SemGAN: A Semantic and Hierarchical Adversarial Network for 3D Human Pose Estimation

Haodong Feng, Yu Xin*, and Guoqing Li

Ningbo University, China
{2411100273,xinyu}@nbu.edu.cn, li241700@126.com
*Corresponding author

Abstract. In recent years, significant progress has been made in estimating 3D human joint positions from monocular 2D images or videos. However, existing approaches that lift 2D to 3D often rely on mean per-joint position error (MPJPE) as explicit supervision, while ignoring implicit semantic consistency across 2D–3D sequences and disregarding human kinematic constraints. As a result, these methods may produce physiologically implausible poses. To address this issue, we propose SemGAN, an adversarial learning framework that integrates explicit supervision with implicit semantic constraints. Specifically, we employ a two-stage adversarial learning approach that complements explicit supervision with motion semantic consistency as an implicit constraint, which is further enforced through an Entropy-based Cross-Frequency Discriminator(ECF-Discriminator). This design jointly optimizes both joint coordinate accuracy and motion semantic consistency. Moreover, guided by human kinematic principles, we construct a Semantic Prototype-guided Hierarchical Generator(SPH-Generator), which enhances local part-level uniformity and global inter-part coordination, thereby alleviating abnormal 3D pose generation. Extensive experiments on the Human3.6M and MPI-INF-3DHP benchmark datasets demonstrate that our method outperforms previous state-of-the-art approaches.

Keywords: 3D Human Pose Estimation · Adversarial Learning · Semantic Consistency

1 Introduction

3D Human Pose estimation, a key task in computer vision, aims to predict the joint positions of the human skeleton in 3D space from a monocular 2D image or video. This technology serves as a core foundation for various applications, such as action recognition [14,15] and human-computer interaction [9,32], and its accuracy has a significant impact on the performance of downstream tasks. As a result, research on 3D human pose estimation has garnered widespread attention in recent years.

The current mainstream "2D-to-3D lifting" approach [16,22,27,35] typically follows a two-stage pipeline: first, a 2D human pose detector [2,17,25] is used to

detect keypoints, and then the 2D coordinates are lifted into 3D space. Based on this pipeline, existing works focus on two main directions. On the one hand, some studies aim to enhance spatio-temporal feature modeling capabilities. For example, [33] separates spatial and temporal attention paths to improve long-range dependency modeling while reducing computational cost, whereas [12] introduces a multi-hypothesis generation mechanism to mitigate depth ambiguity through the interaction and fusion of multiple intermediate hypotheses. On the other hand, certain approaches incorporate prior knowledge for optimization. For instance, [28] leverages anatomical structures and joint motion trajectories as priors, and designs kinematic and trajectory attention modules to strengthen spatio-temporal correlation learning.

Despite the significant progress achieved by these methods [3,18,39], they still face two key challenges. First, most existing methods adopt explicit supervision strategies (e.g., MPJPE loss), which emphasize the positional accuracy of 3D joints but overlook the motion semantic consistency across temporal sequences. Motion semantic consistency refers to ensuring that the motion features (i.e., action expression) of the 2D and 3D sequences of the same action remain consistent over time. Second, hierarchical modeling of the human body is often ignored. Human motion inherently exhibits hierarchical characteristics: joints within each of the five major body parts (e.g., left arm, right arm, left leg, right leg, and torso) should maintain local uniformity, while coordination among different parts is necessary to accomplish complex actions. However, most existing approaches [8, 13, 41] treat all joints as a whole, failing to effectively capture such multi-level motion correlations, which limits their ability to understand complex human actions. In response, this paper focuses on addressing the challenges of lacking kinematic semantic consistency and hierarchical modeling.

Regarding motion semantic consistency, a 2D skeleton sequence and its corresponding 3D skeleton sequence essentially represent the same action expressed in different dimensions, and thus share semantic consistency. By leveraging this cross-dimensional semantic correlation, an implicit association can be established between 2D and 3D sequences. Accordingly, we employ a GAN-based adversarial framework, where a semantic-aware discriminator enforces implicit constraints on 2D–3D sequences to ensure that the generated 3D sequence is consistent with the 2D sequence in terms of motion semantics.

For a hierarchical structure, the core challenge lies in effectively capturing fine-grained motion features within local body parts while maintaining coordination across global parts. To address this problem, we design a hierarchical feature learning and cross-level interaction mechanism. By utilizing semantic prototypes to link local functional parts with global motion patterns, the framework is able to capture relative motion features of joints within each part while preserving inter-part dependencies. This enables multi-granularity motion understanding, ranging from local uniformity to global coordination.

Our main contributions can be summarized as follows:

– We propose a 2D-to-3D sequence GAN framework with a two-stage adversarial mechanism. By incorporating motion semantic consistency be-

tween 2D and 3D sequences as implicit supervision, in conjunction with the explicit MPJPE loss, the generated 3D skeleton sequences achieve both joint coordinate accuracy and motion semantic consistency.

– To capture both local and global dynamics, we design the SPH-Generator. It decouples motion into five Basic Functional Parts and introduces a multi-granularity interaction mechanism that balances local uniformity and global coordination. This hierarchical semantic modeling facilitates a deeper understanding of complex actions.

– We propose an ECF-Discriminator that decomposes motion in the frequency domain to capture overall trends, limb dynamics, and local details, using cross-frequency energy entropy interactions to improve temporal coherence in 3D pose generation.

2 Related Work

2D-to-3D Lifting for Human Pose Estimation. Monocular 3D human pose estimation recovers 3D joint coordinates from 2D observations. Methods are divided into single-stage approaches that directly regress 3D poses from images [10,19,24,26,38], and two-stage "2D-to-3D lifting" approaches [29,33,36,37]. The two-stage strategy, which first uses 2D pose detectors [2, 17, 25] and then lifts the coordinates to 3D, has become the dominant paradigm due to advances in 2D detection. Some works use attention mechanisms to focus on key video frames [16]. Others employ graph convolutional networks (GCNs) [4, 40] to model skeletal topology; for instance, SemGCN [44] and LCN [5] learn spatial joint dependencies. However, these methods are limited in capturing long-range and non-local global dependencies.

