, which enhances the temporal relationship modeling, leads to a further reduction, bringing the MPJPE down to 40.5. Finally, the inclusion of FAR (Feature Aggregation Refinement) optimizes the feature aggregation process, refining spatial-temporal relationships and improving the MPJPE significantly. With all components in place and using 20 hypotheses, we compare our proposed methods: P-Agg, P-Best, J-Agg, and J-Best. The P-Agg setup achieves an MPJPE of 40.0, while the P-Best and J-Agg setups both

Table 2 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 [50] (N=81)	TCSVT'21	87.9	54.0	78.8
PoseFormer [20] (N=9, H=3, P-Agg)	ICCV'21	88.6	56.4	77.1
U-CDGCN [51] (N=96)	MM'21	97.9	69.5	42.5
MHFormer [10] (N=27)	CVPR'22	93.8	63.3	58.0
P-STMO [55] (N=81)	ECCV'22	97.9	75.8	32.2
MixSTE [8] (N=243)	CVPR'22	96.9	75.8	35.4
MHFormer++ [73] (N=9)	PR'23	94.8	65.8	54.0
Uplift & Upsample [58] (N=81)	WACV'23	97.9	75.8	32.2
STRFormer [59] (N=27)	IMAVIS'23	94.8	67.1	54.4
HDFormer [37] (N=32)	arXiv'23	96.8	64.0	51.5
EMHIFormer [74] (N=9)	JVCIR'23	97.1	74.9	33.8
HSTFormer [39] (N=81)	arXiv'23	97.3	71.5	41.4
JoyPose [60]	PR'24	94.1	-	-
DAF-DG [63]	CVPR'24	92.9	60.7	63.1
GLA-GCN [64] (N=27)	ICCV'23	*98.1*	*76.5*	*31.3*
NanoHTNet [65]	TIP'25	86.7	56.1	-
PGFormer [66]	TMM'25	83.9	52.3	-
GraphMLP [67]	PR'25	87.0	54.3	-
DGFormer [68]	PR'25	84.4	52.5	83.9
FGPDFormer (N=81)		***98.7***	***78.0***	***27.2***

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

reach 39.7, showing further improvements. It can be seen that better results are observed with the J-Best setup, which achieves the lowest MPJPE of 39.0, highlighting the competence of our full proposed framework in optimizing 3D human pose estimation.

4.6 Case study analysis

To further evaluate the effectiveness of our proposed method, we conducted a detailed case study analyzing joint-wise 3D pose estimation errors on the Human3.6M dataset, comparing our approach against PoseFormer and MHFormer++ as illustrated in Fig. 5. This analysis provides insight into both overall performance and improvements at challenging body joints. The results indicate that our method consistently achieves lower MPJPE across all major body joints. For instance, the right knee, a relatively stable joint, records an error of 37.3 mm with our method, compared to 40.6 mm (PoseFormer) and 39.5 mm (MHFormer++). The right wrist, which typically exhibits high error due to its greater range of motion and frequent occlusions, demonstrates a significant reduction to 67.5 mm, improving upon PoseFormer (75.5 mm) and MHFormer++ (69.6 mm). Similarly, the neck and head joints achieve MPJPEs of 36.3 mm and 36.5 mm, respectively, outperforming PoseFormer (41.3 mm and 45.1 mm) and MHFormer++ (38.1 mm and 38.9 mm).

Other challenging joints also show notable improvements with our method, demonstrating its ability to handle complex body configurations and refine predictions in difficult pose regions. Specifically, the left wrist achieves 63.6 mm, reducing errors compared to PoseFormer (73.6 mm) and MHFormer++ (66.5 mm). The right shoulder and right elbow record MPJPEs of 37.3 mm and 52.6 mm, respectively, again lower than baseline methods (PoseFormer: 41.9 mm and 60.4 mm; MHFormer++: 39.6 mm and 56.8 mm). These results demonstrate that our approach is robust across joints with both low and high motion variability. At the overall level, our proposed method achieves the lowest average MPJPE of 40.1 mm, compared to 42.5 mm for MHFormer++ and 44.3 mm for PoseFormer. This highlights the consistent advantage of our approach in improving 3D human pose estimation accuracy.

