NanoHTNet: Nano Human Topology Network for Efficient 3D Human Pose Estimation

Jialun Cai, Mengyuan Liu, Hong Liu, *Member, IEEE*, Shuheng Zhou, and Wenhao Li

Abstract*—The widespread application of 3D human pose estimation (HPE) is limited by resource-constrained edge devices like Jetson Nano, requiring more efficient models. A key approach to enhancing efficiency involves designing networks based on the structural characteristics of input data. However, effectively utilizing the structural priors in human skeletal inputs remains challenging. To address this, we leverage both explicit and implicit spatio-temporal priors of the human body through innovative model design and a pre-training proxy task. First, we propose a Nano Human Topology Network (NanoHTNet), a tiny 3D HPE network with stacked Hierarchical Mixers to capture explicit features. Specifically, the spatial Hierarchical Mixer efficiently learns the human physical topology across multiple semantic levels, while the temporal Hierarchical Mixer with discrete cosine transform and low-pass filtering captures local instantaneous movements and global action coherence. Moreover, Efficient Temporal-Spatial Tokenization (ETST) is introduced to enhance spatio-temporal interaction and reduce computational complexity significantly. Second, PoseCLR is proposed as a general pre-training method based on contrastive learning for 3D HPE, aimed at extracting implicit representations of human topology. By aligning 2D poses from diverse viewpoints in the proxy task, PoseCLR aids 3D HPE encoders like NanoHTNet in more effectively capturing the high-dimensional features of the human body, leading to further performance improvements. Extensive experiments verify that NanoHTNet with PoseCLR outperforms other state-of-the-art methods in efficiency, making it ideal for deployment on edge devices like the Jetson Nano. Code and models are available at *https://github.com/vefalun/NanoHTNet

***Index Terms*—3D human pose estimation, human topology, efficient, contrastive learning.**

Received 3 February 2025; revised 6 August 2025; accepted 31 August 2025. Date of publication 17 September 2025; date of current version 13 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62473007, in part by the Natural Science Foundation of Guangdong Province under Grant 2024A1515012089, in part by Shenzhen Innovation in Science and Technology Foundation for the Excellent Youth Scholars under Grant RCYX20231211090248064, and in part by the Ant Group Research Fund. The associate editor coordinating the review of this article and approving it for publication was Prof. Wanqing Li. *(Corresponding author: Mengyuan Liu.)*

Jialun Cai was with the State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Guangdong 518055, China. He is now with Ant Group, Hangzhou 310000, China (e-mail: caijialun.cjl@antgroup.com).

Mengyuan Liu and Hong Liu are with the State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Guangdong 518055, China (e-mail: liumengyuan@pku.edu.cn; hongliu@pku.edu.cn).

Shuheng Zhou is with Ant Group, Hangzhou 310000, China (e-mail: shuheng.zsh@antgroup.com).

Wenhao Li is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wenhao.li@ntu.edu.sg).

Digital Object Identifier 10.1109/TIP.2025.3608662

I. Introduction

3D HUMAN pose estimation (HPE) is a fundamental computer vision task aimed at estimating 3D human joint positions from images or videos. Recently, many effective 3D HPE methods [1], [2], [3], [4], [5] have been proposed and widely applied in human-computer interaction contexts such as health monitoring [6], rehabilitation training [7], and virtual reality [8]. Notably, the deployment of these methods on edge devices like wearables, smart cameras, and mobile phones has significantly expanded the potential applications of 3D vision technologies.

However, deploying 3D HPE methods on these edge devices presents significant efficiency challenges. The mainstream approaches in 3D HPE [9], [10], [11], [12], [13] typically adopt a two-stage pipeline. In the first stage, lightweight 2D HPE models [14], [15], [16] are employed to estimate 2D poses from images or videos in real time. In the second stage, these 2D poses are subsequently lifted to 3D poses. However, as illustrated in Fig. 1, these methods fail to achieve real-time inference in the second stage when deployed on the Jetson Nano, a widely used AI edge device. One key factor contributing to this limitation is that these 2D-to-3D pose lifting methods rely on pixel-based vision approaches that overlook the spatio-temporal priors inherent in human topology, resulting in redundant computations. This leads us to ask:

- *How to better leverage spatio-temporal human priors to enhance data comprehension for more efficient 3D HPE?*

To address this, we explore human priors from the following three perspectives:

(i) Spatial hierarchical structure: The human body can be anatomically divided into multiple hierarchical levels: joints, parts, and the whole body (see in Fig. 2(a)). At the joint level, joints are connected locally through the skeleton; at the part level, the joints in the same part cooperate closely during movement, and they will always advance and retreat together; at the body level, the inter-part joints are also globally connected with each other. For instance, running, a fundamental human activity, is characterized by the natural coordination of contralateral limbs, where the left arm synchronizes with the right leg, rather than ipsilateral limb movement.