Transformer-Based Methods. The Transformer architecture [34], with its strong ability to model long-range dependencies, is widely applied in 3D pose estimation for spatiotemporal correlations [1, 31, 47]. The pioneering PoseFormer [46] decouples spatial and temporal modules. MixSTE [43] constructs a hybrid spatiotemporal encoder with alternating layers, while STCFormer [33] uses a parallel dual-path design to efficiently encode spatial and temporal contexts. More recently, studies have begun to explore hierarchical body structures. StridedTransformer [11] hierarchically aggregates features, and Zhang et al. [42] propose a progressive convolutional–temporal fusion method. However, its fixed hierarchical downsampling strategy struggles to adapt to the temporal variations of complex actions, limiting its ability to understand cooperative human motions.

Frequency-Based Feature Representation. Several studies have incorporated frequency-domain transformations into neural networks to enhance feature extraction and robustness. For instance, converting skeleton sequences into DCT coefficients provides an effective representation of motion dynamics [20, 21], leading to their increased use in the 2D-to-3D lifting task. PoseFormerV2 [45] pioneered this approach, using low-frequency DCT coefficients as Transformer inputs to efficiently handle long sequences and improve noise robustness. Subsequent studies have extended this idea by transforming raw signals into the frequency domain to capture spectral interactions across different poses.

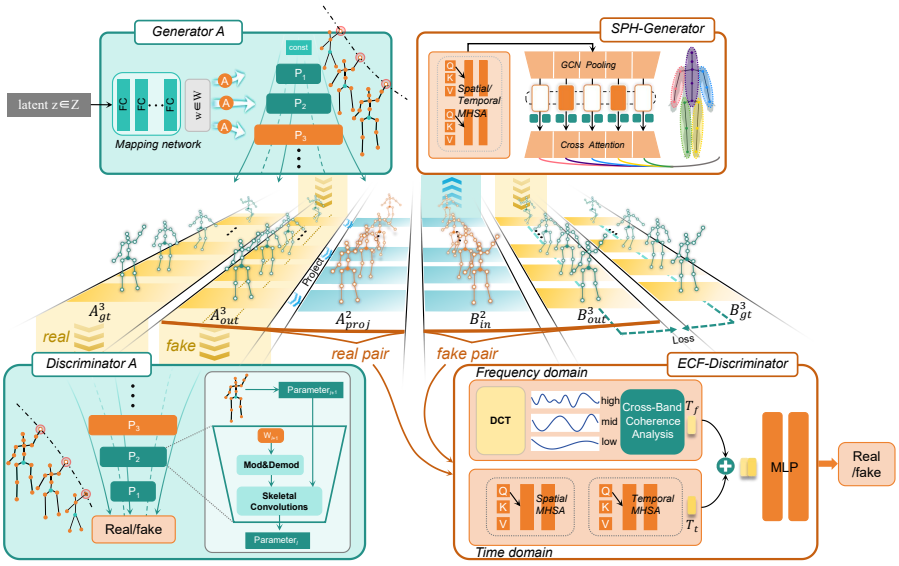

Fig. 1: The overall architecture of our proposed SemGAN model. This framework employs a **two-stage adversarial training** strategy with a **dual-supervision** mechanism. **In the first stage (left)**, a pre-trained GAN is used to generate high-quality 3D motions and their 2D projections, constructing semantically consistent "real pairs." **In the second stage (right)**, the core **SPH-Generator** lifts an input 2D sequence to 3D.

3 Method

3.1 Overall Architecture

2D and 3D sequences of the same action share inherent semantic consistency. Our proposed SemGAN architecture leverages this by combining explicit joint coordinate supervision with implicit, semantic-driven supervision to improve the robustness of 2D-to-3D skeleton reconstruction. As shown in Fig. 1, we introduce a GAN-based framework where a generator and discriminator collaboratively learn motion semantic consistency between 2D and 3D sequences.

Our framework employs a two-stage adversarial training strategy. The first stage generates positive sample pairs with inherent motion semantic consistency, which serve as a reference standard. The second stage uses a dual-supervision mechanism: our Semantic Prototype-guided Hierarchical Generator (SPH-Generator) is supervised by an explicit positional loss (MPJPE), while our Entropy-based Cross-Frequency Discriminator (ECF-Discriminator) provides implicit supervision. The ECF-Discriminator assesses structural and dynamic coherence from both temporal and frequency perspectives, enhancing the semantic consistency and robustness of the generated 3D sequences.

Stage 1: Although existing 3D pose datasets (e.g., Human3.6M) provide accurate 2D-3D ground-truth (GT) pairs, their coverage of action types and

viewpoints is relatively limited, making it difficult to fully encompass the complex and variable human motion patterns in the real world. To break through the bottleneck of limited GT data and enhance the model's generalization ability to unseen actions, we introduce MoDi [30] as a generative module. This module aims to generate diverse and naturally plausible 3D synthetic motion sequences, thereby expanding the motion pattern coverage of the training data. Specifically, Modi's Generator (Generator A) takes a Gaussian noise vector as input, which is transformed into a structured latent variable through a mapping network. Then, a synthesis network, featuring skeleton-aware 3D convolutional modules, hierarchical style injection, and convolutional scaling operators, progressively generates diverse 3D skeleton sequences A^3_{out} in a coarse-to-fine manner. Modi's Discriminator (Discriminator A) employs a hierarchical architecture that is inverse to the generator's synthesis network. It receives either real sequences A^3_{gt} (positive samples) or generated sequences A^3_{out} (negative samples). Through skeleton-aware 3D convolutional modules and downsampling operators, it gradually reduces the temporal and joint resolution to output a discrimination result on the motion's realism, enabling adversarial training. After training, the sequences A^3_{out} generated by Generator A are projected into 2D to form corresponding 2D action sequences A^2_{proj}. Together, A^3_{out} and A^2_{proj} constitute positive sample pairs (A^3_{out}, A^2_{proj}), which provide a real motion reference for the training in the second stage.