Two major factors contribute to these improvements: the integration of fine-grained textual prompts and the diffusion-based refinement mechanism. By encoding descriptive cues about joint motion and spatial relationships, textual prompts guide the network to better understand contextual dependencies among joints, particularly in complex or occluded poses. This enables the model to infer joint positions more accurately, especially for joints that are prone

Table 3 Ablation analysis on various proposed elements is conducted on the Human3.6M dataset using the MPJPE metric

FPP	FPC	PTS	Multi-Level TCP	FAR	Hypothesis	Setup	MPJPE
✓	✗	✗	✗	✗	1	Not Applicable	42.4
✓	✓	✗	✗	✗	1	Not Applicable	42.0
✗	✗	✗	✓	✓	1	Not Applicable	41.8
✗	✗	✗	✓	✓	20	Not Applicable	41.4
✓	✓	✓	✗	✗	1	Not Applicable	41.2
✓	✓	✓	✓	✗	1	Not Applicable	40.5
✓	✓	✓	✓	✓	20	P-Agg	40.0
✓	✓	✓	✓	✓	20	P-Best	*39.7*
✓	✓	✓	✓	✓	20	J-Agg	*39.7*
✓	✓	✓	✓	✓	20	J-Best	***39.0***

The results highlighted in bolditalic and italic indicate the best and second-best outcomes, respectively

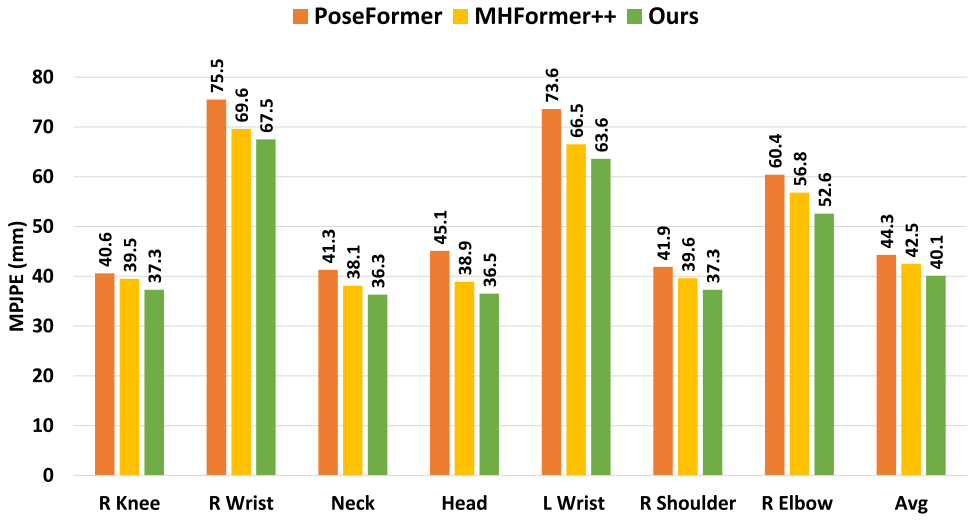

Fig. 5 Comparison of average joint error (MPJPE) with state-of-the-art results on the Human3.6M dataset

Table 4 Comparative analysis of parameters, FLOPs, FPS, and MPJPE for different models evaluated under Protocol-1 on the Human3.6M dataset