(ii) Temporal multi-scale kinematics: Given the kinematic principle, the temporal movement dynamics of human can be divided into local instantaneous features and global coherence features. As shown in Fig. 2(b), local features mainly focus on

1941-0042 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

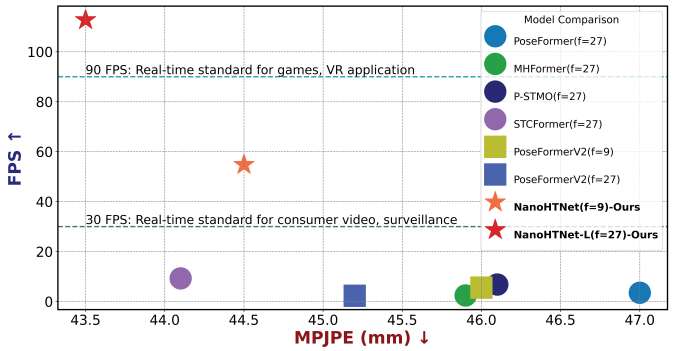


Fig. 1. Comparison of performance (MPJPE) and speed (FPS) between the proposed NanoHTNet and SOTA methods on the Jetson Nano, a compact and power-efficient AI device. NanoHTNet significantly excels in the speed-accuracy trade-off, enabling real-time inference for 3D HPE in edge AI.

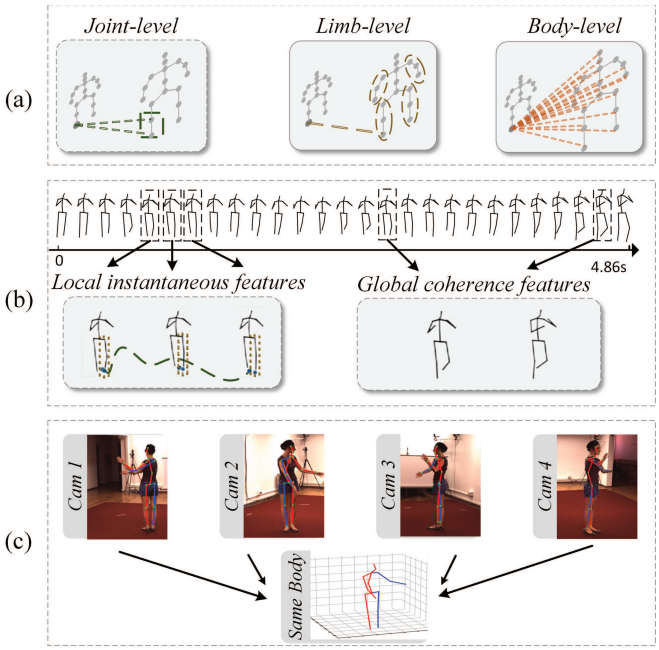


Fig. 2. Spatio-temporal priors of the human skeleton: (a) Spatial hierarchical structure; (b) Temporal multi-scale kinematics; (c) Cross-view spatio-temporal topological consistency.

the immediate movements between adjacent frames, including variations in joint angles, velocity, and acceleration, which are crucial for predicting short-term movements. Meanwhile, global features within the sequence (*e.g.*, a 243-frame input spanning 4.86 seconds) concentrate on the overall coherence and coordination of actions. This enables the comprehension of global movement patterns and rhythms, such as the repetitive and cyclical nature in running, thereby aiding the model in better understanding the consistency of human motion.

(iii) Cross-view spatio-temporal topological consistency: The topological architecture of the human body exhibits consistent spatial representation across all viewpoints at the same moment as shown in Fig. 2(c). Such consistency is crucial for developing robust models that are capable of generalizing across a wide range of poses and orientations. In insight of this, models can more accurately predict joint coordinates by aligning body parts from various perspectives.

Driven by the first two observations, a Nano Human Topology Network (NanoHTNet) is proposed. NanoHTNet utilizes a dual-stream architecture to parallelly extract spatial and temporal kinematic features of human topology through stacked Hierarchical Mixers. Specifically, the spatial Hierarchical Mixers, composed of Local Joint-level Connection (LJC), Intra-Part Constraint (IPC), and Global Body-level Interaction (GBI) modules, capture the hierarchical features of the human skeleton from joints to the entire body. The temporal Hierarchical Mixers, with Instantaneous Motion Extraction (IME) and Global Coherence Perception (GCP) modules, learn both short-term and long-term motion characteristics. Moreover, Discrete Cosine Transform (DCT) separates low-frequency global kinematics from high-frequency noise. A Low-pass Filter (LF) is then applied to remove this noise and reduce data redundancy. To further enhance spatio-temporal feature interaction, Efficient Temporal-Spatial Tokenization (ETST) module is proposed. ETST rearchitects data mapping by creating fused tokens: temporal tokens that encapsulate the full body joints per frame and spatial tokens that represent each joint's entire trajectory. This strategy strengthens feature interaction from the outset while significantly reducing complexity and parameters.

Further inspired by the cross-view topological consistency, we introduce PoseCLR, a self-supervised learning method utilizing contrastive learning to extract structural priors for better model parameter initialization. Specifically, in the pre-training stage, a contrastive proxy task is designed to align the 2D poses of the same person at the same moment across different viewpoints, effectively capturing high-dimensional feature representations of the human body. In the fine-tuning phase, the pre-trained parameters are loaded into the encoder to train for the monocular 3D human pose estimation task. The encoder could be various mainstream 2D-to-3D lifting implementations, including NanoHTNet, GraFormer [17], MixSTE [18], PoseFormerV2 [5], etc.