Stage 2: The SPH-Generator converts a 2D skeleton sequence B^2_{in} into a 3D skeleton sequence B^3_{out}, forming the sample pair (B^2_{in}, B^3_{out}). To construct the adversarial training for the SPH-Generator, we treat (B^2_{in}, B^3_{out}) as a negative sample pair and the (A^3_{out}, A^2_{proj}) pair obtained from the first stage as a positive sample pair. The ECF-Discriminator is then utilized to perform implicit discrimination through its time-domain and frequency-domain branches. In addition, the explicit supervision loss, Mean Per Joint Position Error (MPJPE), is calculated between the 3D skeleton sequence B^3_{out} generated by the SPH-Generator and the ground-truth 3D skeleton sequence B^3_{gt}.

3.2 Semantic Prototype-guided Hierarchical Generator

Based on human kinematics [6], the body can be divided into five Basic Functional Parts (BFPs): the trunk, arms, and legs. Human motion is characterized by two key principles: 1) **Intra-BFP motion uniformity**, where joints within a part move cohesively, and 2) **Inter-BFP motion coordination**, where different parts interact to maintain overall balance and naturalness.

Generating coherent 3D skeleton sequences requires a model that can consistently handle both fine-grained local actions within BFPs and coarse-grained global interactions between them. To achieve this, we design the Semantic Prototype-guided Hierarchical Generator (SPH-Generator), shown in Fig. 2. The generator takes a 2D skeleton sequence B^2_{in} as input, which first undergoes feature extraction using stacked spatio-temporal Transformers.

Based on human kinematics principles, we divide the J joints in the feature set F_{st} into five functional parts (left arm, right arm, left leg, right leg, and trunk). For the feature subset of each part, F_i (where i $\in$ {1, ..., 5}), we use a

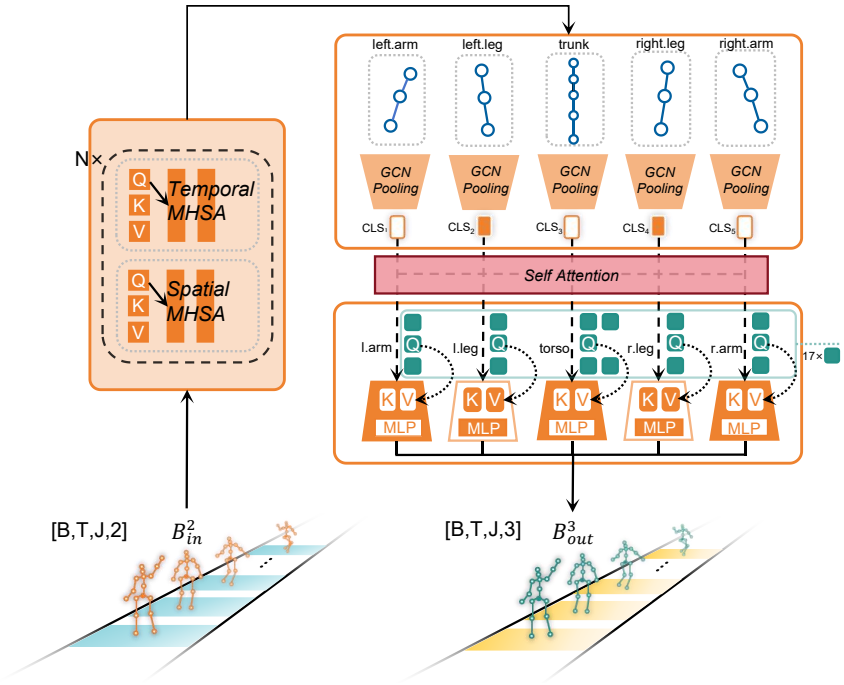

Fig. 2: Architecture of the SPH-Generator

Graph Convolutional Network (GCN) to further strengthen the collaborative modeling of its internal joints and then use a pooling operation to aggregate it into a semantic prototype, cls_i, which represents the overall motion state of the part.

$$cls_i = \text{Pooling}(\text{GCN}(F_i)) \tag{1}$$

This process compresses the complex dynamic information of multiple joints within each part into a single, high-dimensional feature vector. Ultimately, we obtain an initial set of semantic prototypes $cls_{\text{initial}} = \{cls_1, cls_2, cls_3, cls_4, cls_5\}$ representing the five major body parts.

To capture the dynamic synergistic relationships between different body regions, such as the coordinated swinging of arms and legs during walking, we introduce a self-attention mechanism among the five semantic prototypes. This module takes the entire set of prototypes, cls_{initial}, as input to compute the Query, Key, and Value matrices:

$$Q = cls_{\text{initial}}W^Q, \quad K = cls_{\text{initial}}W^K, \quad V = cls_{\text{initial}}W^V \tag{2}$$

where W^Q, W^K, W^V are learnable weight matrices. Subsequently, the prototypes are updated through the standard scaled dot-product attention operation, allowing

each prototype to perceive the state of all other prototypes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

After processing by this module, we obtain a set of enhanced semantic prototypes that have undergone global information interaction, denoted as $\text{CLS}_{\text{updated}}$. This set contains five updated prototypes, each representing a specific body part (referred to as l.arm, l.leg, torso, r.leg, and r.arm in the architecture of the SPH-Generator).

Finally, the model utilizes the enhanced semantic prototypes to guide the generation of 3D coordinates for each joint within its corresponding part. Specifically, for each part, we initialize a set of learnable embeddings to serve as Queries, where the number of Queries strictly matches the number of constituent joints in that specific part (for instance, 3 learnable Queries are assigned to the left arm, denoted as l.arm, corresponding to its 3 joints), and the corresponding enhanced semantic prototype as the Key and Value to perform a cross-attention calculation. This allows the final feature of each joint to be precisely adjusted according to the overall motion state of its part, as encoded by the semantic prototype. Ultimately, the features of all 17 joints across all parts are passed through a regression head to output their coordinates in 3D space, forming the complete 3D skeleton sequence B^3_{out}.