Methods	Parameter ↓	FLOPs ↓	FPS	MPJPE ↓
VideoPose3D [46] (N=243)	16.9M	33 M	863	46.8
PoseFormer [20] (N=81)	9.6M	1358 M	269	44.3
Anatomy [50] (N=243)	58.1M	656 M	264	44.1
P-STMO [55] (N=243)	6.2M	1350 M	3040	42.8
FGPDFormer (Ours) (N=243)	5.9M	625 M	3380	39.0

to high variability, such as wrists and elbows. In addition, the incorporation of the diffusion-based refinement process further enhances prediction stability and precision. The diffusion module iteratively denoises intermediate pose representations, allowing the model to correct local inconsistencies and progressively converge toward anatomically plausible joint configurations. This is especially beneficial for correcting subtle errors in small or fast-moving joints, where frame-to-frame variations often introduce noise into the estimation process. Together, the fine-grained textual prompts and the diffusion mechanism improve both interpretability and robustness. The prompts provide semantic guidance on joint behavior, while the diffusion process enforces structural consistency, resulting in more accurate and reliable 3D human pose estimation across diverse motion patterns.

4.7 Parameter analysis

As presented in Table 4, our proposed FGPDFormer demonstrates outstanding performance and computational efficiency compared to existing state-of-the-art methods on the Human3.6M dataset under Protocol-1. Specifically, FGPDFormer achieves the lowest MPJPE of 39.0 mm, outperforming previous models such as VideoPose3D (46.8 mm), PoseFormer (44.3 mm), Anatomy (44.1 mm), and P-STMO (42.8 mm). This substantial reduction in error highlights the superior pose estimation capability of our framework. Moreover, FGPDFormer attains this accuracy with only 5.9M parameters, which is considerably fewer than all compared methods–nearly 65% fewer parameters than PoseFormer and an order of magnitude smaller than Anatomy. This indicates that the proposed model is highly lightweight and memory-efficient, making it well-suited for deployment on low-resource or embedded systems. In terms of computational complexity, FGPDFormer records 625 M FLOPs, which is significantly lower than the transformer-based counterparts PoseFormer (1358 M) and P-STMO (1350 M). The reduced computational load directly translates to faster inference, achieving an impressive 3380 FPS, the highest among all compared methods. This efficiency demonstrates the effectiveness of the Feature-Guided Pose Diffusion (FGPD) module in refining pose features without incurring additional computational overhead. Overall, the experimental analysis confirms that FGPDFormer achieves the best balance between model size, computational cost, and prediction accuracy. The framework not only produces highly precise 3D pose estimations but also exhibits remarkable efficiency, validating its potential for real-time 3D human pose estimation in practical applications such as motion analysis, human-robot interaction, and surveillance systems.

4.8 Qualitative evaluation

To further assess the effectiveness of our proposed method, we perform a qualitative analysis by comparing it with state-of-the-art approaches, including MHFormer, GraphMLP, HSTFormer, and STCFormer. The comparative results, illustrated in Fig. 6, highlight instances of deviated 3D pose predictions, which are marked with dotted black circles, emphasizing the superiority of our approach