Therefore, by leveraging unlabeled multi-view data or the inherent viewpoint information in existing datasets [19], PoseCLR could introduce better initialization parameters that represent the implicit structural prior of human topology for 3D HPE encoders. This approach achieves a generic pre-training approach for 2D-to-3D lifting HPE without extra labeled data and model size increase.

In summary, the main contributions are as follows:

- A Nano Human Topology Network (NanoHTNet) is proposed, which can infer on edge AI devices like Jetson Nano with excellent efficiency.
- Spatial and temporal Hierarchical Mixers with a channel-split progressive structure are designed to efficiently learn multilevel spatio-temporal features of the human body.
- To the best of our knowledge, PoseCLR is the first to introduce a multi-view contrastive learning framework to the 2D-to-3D HPE task, which utilizes cross-view consistency of the human body to optimize model initialization.
- Experiments show that NanoHTNet effectively extracts explicit spatio-temporal human features, while PoseCLR, as a general pre-training method for 2D-to-3D HPE,

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

further captures implicit features for better comprehension of the human body, leading to efficient 3D HPE.

This paper is the extension of our conference paper [20], and NanoHTNet extends in the following aspects: (1) We introduce a Temporal Hierarchical Mixer augmented with DCT and LF to efficiently model multi-scale motion from video, a capability largely absent in the original spatial-focused work. (2) We propose a fundamentally lighter architecture, NanoHTNet, which utilizes a novel Efficient Temporal-Spatial Tokenization (ETST) module to drastically reduce model complexity for real-time edge deployment. (3) We introduce PoseCLR, a new contrastive learning pre-training method that captures implicit structural priors from multi-view data, a paradigm not present in the conference version.

II. Related Work

A. Efficient 3D HPE

The deployment of real-time and accurate 3D HPE algorithms on edge devices could significantly promote the widespread application of 3D human vision technology. However, the hardware limitations of these devices impose constraints on model size and computational complexity. To address this challenge, many methods try to enhance efficiency by data optimization. DeciWatch [21] introduces a sample-denoise-recover pipeline, achieving 10× efficiency improvement on edge devices by estimating from less than sampled 10% of video frames. HoT [22], as a plug-and-play framework, utilizes clustering algorithms to select a few representative tokens from redundant video sequences. PoseFormerV2 [5] explores the frequency domain by using a few low-frequency Discrete Cosine Transform (DCT) coefficients to create a compact representation of long skeleton sequences. This approach fuses global frequency-domain features with time-domain features from a few central frames within a specialized Time-Frequency Feature Fusion module. Our NanoHTNet also utilizes DCT and low-pass filtering for data optimization. However, instead of a fusion module, we employ a dual-stream architecture where the filtered frequency features are processed by a dedicated temporal mixer in parallel with a spatial mixer, allowing for specialized extraction of kinematic and physical topology.

Other approaches for efficient HPE are based on the model lightweight. MobileHumanPose [23] significantly reduces model parameters and FLOPs via skip cascading and residuals, facilitating deployment on mobile devices like Galaxy S20. MixSynthFormer [24] incorporates a reduction factor in an MLP-based mixed synthetic attention matrix generator, enabling dynamic fusion of temporal and spatial features and achieving real-time inference on an 8-core CPU Apple M1 Pro. Similarly, DSPNet [25] proposes a deep supervision pyramid network, which utilizes a lightweight up-sampling unit and a novel pyramid architecture to reduce computational load, achieving a remarkable balance between accuracy and efficiency. Our NanoHTNet integrates these methodologies by (1) incorporating DCT in data optimization to filter key features and (2) employing a channel-split progressive architecture and ETST module in model lightweight to reduce model complexity, resulting in real-time inference on Jetson Nano.

B. Human Topology Modeling in 3D HPE

Unlike pixel-based vision tasks, 2D-3D lifting HPE takes human topology composed of sparse and highly structured joints as input, which integrates both the spatial, physical structures and temporal kinematic relationships [26], [27], [28]. Exploiting these inherent priors could boost estimation efficiency.

Many recent works [29], [30], [31] incorporate spatial physical structures from different body levels into the model design. Some studies [32], [33], [34] utilize Graph Convolutional Neural Networks (GCNs) to model joint-level features, employing adjacency matrices to simulate the anatomical connections of joints and aggregate local features. Addressing limb movement inconsistency, other works [29], [30], [31] extract features at the limb level. For example, SRNet [29] splits the joints of the human body into local groups and recombines them. Wang et al. [30] treat body joints unequally and capture the relationships among the body parts with different DOFs. Wu et al. [31] design a limb pose-aware framework through angular relations to provide constraints for the limbs. Furthermore, vision Transformer [35], [36] enable encoding body-level dependencies via self-attention. This approach [24], [37] captures interactions between the whole body, enhancing expressiveness in skeleton sequences. More recently, diffusion models [38], [39], [40] have been used to tackle pose ambiguity by generating multiple topological hypotheses through a denoising process.