3.3 Entropy-based Cross-Frequency Discriminator

A 3D skeleton sequence (A^3_{out}) and its 2D projected sequence (A^2_{proj}) for the same action are semantically consistent at their core. Through a pairwise comparison of the 2D-3D sample features, it becomes easier to discriminate whether A^3_{out} and A^2_{proj} belong to the same action. Furthermore, the 3D and 2D sequences of the same action exhibit strong consistency in both the temporal and frequency domains. Therefore, fusing the temporal and frequency domain features of the 2D-3D samples can fully express the uniqueness of the action, which is more effective for discriminating them as a corresponding pair.

When designing an action discriminator, this allows for the creation of separate branches to discriminate motion semantics in the time and frequency domains:

1. **Temporal Branch:** This branch needs to capture the structural differences between the 2D-3D sequences in both space and time. By learning spatio-temporal features along the temporal trajectory and in the spatial topology, the discriminator can enhance the feature discriminability between real action pairs and fake action pairs.
2. **Frequency Branch:** This branch needs to establish a multi-scale feature extraction mechanism in the frequency domain, based on the motion semantics of human actions in different frequency bands. It should also utilize a cross-band interaction mechanism to correlate multiple frequency bands, thereby improving its ability to detect abnormal frequency distributions in fake action pairs.

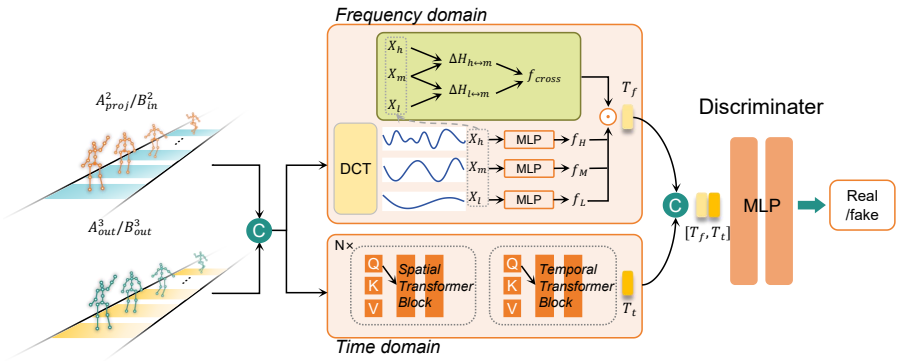

Fig. 3: The overall architecture of the ECF-Discriminator, featuring the Frequency-domain branch above and the Time-domain branch below.

We designed the ECF-Discriminator as shown in Fig. 3. Its input is a 2D-3D sample pair—either a positive pair $p(A^3_{out}, A^2_{proj})$ or a negative one $p(B^2_{in}, B^3_{out})$. The 2D ($P_{2D} \in \mathbb{R}^{T\times J\times 2}$) and 3D ($P_{3D} \in \mathbb{R}^{T\times J\times 3}$) sequences (where T is the number of time steps and J is the number of joints) are first concatenated along the coordinate dimension.

$$P_{\text{combined}} = \text{Concat}(P_{2D}, P_{3D}) \in \mathbb{R}^{T\times J\times 5} \tag{4}$$

A Discrete Cosine Transform (DCT-II) is applied along the time dimension to the combined input, yielding a frequency-domain representation $X \in \mathbb{R}^{T\times J\times 5}$. The transformation is as follows:

$$X_{k,j,c} = \sum_{t=0}^{T-1} P_{\text{combined}}(t,j,c) \cos\left(\frac{\pi}{T}\left(t+\frac{1}{2}\right)k\right) \tag{5}$$

Here, k represents the frequency index, t the time step, j the joint index, and c the coordinate dimension. Subsequently, the DCT coefficients are divided into three non-overlapping frequency bands based on the spectral energy distribution characteristics of human motion. Since human motion is primarily dominated by low-frequency components, whereas high-frequency components typically contain subtle joint jitters or noise, we employ a non-uniform frequency band division strategy to partition the coefficients into: a low-frequency component X_l for the overall motion trend, a mid-frequency component X_m for inter-limb coordination, and a high-frequency component X_h for fine-grained realism.

Independent Band Encoding: The coefficients of each band are averaged along the frequency dimension, flattened, and passed through separate MLP encoders to yield feature vectors $f_L, f_M,$ and f_H.

$$f_B = \text{Encoder}_B(\text{Flatten}(\text{Mean}(X_B))), \quad B \in \{l, m, h\} \tag{6}$$

Cross-band Features: To quantify the smoothness of energy flow between bands, we compute the energy entropy difference. First, the energy E_B for each

band is calculated:

$$E_B = \sum_{k \in B} \sum_{j,c} |X(k,j,c)|^2 \tag{7}$$

The energy at each frequency point is then normalized to form a probability distribution $p_B(k) = \frac{\sum_{j,c} |X(k,j,c)|^2}{E_B}$, allowing for the calculation of the Shannon entropy H_B for the band:

$$H_B = -\sum_{k \in B} p_B(k) \log_2(p_B(k)) \tag{8}$$

The energy entropy difference (ΔH) is then calculated between adjacent bands. A smaller difference indicates a more natural energy transfer.

$$\Delta H_{L \leftrightarrow M} = |H_L - H_M| \tag{9}$$

$$\Delta H_{M \leftrightarrow H} = |H_M - H_H| \tag{10}$$

These differences are encoded via an independent MLP to produce a cross-band feature vector, f_{cross}.

$$f_{\text{cross}} = \text{Encoder}_{\text{cross}}([\Delta H_{L \leftrightarrow M}, \Delta H_{M \leftrightarrow H}]) \tag{11}$$

Specifically, $\text{Encoder}_{\text{cross}}$ is designed as a lightweight three-layer Multilayer Perceptron (MLP) that maps the low-dimensional scalar difference into a high-dimensional feature vector, f_{cross}. This projection aligns f_{cross} within the same dimensional space as the frequency-domain features (f_l, f_m, and f_h). Subsequently, these four features are fused via an attention-weighted mechanism to form the final frequency-domain token (T_f).

$$T_f = W_a \odot [f_L; f_M; f_H; f_{\text{cross}}] \tag{12}$$

Here, $[;]$ denotes concatenation, W_a is an adaptive weight vector generated by an MLP to dynamically adjust the importance of each feature, and $\odot$ represents element-wise multiplication.