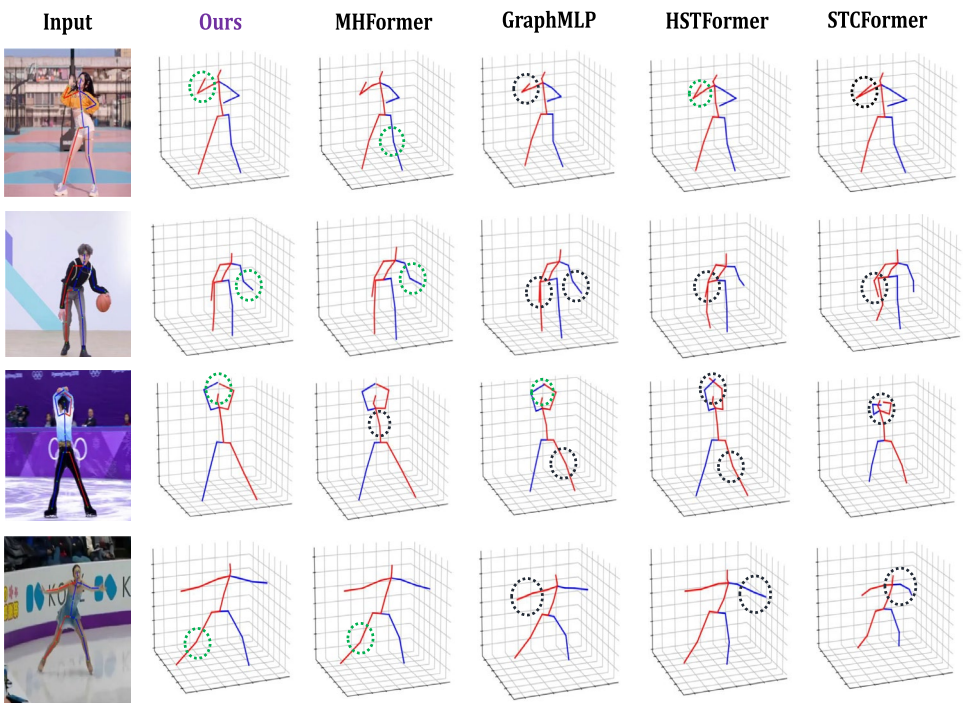

Fig. 6 A qualitative evaluation of in-the-wild videos. The green circle represents the correctly predicted poses, whereas the black circle denotes the incorrectly predicted poses

in handling complex and unconstrained environments. In contrast, green circles indicate regions where our method demonstrates superior accuracy. To ensure a fair evaluation, we utilize the CPN 2D detector [76] to obtain the 2D poses, which are then input into each model. Despite the complexity of dynamic actions and rapid movements, our approach consistently produces more realistic and plausible 3D pose estimations, outperforming previous methods. These findings emphasize the resilience of our model in managing partial occlusions and its proficiency in resolving depth ambiguities efficiently.

5 Discussion

In this study, the FGPD framework combined with the RTCPT architecture demonstrates strong performance on standard benchmarks. However, real-world 3D HPE systems encounter significant domain shifts due to changes in lighting, background, clothing, occlusion, and body appearance. To address this, the proposed fine-grained part-aware prompts provide stable semantic and anatomical priors that remain consistent across domains, thereby supporting robust pose estimation even when visual cues become unreliable. Moreover, the prompt–based communication design enhances structural consistency, which is especially important in low-resource settings where available 3D annotations are limited or noisy. Specifically, the FGPD framework demonstrates improved robustness, and additional progress can be achieved by incorporating domain adaptation techniques that align prompt-based representations with diverse target environments. Additionally, synthetic data present another avenue, since prompt-guided conditioning reduces the gap between real and synthetic distributions. Overall, the proposed method establishes a foundation for robust generalization in real-world and data-limited scenarios and highlights several promising directions for future research.

6 Conclusion

In conclusion, our proposed improved fine-grained prompt-driven denoiser with the refined TCP transformer significantly improves 3D human pose estimation by effectively integrating textual information and human knowledge into the learning process. By leveraging FPP, FPC, and PTS, our method improves the denoising process within the diffusion model, enabling more accurate and robust pose predictions. Specifically, the TCP framework and FAR module collectively enhance spatio-temporal modeling and address depth ambiguity. These components significantly improve motion representation and semantic coherence by dynamically refining temporal dependencies and enabling continuous feature interaction across layers. Experimental evaluations on commonly used datasets demonstrate that our approach outperforms existing methods, producing more realistic and precise 3D pose estimations. For future work, exploring domain adaptation and the use of synthetic data presents promising avenues to enhance the generalization capability

of our approach. Additionally, we aim to integrate more advanced MLPs or GCNs to further boost performance.

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 Conceptualization, methodology, writing - original draft preparation - SA; review, editing, and investigation - SA, SM; Resources - JHJ; Funding acquisition - JHJ, supervision - JHJ

Data availability The data used in this study are available upon reasonable request.