Another line of works [18], [41], [42] investigates temporal kinematics for better estimation. Tekin et al. [41] pioneers the integration of video sequences into 3D HPE based on the Seq2Frame pipeline, which only regresses the center frame, effectively leveraging motion information to resolve ambiguities. MixSTE [18] introduces the Seq2Seq framework to the Transformer architecture for 3D HPE. It alternately extracts temporal and spatial relationships and simultaneously outputs 3D poses for all frames, enhancing motion consistency. KTPFormer [42] constructs kinematic topology and motion trajectory separately, which effectively enhances the global correlation modeling capability of human topology. However, these methods can not simultaneously take full advantage of the spatio-temporal priors of human topology, leading to specific estimation errors and significant computational redundancy.

C. Self-Supervised Learning in 3D HPE

Beyond the inherent spatio-temporal priors of human topology, self-supervised learning can further capture high-dimensional and implicit representations of the human skeleton via unlabeled pre-training. Current mainstream self-supervised learning methods can be categorized into generative proxy tasks [43], [44], [45] and contrastive proxy tasks [46], [47], [48], [49]. The former masks a portion of the input sequence and trains the model to predict the masked content. For example, P-STMO [3] employs masked pose modeling to pre-train the model by reconstructing randomly masked joints, thus enhancing the understanding of human body structures. MotionBERT [37] extends the mask autoencoders

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

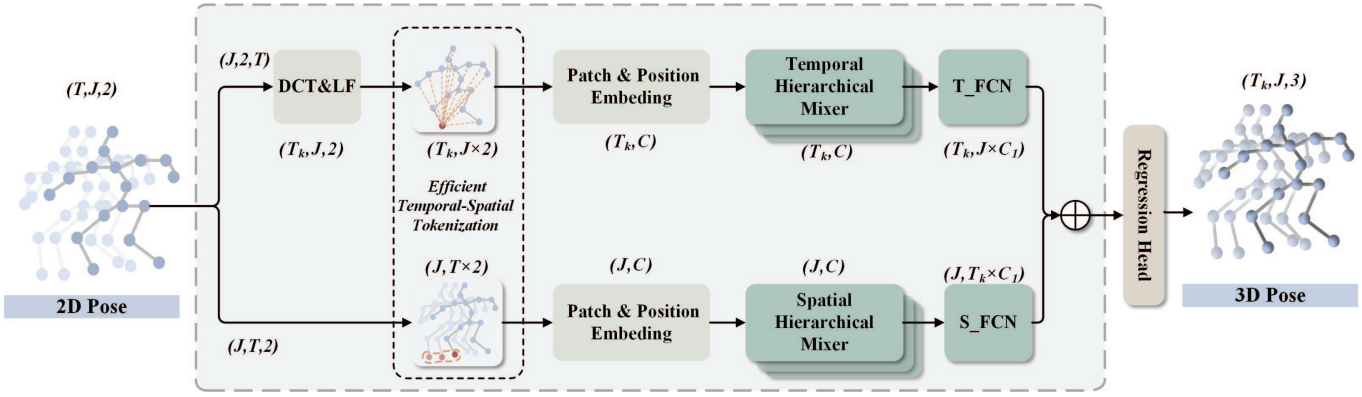


Fig. 3. Overview of our encoding network NanoHTNet. A two-stream structure is designed for independently extracting the spatial-domain physical topology and the temporal-domain kinematic topology of the human body. T denotes input frames, while T_k represents the reduced frames after applying DCT and LF. C is the main channel dimension after patch embedding, and C_l is the channel dimension after MLP networks T_FCN and S_FCN.

framework to human motion representations, utilizing self-supervised pre-training to capture geometric, kinematic, and physical knowledge for multiple downstream tasks.

Contrastive learning captures the intrinsic characteristics of data by learning the relative relationships among samples within the dataset, providing a comprehensive understanding of the human body for skeleton-based tasks [50], [51]. For example, UniHPE [52] aligns images, corresponding 2D poses, and 3D poses as mutual positive samples for training, facilitating constraints and information sharing across different modalities. ORC-Pose [53] uses unoccluded 2D poses as positive samples and occluded ones as negatives to ensure topology invariance, effectively mitigating the occlusion problem. However, these methods overlook the multi-view data in datasets and the cross-view consistent topological patterns of the human topology. In this work, PoseCLR is designed to exploit these relative topological relationships to build contrastive pre-training tasks, thereby enhancing the extraction of topological priors.

III. Methods

Our method consists of two stages, each stage focusing on the extraction of topological features. In the pre-training phase, we implement a contrastive proxy task based on cross-view topological consistency to extract implicit and high-dimensional features without 3D labels; In the fine-tuning phase, we design the encoder architecture to extract explicit hierarchical representations of human topology. Initialized with pre-trained parameters, the model is then fine-tuned using 3D ground truth supervision. In Sec. III-A, we introduce the overview of NanoHTNet as the encoder. Subsequently, Hierarchical Mixers, the core elements of NanoHTNet, are detailed in Sec. III-B. Finally, the contrastive learning-based method PoseCLR is presented in Sec. III-C.