The **time-domain branch** uses alternating Spatial-MHSA and Temporal-MHSA layers. Spatial-MHSA focuses on spatial relationships between joints within a frame, while Temporal-MHSA captures temporal dependencies of a joint across frames. This stacked design validates the spatio-temporal coherence and outputs a time-domain feature token (T_t).

Finally, the frequency token T_f and time token T_t are concatenated into a joint feature $T = [T_f, T_t]$ and fed into an MLP to output a discrimination probability. This dual-branch mechanism ensures the generated sequence conforms to realistic motion patterns across its overall trend, limb coordination, and fine-grained dynamics.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct a comprehensive evaluation of our experiments on two public datasets: Human3.6M [7] and MPI-INF-3DHP [23].

Human3.6M is the most representative benchmark dataset in the field of 3D human pose estimation. It contains approximately 3.6 million video frames collected in an indoor environment, captured by four synchronized cameras from different viewpoints. The dataset involves 11 professional actors performing 15 common activities. Following previous studies [12, 46], we use data from five subjects (S1, S5, S6, S7, S8) for training and data from two subjects (S9, S11) for testing. We use the most common metrics to evaluate our proposed method: Protocol 1 is the Mean Per Joint Position Error (MPJPE), which calculates the error between the predicted and ground-truth joint coordinates; Protocol 2 is the Procrustes-aligned Mean Per Joint Position Error (P-MPJPE), which refers to the MPJPE calculated after further aligning the predicted joint coordinates with the ground-truth coordinates.

MPI-INF-3DHP is a large-scale 3D pose dataset that includes three scenes: green screen, non-green screen, and outdoor. It contains 1.3 million frames from 8 subjects, with the training set including 8 activities and the test set including 7 activities. Following previous settings [12, 28], we use three evaluation metrics: Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints (PCK) within a 150mm range, and Area Under the Curve (AUC).

4.2 Implementation Details

We implement our model based on the PyTorch framework and conduct training and evaluation on a server equipped with four NVIDIA A6000 GPUs. During model training, we set the number of stacked spatial-temporal Transformers (N) to 7 and use the Adam optimizer to optimize the model. We train for 80 epochs with an initial learning rate of 8e-5, which is decayed by a factor of 0.99 each epoch. For Human3.6M, following established settings [12, 28], there are two types of input 2D joint sequences: 2D keypoints detected by a pre-trained CPN model and 2D ground truth (GT) labels. For MPI-INF-3DHP, following previous settings, we directly use the 2D ground truth (GT) labels as input. During the training and testing phases, we follow the same settings as in prior works and use horizontally flipped pose data for data augmentation.

4.3 Comparison with State-of-the-Art Methods

Results on the Human3.6M Dataset. We evaluated our method on the Human3.6M dataset using CPN-detected 2D poses, with results shown in Table 1. With an input of $T = 243$ frames, our method achieves a state-of-the-art MPJPE of 39.7mm, outperforming methods like STCFormer (40.5mm) and MixSTE (40.9mm). Our model also demonstrates high efficiency; with a shorter input

Table 1: Quantitative comparison results with the state-of-the-art methods on Human3.6M. The 2D poses obtained by CPN [2] are used as inputs. Bold: best; Underlined: second best.

MPJPE (CPN)

Method	Publication	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
UGCN [35](T=96)	ECCV'20	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
PoseFormer [46](T=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
StridedFormer [11](T=351)	TMM'22	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
MHFormer [12](T=351)	CVPR'22	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
P-STMO [31](T=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 [43](T=81)	CVPR'22	39.8	43.0	38.6	40.1	43.4	50.6	40.6	41.4	52.2	56.7	43.8	40.8	43.9	29.4	30.3	42.4
MixSTE [43](T=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
GLA-GCN [40](T=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
HDformer [1](T=96)	IJCAI'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
STCFormer [33](T=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
STCFormer [33](T=243)	CVPR'23	38.4	41.2	36.8	38.0	42.7	50.5	38.7	**38.2**	52.5	56.8	41.8	**38.4**	40.2	**26.2**	27.7	40.5
KTPFormer [28](T=81)	CVPR'24	39.1	41.9	37.3	40.1	44.0	51.3	39.8	41.0	51.4	56.0	43.0	41.0	42.6	28.8	29.5	41.8
KTPFormer [28](T=243)	CVPR'24	37.3	39.2	35.9	37.6	42.5	**48.2**	**38.6**	39.0	51.4	55.9	41.6	39.0	40.0	27.0	27.4	40.1
Ours (T=81)		38.8	41.0	36.6	38.8	43.7	51.6	40.8	40.2	52.0	58.0	42.4	40.9	41.8	27.9	28.3	41.5
Ours (T=243)		**36.9**	**39.0**	**35.1**	**36.7**	**41.9**	48.5	38.8	38.6	**51.3**	56.3	**41.3**	38.8	**39.4**	26.7	**26.5**	**39.7**

P-MPJPE (CPN)

Method	Publication	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
UGCN [35](T=96)	ECCV'20	31.8	34.3	35.4	33.5	35.4	41.7	31.1	31.6	44.4	49.0	36.4	32.2	35.0	24.9	23.0	34.5
PoseFormer [46](T=81)	ICCV'21	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
StridedFormer [11](T=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
MHFormer [12](T=351)	CVPR'22	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
P-STMO [31](T=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 [43](T=81)	CVPR'22	32.0	34.2	31.7	33.7	34.4	39.2	32.0	31.8	42.9	46.9	35.5	32.0	34.4	23.6	25.2	33.9
MixSTE [43](T=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
GLA-GCN [40](T=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
HDformer [1](T=96)	IJCAI'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
STCFormer [33](T=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
STCFormer [33](T=243)	CVPR'23	**29.3**	33.0	30.7	**30.6**	32.7	38.2	**29.7**	**28.8**	42.2	45.0	**33.3**	29.4	31.5	20.9	**22.3**	31.8
KTPFormer [28](T=81)	CVPR'24	30.6	33.4	30.1	31.9	33.7	38.2	30.6	30.7	40.9	44.8	34.4	30.5	32.7	22.3	24.0	32.6
KTPFormer [28](T=243)	CVPR'24	30.1	32.3	29.6	30.8	**32.3**	37.3	30.0	30.2	41.0	45.3	33.6	29.3	31.4	21.5	22.6	31.9
Ours (T=81)		29.9	32.7	29.8	31.6	33.9	38.5	30.9	30.5	41.5	45.6	34.2	30.2	32.6	22.0	23.7	32.5
Ours (T=243)		29.5	**32.1**	**29.3**	30.7	32.6	**37.1**	30.4	29.6	**40.6**	45.1	33.7	**29.2**	**31.1**	**20.6**	22.4	**31.6**