Declarations

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

References

1. Arthanari, S., Elayaperumal, D., Joo, Y.H.: Learning temporal regularized spatial-aware deep correlation filter tracking via adaptive channel selection. Neural Networks **186**, 107210 (2025)
2. Moorthy, S., KS, S.S., Arthanari, S., Jeong, J.H., Joo, Y.H.: Learning disruptor-suppressed response variation-aware multi-regularized correlation filter for visual tracking. J. Vis. Commun. Image Represent., 104458 (2025)
3. KS, S.S., Jeong, J.H., Joo, Y.H.: A multi-level hybrid siamese network using box adaptive and classification approach for robust tracking. Multimed. Tools Appl., 1–26 (2024)
4. Elayaperumal, D., Joo, Y.H.: Learning spatial variance-key surrounding-aware tracking via multi-expert deep feature fusion. Inf. Sci. **629**, 502–519 (2023)
5. Arthanari, S., Moorthy, S., Jeong, J.H., Joo, Y.H.: Adaptive spatially regularized target attribute-aware background suppressed deep correlation filter for object tracking. Signal Proces.: Image Commun. **136**, 117305 (2025)
6. Zhou, L., Chen, Y., Wang, J.: Dual-path transformer for 3d human pose estimation. IEEE Trans. Circuits Syst. Video Technol. **34**(5), 3260–3270 (2023)
7. Xie, B., Liu, G., Deng, F., Lu, M.: Aitepose: Learning an end-to-end monocular 3d human pose estimator via auxiliary-information-driven training enhancement. IEEE Trans. Circ. Syst. Video Technol. (2025)
8. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13232–13242 (2022)
9. Arthanari, S., Jeong, J.H., Joo, Y.H.: Exploiting multi-transformer encoder with multiple-hypothesis aggregation via diffusion model for 3d human pose estimation. Multimed. Tools Appl., 1–29 (2024)
10. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13156 (2022)
11. Arthanari, S., Jeong, J.H., Joo, Y.H.: Exploring multi-level transformers with feature frame padding network for 3d human pose estimation. Multimedia Syst. **30**(5), 243 (2024)
12. Arthanari, S., Moorthy, S., Jeong, J.H., Joo, Y.H.: Adaptive spatially regularized target attribute-aware background suppressed deep correlation filter for object tracking. Signal Processing: Image Commun. **136**, 117305 (2025)
13. Kuppusami Sakthivel, S.S., Moorthy, S., Arthanari, S., Jeong, J.H., Joo, Y.H.: Learning a context-aware environmental residual correlation filter via deep convolution features for visual object tracking. Math. **12**(14), 2279 (2024)
14. Elayaperumal, D., Joo, Y.H.: Robust visual object tracking using context-based spatial variation via multi-feature fusion. Inf. Sci. **577**, 467–482 (2021)
15. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.-J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2272–2281 (2019)
16. Liu, J., Rojas, J., Li, Y., Liang, Z., Guan, Y., Xi, N., Zhu, H.: A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 3374–3380 (2021). IEEE
17. Wu, Y., Kong, D., Wang, S., Li, J., Yin, B.: Hpgcn: Hierarchical poselet-guided graph convolutional network for 3d pose estimation. Neurocomputing **487**, 243–256 (2022)
18. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.-h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8818–8829 (2023)
19. Hassan, M.T., Ben Hamza, A.: Regular splitting graph network for 3d human pose estimation. IEEE Trans. Image Process. **32**, 4212–4222 (2023). https://doi.org/10.1109/TIP.2023.3275914
20. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11656–11665 (2021)
21. Xie, Y., Hong, C., Zhuang, W., Liu, L., Li, J.: Hogformer: high-order graph convolution transformer for 3d human pose estimation. Int. J. Machine Learning and Cybernetics, 1–12 (2024)
22. Chen, Z., Dai, J., Bai, J., Pan, J.: Dgformer: Dynamic graph transformer for 3d human pose estimation. Pattern Recogn. **152**, 110446 (2024)
23. Moorthy, S., KS, S.S., Arthanari, S., Jeong, J.H., Joo, Y.H.: Hybrid multi-attention transformer for robust video object detection. Eng. Appl. Artif. Intell. **139**, 109606 (2025)
24. Xu, J., Guo, Y., Peng, Y.: Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 561–570 (2024)
25. Shan, W., Zhang, Y., Zhang, X., Wang, S., Zhou, X., Ma, S., Gao, W.: Diffusion-based hypotheses generation and joint-level hypotheses aggregation for 3d human pose estimation. IEEE Trans. Circuits Syst. Video Tech. (2024)
26. Tang, Z., Hao, Y., Li, J., Hong, R.: Ftcm: Frequency-temporal collaborative module for efficient 3d human pose estimation in video. IEEE Trans. Circuits Syst. Video Technol. **34**(2), 911–923 (2023)
27. Ma, X., Su, J., Wang, C., Ci, H., Wang, Y.: Context modeling in 3d human pose estimation: A unified perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6238–6247 (2021)
28. Fang, H.-S., Xu, Y., Wang, W., Liu, X., Zhu, S.-C.: Learning pose grammar to encode human body configuration for 3d pose

estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
29. Shuai, H., Wu, L., Liu, Q.: Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4122–4135 (2023). https://doi.org/10.1109/TPAMI.2022.3188716
30. Hua, G., Liu, H., Li, W., Zhang, Q., Ding, R., Xu, X.: Weakly-supervised 3d human pose estimation with cross-view u-shaped graph convolutional network. IEEE Trans. Multimed. **25**, 1832–1843 (2023). https://doi.org/10.1109/TMM.2022.3171102
31. Li, W., Liu, M., Liu, H., Guo, T., Wang, T., Tang, H., Sebe, N.: Graphmlp: A graph mlp-like architecture for 3d human pose estimation. Pattern Recogn. **158**, 110925 (2025)
32. Lu, J., Lin, J., Dou, H., Zeng, A., Deng, Y., Zhang, Y., Wang, H.: Dposer: Diffusion model as robust 3d human pose prior. arXiv preprint arXiv:2312.05541 (2023)
33. Cai, Q., Hu, X., Hou, S., Yao, L., Huang, Y.: Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 882–890 (2024)
34. Chen, Z., Dai, J., Pan, J., Zhou, F.: Diffusion model with temporal constraint for 3d human pose estimation. Vis. Comput., 1–17 (2024)
35. Zheng, H., Li, H., Shi, B., Dai, W., Wang, B., Sun, Y., Guo, M., Xiong, H.: Actionprompt: Action-guided 3d human pose estimation with text and pose prompting. In: 2023 IEEE International Conference on Multimedia and Expo (ICME), pp. 2657–2662 (2023). IEEE
36. Hu, S., Zheng, C., Zhou, Z., Chen, C., Sukthankar, G.: Lamp: Leveraging language prompts for multi-person pose estimation. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3759–3766 (2023). IEEE
37. Chen, H., He, J.-Y., Xiang, W., Cheng, Z.-Q., Liu, W., Liu, H., Luo, B., Geng, Y., Xie, X.: Hdformer: High-order directed transformer for 3d human pose estimation (2023). arXiv preprint arXiv:2302.01825
38. Moorthy, S., Moon, Y.-K.: Hybrid multi-attention network for audio-visual emotion recognition through multimodal feature fusion. Mathematics **13**(7), 1100 (2025)
39. Qian, X., Tang, Y., Zhang, N., Han, M., Xiao, J., Huang, M.-C., Lin, R.-S.: Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation (2023). arXiv preprint arXiv:2301.07322
40. Zhang, Y., Lu, Y., Liu, B., Zhao, Z., Chu, Q., Yu, N.: Evopose: A recursive transformer for 3d human pose estimation with kinematic structure priors. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). IEEE
41. Diaz-Arias, A., Shin, D.: Convformer: parameter reduction in transformer models for 3d human pose estimation by leveraging dynamic multi-headed convolutional attention. Vis. Comput. **40**(4), 2555–2569 (2024)
42. Choi, J., Shim, D., Kim, H.J.: 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 (2023). IEEE
43. Kang, H., Wang, Y., Liu, M., Wu, D., Liu, P., Yuan, X., Yang, W.: 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), pp. 5130–5134 (2024). IEEE
44. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
45. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model (2022). arXiv preprint arXiv:2208.15001
46. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7753–7762 (2019)
47. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pp. 507–523 (2020). Springer
48. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11656–11665 (2021)
49. Shan, W., Lu, H., Wang, S., Zhang, X., Gao, W.: Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3446–3454 (2021)
50. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE Trans. Circuits Syst. Video Technol. **32**(1), 198–209 (2021)
51. Hu, W., Zhang, C., Zhan, F., Zhang, L., Wong, T.-T.: Conditional directed graph convolution for 3d human pose estimation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 602–611 (2021)
52. Zhan, Y., Li, F., Weng, R., Choi, W.: Ray3d: ray-based 3d human pose estimation for monocular absolute 3d localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13116–13125 (2022)
53. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Trans. Multimedia **25**, 1282–1293 (2022)
54. Xue, Y., Chen, J., Gu, X., Ma, H., Ma, H.: Boosting monocular 3d human pose estimation with part aware attention. IEEE Trans. Image Process. **31**, 4278–4291 (2022)
55. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision, pp. 461–478 (2022). Springer
56. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: 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 (2022)
57. Tang, Z., Li, J., Hao, Y., Hong, R.: Mlp-jcg: Multi-layer perceptron with joint-coordinate gating for efficient 3d human pose estimation. IEEE Trans. Multimed. **25**, 8712–8724 (2023). https://doi.org/10.1109/TMM.2023.3240455
58. Einfalt, M., Ludwig, K., Lienhart, R.: Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2903–2913 (2023)
59. Liu, X., Tang, H.: Strformer: Spatial-temporal-retemporal transformer for 3d human pose estimation. Image Vis. Comput. **140**, 104863 (2023)
60. Du, S., Yuan, Z., Lai, P., Ikenaga, T.: Joypose: Jointly learning evolutionary data augmentation and anatomy-aware global-local representation for 3d human pose estimation. Pattern Recogn. **147**, 110116 (2024)
61. Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 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 (2023)
62. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human

pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8877–8886 (2023)
63. Peng, Q., Zheng, C., Chen, C.: A dual-augmentor framework for domain generalization in 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2240–2249 (2024)
64. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.-h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8818–8829 (2023)
65. Cai, J., Liu, M., Liu, H., Zhou, S., Li, W.: Nanohtnet: Nano human topology network for efficient 3d human pose estimation. IEEE Transactions on Image Processing (2025)
66. Wei, M., Xie, X., Zhong, Y., Shi, G.: Learning pyramid-structured long-range dependencies for 3d human pose estimation. IEEE Transactions on Multimedia (2025)
67. Li, W., Liu, M., Liu, H., Guo, T., Wang, T., Tang, H., Sebe, N.: Graphmlp: A graph mlp-like architecture for 3d human pose estimation. Pattern Recogn. **158**, 110925 (2025)
68. Chen, Z., Dai, J., Bai, J., Pan, J.: Dgformer: Dynamic graph transformer for 3d human pose estimation. Pattern Recogn. **152**, 110446 (2024)
69. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3d human pose estimation by generation and ordinal ranking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2325–2334 (2019)
70. Li, C., Lee, G.H.: Weakly supervised generative network for multiple 3d human pose hypotheses (2020). arXiv preprint arXiv:2008.05770
71. Oikarinen, T., Hannah, D., Kazerounian, S.: Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2021). IEEE
72. Wehrbein, T., Rudolph, M., Rosenhahn, B., Wandt, B.: Probabilistic monocular 3d human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11199–11208 (2021)
73. Li, W., Liu, H., Tang, H., Wang, P.: Multi-hypothesis representation learning for transformer-based 3d human pose estimation. Pattern Recogn. **141**, 109631 (2023)
74. Xiang, X., Zhang, K., Qiao, Y., El Saddik, A.: Emhiformer: An enhanced multi-hypothesis interaction transformer for 3d human pose estimation in video. J. Vis. Commun. Image Represent. **95**, 103890 (2023)
75. Li, C., Lee, G.H.: 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 (2019)
76. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)