A. Encoding Network Structure

As the encoder for the 3D HPE task, the overall structure of NanoHTNet is shown in Fig. 3. Given a 2D pose sequence $X \in \mathbb{R}^{T\times J\times 2}$ with T frames and J joints, NanoHTNet adopts a tiny dual-stream architecture to extract and fuse spatial and temporal features in parallel.

1) Spatial-Domain: Referencing Transformer-based methods, $X \in \mathbb{R}^{T\times J\times 2}$ is initially reshaped to $X_s \in \mathbb{R}^{J\times(T\cdot 2)}$, allowing each joint to be treated as an individual token. These tokens are mapped to a high dimension $X_s' \in \mathbb{R}^{J\times(T\cdot C)}$ with a linear layer in the patch embedding, where C is the channel dimension. Then, we embed X' with a learnable positional matrix $E_{pos} \in \mathbb{R}^{1\times(T\cdot C)}$ in the position embedding to obtain the embedded features X^l, where l is the index of spatial Hierarchical Mixers. These Hierarchical Mixers, which are described in detail in Sec. III-B, play a crucial role in the extraction of the human topology. Finally, a channel MLP [54] called S_FCN is adopted to aggregate information between different channels.

2) Temporal-Domain: Long sequences, such as 351 frames [2], [55] and 243 frames [4], [37], enhance performance by expanding the receptive field. However, these extended sequences introduce substantial redundancy and lead to a quadratic increase in computational complexity. To mitigate video input redundancy while maintaining a large receptive field, we incorporate a Discrete Cosine Transform (DCT) and a low-pass filter before the patch embedding for data compression and noise reduction. First, we utilize DCT to transform the input data from the time domain to the frequency domain as:

$$X_\phi = c(f)\sum_{t=0}^{T-1} X_t \cos\left[\frac{\pi(2T+1)f}{2T}\right], f = 0, 1, \ldots, T-1, \quad (1)$$

where $X_t' \in \mathbb{R}^{J\times C\times T}$ is the input with T frames, f represents the index of the frequency component, and $c(f)$ is a normalization coefficient used to ensure the orthogonality and energy conservation of DCT. Then, we utilize low-pass filtering to reduce redundancy as:

$$X^k{}_\phi = \begin{cases} X_\phi, & \text{if } f < k \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where k is the cutoff frequency, and $X^k{}_\phi$ represents the frequency domain feature after filtering. Low-frequency coefficients in the frequency domain focus primarily on global movement trends and major poses, while high-frequency coefficients contain noise and minor variations. Through low-pass

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I
COMPLEXITY COMPARISON BETWEEN DIFFERENT TOKENIZATIONS

Method	Domain	Tokens	Complexity
DTST	Temporal	T	$O(J \times T^2 \times C)$
	Spatial	J	$O(T \times J^2 \times C)$
ETST (Ours)	Temporal	T_k	$O(T_k^2 \times C)$
	Spatial	J	$O(J^2 \times C)$

filtering, our model focuses on the main movements and poses, boosting computational efficiency and generalization within a large receptive field. The subsequent steps mirror those in the spatial domain, but each frequency domain feature is treated as a token. These tokens are processed through the Temporal Hierarchical Mixer and T_FCN.

Finally, we concatenate the spatio-temporal representations in the dual-stream network and obtain the 3D pose $Y \in \mathbb{R}^{T_k \times J \times 3}$ through a regression head.

3) Efficient Temporal-Spatial Tokenization: Existing Video Pose Transformers [1], [2], [4], [18], [37] construct tokens independently within spatial and temporal dimensions (Spatially, $X \in \mathbb{R}^{T \times J \times 2} \rightarrow X_s' \in \mathbb{R}^{T \times J \times C}$); Temporally, $X \in \mathbb{R}^{T \times J \times 2} \rightarrow X_t' \in \mathbb{R}^{J \times T \times C}$). These approaches, we termed Decoupled Temporal-Spatio Tokenization (DTST), fail to capture the interaction of spatio-temporal information, and the computational complexity increases quadratically with the number of tokens as shown in Tab. I.

To construct a tiny and more efficient network, we introduce Efficient Temporal-Spatial Tokenization (ETST). Specifically, in the temporal processing stream, $X \in \mathbb{R}^{T \times J \times 2}$ is reshaped to $X_t \in \mathbb{R}^{T \times (J \cdot 2)}$ and subsequently embedded to $X_t' \in \mathbb{R}^{T \times C}$. This enables the model to track the global motion trends of all joints across frames, enhancing spatio-temporal interactions and the understanding of human motion. Similarly, in the spatial stream, $X \in \mathbb{R}^{T \times J \times 2}$ is reshaped to $X_s \in \mathbb{R}^{J \times (T \cdot 2)}$ and then embedded to $X_s' \in \mathbb{R}^{J \times C}$, allowing the model to focus on spatial interactions between joints over T frames.