of $T = 81$, its 41.5mm MPJPE not only improves upon STCFormer (42.0mm) but also surpasses MHFormer (43.0mm), which uses a much longer sequence of $T = 351$ frames.To evaluate the upper-bound performance, we used ground-truth 2D poses as input (Table 2). In this setting, SemGAN achieves a new state-of-the-art MPJPE of 18.8mm, outperforming the previous best competitor, KTPFormer.

Results based on MPI-INF-3DHP. To verify the generalization ability of our method, we conducted evaluations on the more challenging MPI-INF-3DHP dataset, which features more complex backgrounds and outdoor scenes. Following the setup of previous studies [28, 33], we used ground-truth 2D poses as input. Considering the shorter video sequences in this dataset, we set the input frame counts to 27 and 81. As shown in Table 3, our method achieves excellent performance at $T = 81$, with a PCK of 98.7%, an AUC of 86.2%, and an MPJPE of 16.5mm. Compared to the current state-of-the-art method, KTPFormer, our model achieves a 0.3% improvement in AUC and a 0.2mm reduction in MPJPE, establishing a new state-of-the-art result of 16.5mm in this key metric. Furthermore, our method remains highly competitive with an input frame count of $T = 27$, achieving an AUC of 84.8% and an MPJPE of 19.3mm, which proves its effectiveness in handling shorter sequences.

Table 2: Quantitative comparison results of MPJPE (mm) with the state-of-the-art methods on Human3.6M using ground-truth (GT) 2D poses as inputs. T is the number of input frames.

		MPJPE (GT)															
Method	Publication	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
UGCN [35](T=96)	ECCV'20	23.0	25.7	22.8	22.6	24.1	30.6	24.9	24.5	31.1	35.0	25.6	24.3	25.1	19.8	18.4	25.6
PoseFormer [46](T=81)	ICCV'21	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
StridedFormer [11](T=351)	TMM'22	27.1	29.4	26.5	27.1	28.6	33.0	30.7	26.8	38.2	34.7	29.1	29.8	26.8	19.1	19.8	28.5
MHFormer [12](T=351)	CVPR'22	27.7	32.1	29.1	28.9	30.0	33.9	33.0	31.2	37.0	39.3	30.0	31.0	29.4	22.2	23.0	30.5
P-STMO [31](T=243)	ECCV'22	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	18.7	29.0
MixSTE [43](T=81)	CVPR'22	25.6	27.8	24.5	25.7	24.9	29.9	28.6	27.4	29.9	29.0	26.1	25.0	25.2	18.7	19.9	25.9
MixSTE [43](T=243)	CVPR'22	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
GLA-GCN [40](T=243)	ICCV'23	20.1	21.2	20.0	19.6	21.5	26.7	23.3	19.8	27.0	29.4	20.8	20.1	19.2	**12.8**	**13.8**	21.0
STCFormer [33](T=81)	CVPR'23	26.2	26.5	23.4	24.6	25.0	28.6	28.3	24.6	30.9	33.7	25.7	25.3	24.6	18.6	19.7	25.7
STCFormer [33](T=243)	CVPR'23	21.4	22.6	21.0	21.3	23.8	26.0	24.2	20.0	28.9	28.0	22.3	21.4	20.1	14.2	15.0	22.0
KTPFormer [28](T=81)	CVPR'24	22.5	22.4	21.3	21.4	21.2	25.5	24.2	22.4	24.4	27.5	22.7	21.4	21.7	16.3	17.3	22.2
KTPFormer [28](T=243)	CVPR'24	19.6	**18.6**	18.5	18.1	18.7	22.1	20.8	18.3	22.8	**22.4**	**18.8**	18.1	18.4	13.9	15.2	19.0
Ours (T=81)		21.9	22.2	20.8	20.5	21.0	25.6	23.8	21.6	24.7	27.3	21.5	20.7	21.4	14.8	16.4	21.6
Ours (T=243)		**19.4**	18.8	**18.3**	**18.0**	**18.5**	**21.7**	**20.5**	**18.0**	**22.6**	22.7	19.1	**18.0**	**18.2**	13.5	14.8	**18.8**

Table 3: Performance comparisons on MPI-INF-3DHP

Method	**Publication**	**PCK↑**	**AUC↑**	**MPJPE↓**
UGCN [35] (T=96)	ECCV'20	86.9	62.1	68.1
PoseFormer [46] (T=9)	ICCV'21	88.6	56.4	77.1
MHFormer [12] (T=9)	CVPR'22	93.8	63.3	58.0
MixSTE [43] (T=27)	CVPR'22	94.4	66.5	54.9
P-STMO [31] (T=81)	ECCV'22	97.9	75.8	32.2
PoseFormerV2 [45] (T=81)	CVPR'23	97.9	78.8	27.8
GLA-GCN [40] (T=81)	ICCV'23	98.5	79.1	27.7
STCFormer [33] (T=27)	CVPR'23	88.4	83.4	24.2
STCFormer [33] (T=81)	CVPR'23	98.7	83.9	23.1
KTPFormer [28] (T=27)	CVPR'24	**98.9**	84.4	19.2
KTPFormer [28] (T=81)	CVPR'24	**98.9**	85.9	16.7
Ours (T=27)		98.6	84.8	19.3
Ours (T=81)		98.7	**86.2**	**16.5**

Metric ↑ denotes the higher, the better; ↓ denotes the lower, the better.

Table 4: Ablation study on different components of our SemGAN.

Method	SPH-Generator	Time Branch	Frequency Branch	MPJPE
Baseline	✗	✗	✗	51.2
SPH-Time	✓	✓	✗	44.2
SPH-Freq	✓	✗	✓	44.5
Time-Freq	✗	✓	✓	42.4
SemGAN	✓	✓	✓	39.7

4.4 Ablation Study

To validate the effectiveness of the core components in our proposed SemGAN, we conducted a series of ablation studies on the Human3.6M dataset. All experiments use 2D poses detected by CPN as input, with a fixed input frame count of $T = 243$.

As shown in Table 4, we progressively dismantled the model to quantify the contribution of each innovative design. Our baseline model (a) removes the entire GAN framework, retaining only the SPH-Generator with pure explicit supervision using MPJPE.

Effectiveness of the SPH-Generator. In the Time-Freq experiment, we removed the Semantic Prototype-guided Hierarchical Generator (SPH-Generator). Compared to the full model (SemGAN), the MPJPE increased by 1.5mm (from 43.9mm to 45.4mm). This indicates that by partitioning the body into five Basic Functional Parts (BFPs) and using semantic prototypes to constrain their internal and external motor coordination, the SPH-Generator can more effectively model the intrinsic kinematic structure of the human body, thereby enhancing the overall plausibility of the generated actions.

Effectiveness of the Dual Branches. In the SPH-Time and SPH-Freq experiments, we removed the frequency-domain and time-domain branches, respectively. The results show that removing either branch leads to performance degradation. Specifically, removing the frequency branch (SPH-Time) caused the MPJPE to increase by 0.3mm, while removing the time branch (SPH-Freq) led to a 0.6mm increase. This confirms that the time and frequency domains provide complementary perspectives for action discrimination: the time-domain branch focuses on the coherence of spatio-temporal topology, while the frequency-domain branch examines motion rhythms at different scales. Modeling both in parallel allows for the construction of a more robust and comprehensive discrimination criterion.

Impact of Frequency Band Division Strategy: To verify the rationality of the non-uniform frequency band division in the ECF-Discriminator, we compared different division ratios, as shown in Table 5. Experimental results demonstrate that division strategies deviating from the energy distribution of human motion lead to performance degradation. Specifically, the uniform split (a) broadens the low-frequency range to 33%, causing mid-frequency signals (e.g., limb coordination) to mix into global trajectory features, which weakens the discriminator's ability to capture macroscopic motion trajectories. The low-frequency bias (b) ensures trajectory smoothness but struggles to accurately capture the subtle jitters of distal joints due to the limited high-frequency bandwidth (only 20%). The high-frequency bias (c) loses the low-frequency information containing the primary motion components due to an overemphasis on details, resulting in significant distortions in the overall skeletal structure. In contrast, our strategy (10%-20%-70%) effectively aligns with the natural energy distribution of human motion. It utilizes a narrower band to precisely capture high-energy primary motion components while retaining a sufficiently wide high-frequency band to monitor generative noise.

Table 5: Ablation study on different frequency band division strategies.

Strategy	Low Freq (X_l)	Mid Freq (X_m)	High Freq (X_h)	MPJPE	P-MPJPE
(a) Uniform Split	33%	33%	33%	40.5	33.1
(b) Low-Frequency Bias	50%	30%	20%	40.2	32.3
(c) High-Frequency Bias	5%	10%	85%	41.5	34.1
(d) Ours (Physiological)	**10%**	**20%**	**70%**	**39.7**	**31.6**

Parameter-matched comparison. To further verify that the improvement of SemGAN does not simply come from increased model capacity, we compare it with KTPFormer and a parameter-matched flat Transformer baseline under the same setting. As shown in Table 6, the flat Transformer has comparable parameters and FLOPs to SemGAN, but obtains a worse MPJPE. This indicates that the performance gain mainly stems from the hierarchical semantic modeling and cross-frequency motion constraints, rather than merely increasing the number of parameters.

Table 6: Complexity and parameter-matched comparison on Human3.6M with CPN 2D inputs.

Method	Params↓	FLOPs↓	MPJPE↓
KTPFormer [28]	33.652M	139.059G	40.1
Flat Trans.	36.741M	145.214G	40.4
SemGAN	36.569M	144.662G	**39.7**

4.5 Qualitative Analysis

To intuitively evaluate the performance and robustness of our method, we show a qualitative comparison of SemGAN against KTPFormer on the Human3.6M dataset in Fig. 4. The figure displays two representative actions: "Eating" (first row) and "SittingDown" (second row). These poses are typically challenging for 2D-to-3D lifting due to their complex body articulations and significant self-occlusions (e.g., occlusion of arms and legs), which introduces significant depth ambiguity. As can be seen, our method generates 3D poses that are more consistent with the Ground Truth than KTPFormer, faithfully capturing correct spatial relationships. Particularly for the complex "SittingDown" action, where the lower body is heavily occluded, KTPFormer produces noticeable artifacts. In contrast, SemGAN still estimates the 3D coordinates accurately and reconstructs a structurally plausible and precise 3D pose, demonstrating its superior effectiveness in resolving ambiguities.