ETST facilitates the fusion of spatio-temporal information while effectively reducing computational complexity from $O(J^2 \times T \times C) + O(T^2 \times J \times C)$ to $O(J^2 \times C) + O(T^2 \times C)$. By integrating the Low-pass Filter, the computational complexity can be further reduced to $O(J^2 \times C) + O(T_k^2 \times C)$. As the number of temporal input frames increases (like PoseFomer [1] with 243 frames and StridedFormer [55] with 351 frames), the efficiency benefits become more apparent, making it suitable for deployment on resource-constrained edge devices like the Jetson Nano.

B. Hierarchical Mixers

The core of our encoder is composed of stacked Hierarchical Mixers. We design spatial Hierarchical Mixers and temporal Hierarchical Mixers to capture the spatial anatomical topology and temporal kinematic topology, respectively.

1) Spatial Hierarchical Mixers: To learn hierarchical human topology representations, the input X_s is split into three parts: X_{LJC}^s, X_{IPC}^s, and X_{GBI}^s, which have equal dimensions $C' = C/3$ and are fed into corresponding modules: Local Joint-level Connection (LJC), Intra-Part Constraint (IPC), and Global Body-level Interaction (GBI) as shown in Fig. 4. These modules construct the human topology from multiple semantic levels: joint, part, and body.

a) Local joint-level connection: LJC mainly focuses on joint-level features by leveraging GCNs [56] to model the natural relations among joints locally. The graph-structured human skeleton can be defined as $G = (V, A)$, where V is a set of J joints and $A \in \{0,1\}^{J \times J}$ is an adjacency matrix representing the connection between joints. Given the input $X_{LJC}^s \in \mathbb{R}^{J \times C'}$, LJC aggregates neighboring joint features into Y_{LJC}^s, given by:

$$GCN(X_{LJC}^s) = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X_{LJC}^s W, \tag{3}$$

$$Y_{LJC}^s = X_{LJC}^s + GCN(\sigma(GCN(X_{LJC}^s))), \tag{4}$$

where $\tilde{A} = A + I$, $\tilde{D}$ is the diagonal node degree matrix, W is the weight matrix, and $\sigma(\cdot)$ denotes the *GELU* activation function.

b) Intra-part constraint: Neglecting part-level features of the human body often leads to special estimation errors. Taking Fig. 5(a) as an example, the previous state-of-the-art model [57] suffers from significant estimation errors on limb-end joints, and these errors are regular for joints with different Limb Degree of Freedoms (LDoFs), where LDoF denotes the distance between the joints and the torso, *e.g.*, the wrist with 3-LDoF has a larger error than the elbow with 2-LDoF. A novel IPC (see in Fig. 5(b)) is designed, which mainly focuses on the part-level features within each limb through topological constraints. Given the input $\tilde{X}_{IPC}^s = X_{IPC}^s + Y_{LJC}^s$, topological constraints perform in two sets of joints: *(I)* $X_1^s \in \mathbb{R}^{8 \times C'}$ consists of end and 2-LDoF joints; *(II)* $X_2^s \in \mathbb{R}^{12 \times C'}$ consists of all limb joints. Next, X_j^s, $j \in \{1, 2\}$ are fed into two convolution layers to generate limb features $\mathcal{F}_j^s$. Then, a channel MLP [54] is adopted to aggregate information between different channels for each limb feature:

$$\mathcal{F}_j^s = \sigma(Conv_j(X_j^s)), j \in \{1, 2\}, \tag{5}$$

$$\tilde{\mathcal{F}}_j^s = \mathcal{F}_j^s + MLP(LN(\mathcal{F}_j^s)), \tag{6}$$

where $\tilde{\mathcal{F}}_j^s$ are aggregated limb features, $LN(\cdot)$ denotes Layer Normalization, $Conv_1$ and $Conv_2$ are convolution layers with kernel sizes of 2 and 3, respectively. Based on the above $\tilde{\mathcal{F}}_j^s$, we construct two topological constraints $\mathcal{R}_j$ in X_j^s: *(I)* In $\mathcal{R}_1$, 3-LDoF joints in X_1^s are replaced with $\tilde{\mathcal{F}}_1^s$; *(II)* In $\mathcal{R}_2$, 3-LDoF and 2-LDoF joints in X_2^s are replaced with $\tilde{\mathcal{F}}_2^s$. Then, the unprocessed joints are filled into the replaced X_j^s, and Y_{IPC}^s, the final output of IPC, can be represented as follows:

$$Y_{IPC}^s = \tilde{X}_{IPC}^s + \mathcal{R}_1(\tilde{X}_{IPC}^s, \tilde{\mathcal{F}}_1^s) + \mathcal{R}_2(\tilde{X}_{IPC}^s, \tilde{\mathcal{F}}_2^s). \tag{7}$$

The aggregated limb feature $\tilde{\mathcal{F}}_j^s$ provides a holistic and more stable representation of the limb's motion. The Replacement operation $\mathcal{R}_j$ uses this robust feature to constrain the error-prone distal joints, thereby severing the error propagation path. Finally, the residual connection allows the network to learn a corrective update based on this topological prior, leading to a more kinematically consistent and accurate pose estimation.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

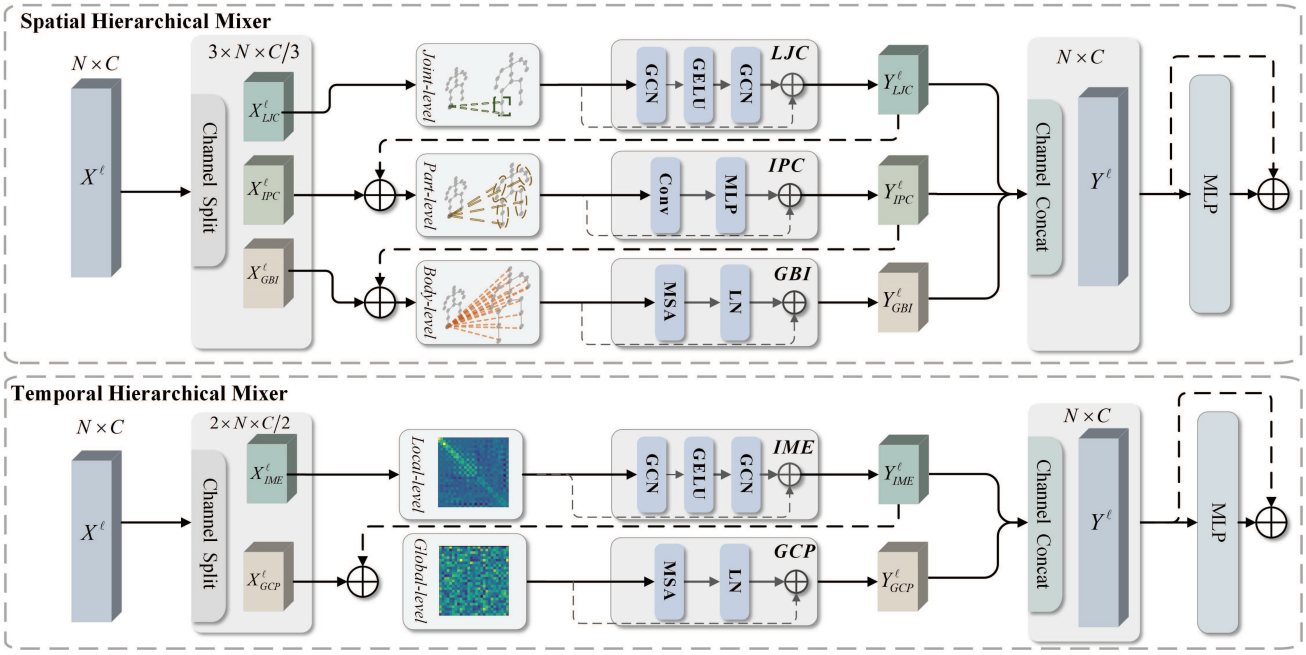


Fig. 4. Stacked Hierarchical Mixers are key in our encoder. Each Spatial Hierarchical Mixer consists of Local Joint-level Connection (LJC), Intra-Part Constraint (IPC), and Global Body-level Interaction (GBI). Each Temporal Hierarchical Mixer effectively captures temporal features from local to global extents by Instantaneous Motion Extraction (IME) and Global Coherence Perception (GCP).

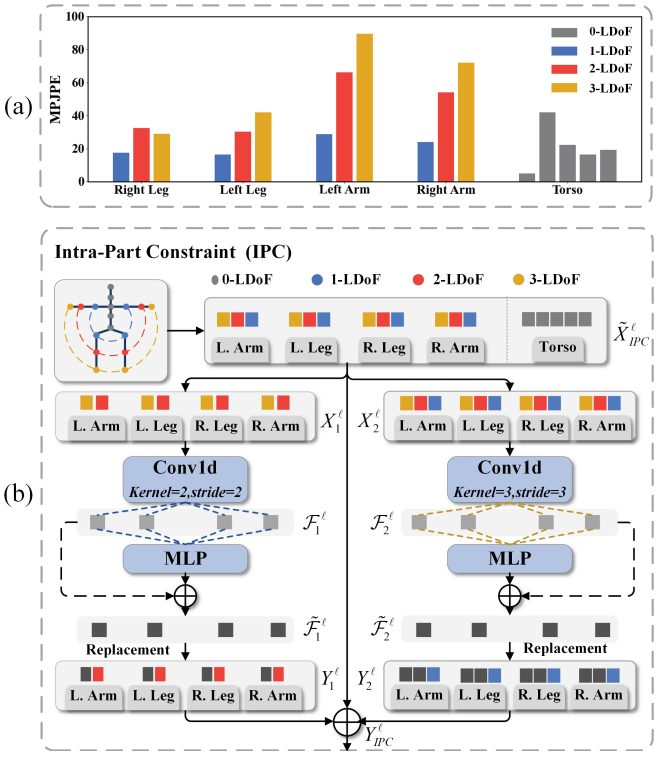


Fig. 5. (a) Specific estimation error distribution caused by neglect of part-level features; (b) Structure of Intra-Part Constraint (IPC) module.

c) Global body-level interaction: GBI is built by the self-attention of Transformer, which can capture global contexts across the whole human body [2], [17]. Given $\tilde{X}^s_{GBI} = X^s_{GBI} + Y^s_{IPC}$ as input, the GBI module can be formulated as:

$$H_i = Softmax(Q^s K^{sT} / \sqrt{C'}) V^s, i \in \{1, \ldots, h\} \quad (8)$$

$$MSA(\tilde{X}^s_{GBI}) = \text{Concat}(H_1, H_2, \ldots, H_h) W_{out}, \quad (9)$$

$$\tilde{Y}^s_{GBI} = \tilde{X}^s_{GBI} + LN(MSA(\tilde{X}^s_{GBI})), \quad (10)$$

where h is the number of attention heads, Q^s, K^s, V^s are query, key, and value matrices, which are calculated from $\tilde{X}^s_{GBI}$ by linear transformations, $\tilde{Y}^s_{GBI}$ is the output of the GBI module.