5 Conclusion

In this paper, we propose an adversarial learning framework that integrates explicit supervision with implicit semantic constraints to synergistically optimize both joint coordinate accuracy and motion semantic consistency. We design a SPH-Generator that enhances the understanding of complex actions by decoupling the body's functional parts and establishing a multi-granularity interaction mechanism. Additionally, we construct a ECF-Discriminator that implements multi-scale semantic discrimination by analyzing signals at low, medium, and

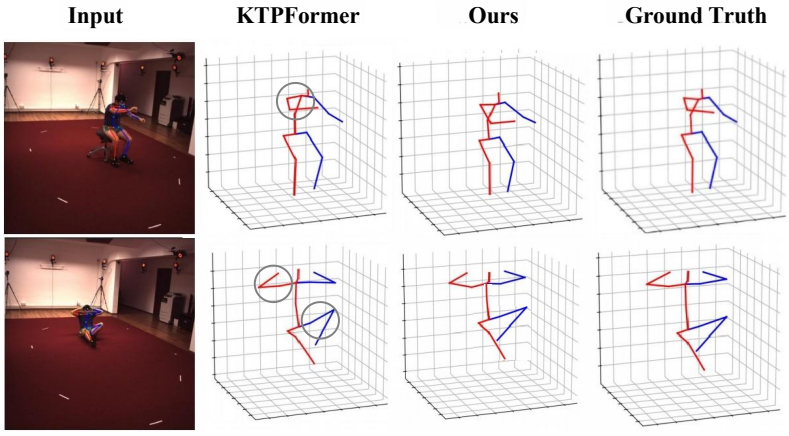

Fig. 4: Visualization of 3D pose estimation results comparing SemGAN and KTPFormer on the Human3.6M dataset.

high-frequency levels through a cross-frequency interaction strategy. Experiments on two benchmarks demonstrate SemGAN's effectiveness and good generalization ability compared to state-of-the-art techniques.

Acknowledgements

We acknowledge the support of the Natural Science Foundation of Zhejiang Province, China (Grant No. LY22F020001, No.LZ20F020001), the 3315 Plan Foundation of Ningbo (Grant No. 2019B-18-G), China Natural Science Foundation under Grant 62271274, the support of Research and Application of A Multi Billion Parameter Monitoring Video Model for domestic full stack AI infrastructure, China (Grant No. 2024Z004) and Ningbo "Kechuang Yongjiang 2035" Key Technology (2024Z245).

References

1. Chen, H., He, J.Y., Xiang, W., Cheng, Z.Q., Liu, W., Liu, H., Luo, B., Geng, Y., Xie, X.: Hdformer: High-order directed transformer for 3d human pose estimation. arXiv preprint arXiv:2302.01825 (2023)
2. 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)
3. Chen, Z., Wan, X., Bao, Y., Zhao, X.: Joint-limb compound triangulation with co-fixing for stereoscopic human pose estimation. IEEE Transactions on Multimedia **26**, 10708–10719 (2024)
4. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision. pp. 769–787. Springer (2020)

5. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2262–2271 (2019)
6. Fohanno, V., Begon, M., Lacouture, P., Colloud, F.: Estimating joint kinematics of a whole body chain model with closed-loop constraints. Multibody System Dynamics **31**(4), 433–449 (2014)
7. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
8. Kang, H., Wang, Y., Liu, M., Wu, D., Liu, P., Yang, W.: Double-chain constraints for 3d human pose estimation in images and videos. arXiv preprint arXiv:2308.05298 (2023)
9. Kisacanin, B., Pavlovic, V., Huang, T.S.: Real-time vision for human-computer interaction. Springer Science & Business Media (2005)
10. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian conference on computer vision. pp. 332–347. Springer (2014)
11. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia **25**, 1282–1293 (2022)
12. 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)
13. Liu, H., Xiang, W., He, J.Y., Cheng, Z.Q., Luo, B., Geng, Y., Xie, X.: Refined temporal pyramidal compression-and-amplification transformer for 3d human pose estimation. arXiv preprint arXiv:2309.01365 (2023)
14. Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition **68**, 346–362 (2017)
15. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1159–1168 (2018)
16. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: 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 (2020)
17. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., et al.: Mediapipe: A framework for perceiving and processing reality. In: Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR). vol. 2019 (2019)
18. Ma, H., Wang, Z., Chen, Y., Kong, D., Chen, L., Liu, X., Yan, X., Tang, H., Xie, X.: Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In: European Conference on Computer Vision. pp. 424–442. Springer (2022)
19. 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)
20. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: European Conference on Computer Vision. pp. 474–489. Springer (2020)

21. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019)
22. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 2640–2649 (2017)
23. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017)
24. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10133–10142 (2019)
25. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
26. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2d pose information. In: European Conference on Computer Vision. pp. 156–169. Springer (2016)
27. 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)
28. Peng, J., Zhou, Y., Mok, P.: Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1123–1132 (2024)
29. Qian, X., Tang, Y., Zhang, N., Han, M., Xiao, J., Huang, M.C., Lin, R.S.: Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. arXiv preprint arXiv:2301.07322 (2023)
30. Raab, S., Leibovitch, I., Li, P., Aberman, K., Sorkine-Hornung, O., Cohen-Or, D.: Modi: Unconditional motion synthesis from diverse data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13873–13883 (2023)
31. 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. Springer (2022)
32. Svenstrup, M., Tranberg, S., Andersen, H.J., Bak, T.: Pose estimation and adaptive robot behaviour for human-robot interaction. In: 2009 IEEE International Conference on Robotics and Automation. pp. 3571–3576. IEEE (2009)
33. 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)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3d pose estimation from videos. In: European conference on computer vision. pp. 764–780. Springer (2020)
36. Wang, R., Ying, X., Xing, B.: Exploiting temporal correlations for 3d human pose estimation. IEEE Transactions on Multimedia **26**, 4527–4539 (2023)
37. Wang, Y., Kang, H., Wu, D., Yang, W., Zhang, L.: Global and local spatio-temporal encoder for 3d human pose estimation. IEEE Transactions on Multimedia **26**, 4039–4049 (2023)

38. 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)
39. Wu, L., Yu, Z., Liu, Y., Liu, Q.: Limb pose aware networks for monocular 3d pose estimation. IEEE Transactions on Image Processing **31**, 906–917 (2021)
40. 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)
41. Zeng, A., Sun, X., Yang, L., Zhao, N., Liu, M., Xu, Q.: Learning skeletal graph neural networks for hard 3d pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11436–11445 (2021)
42. Zhang, H., Hu, Z., Sun, Z., Zhao, M., Bi, S., Di, J.: A fused convolutional spatio-temporal progressive approach for 3d human pose estimation. The Visual Computer **40**(6), 4387–4399 (2024)
43. 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)
44. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3425–3435 (2019)
45. 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)
46. 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)
47. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15085–15099 (2023)