2) Temporal Hierarchical Mixers: To effectively extract temporal multi-scale kinematic features, we design the temporal Hierarchical Mixers. This design includes two components: *(i)* Instantaneous Motion Extraction (IME) module, based on graph convolutional networks, captures local temporal features such as acceleration and velocity between adjacent frames; *(ii)* Global Coherence Perception (GCP) module, based on self-attention mechanisms, learns periodic changes and inherent coherence of motion, thereby fully leveraging the kinematic features within long-term sequences.

Significantly, these mixers adopt a channel-split progressive architecture rather than a simple serial or parallel structure for efficient feature extraction. Specifically, channels are split into several parts and allocated to distinct sub-modules (*e.g.*, LJC, IPC, GBI, IME, and GCP). Meanwhile, the output from the previous sub-module is input into the subsequent module via residual-like connections, learning the human topological features from local to global. Finally, the output of each sub-module is concatenated and fed into an MLP network for aggregation. To summarize, Hierarchical Mixers can aggregate features from different levels and obtain a more expressive representation, achieving a balance between feature extraction and computational efficiency.

C. PoseCLR

In addition to the above explicit extraction of kinematic and physical features of the human body, we observe that

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

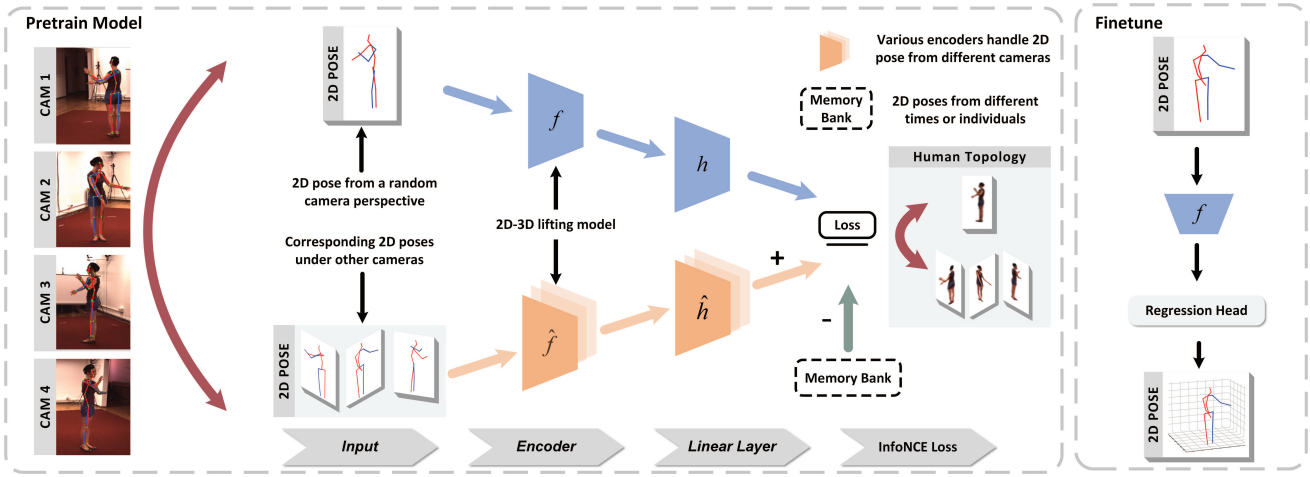


Fig. 6. PoseCLR is a contrastive learning architecture based on cross-view topological consistency. 2D poses of the same individual captured simultaneously from different cameras are treated as positive pairs, while other 2D poses as negative pairs.

human topology exhibits consistency across multiple views simultaneously. Building on this insight, we propose PoseCLR, a multi-view contrastive learning framework for implicit extraction of human topological features (see Fig. 6). This self-supervised learning approach consists of pre-training and fine-tuning stages, extracting universal skeletal priors from unlabeled multi-view 2D poses for more accurate 3D HPE.

1) Pre-training Stage: During the pre-training stage, we utilize a memory-augmented architecture [47], [50] to align 2D poses from different camera views. The goal is to maximize the similarity between 2D poses of the same individual captured simultaneously from various camera angles, as these poses correspond to the same human topology in world coordinates. We take the Human3.6M dataset for illustration, which includes four camera viewpoints. In each iteration, we select a 2D pose from a random viewpoint as the input into encoder f. We employ NanoHTNet in Sec. III-A as the encoder, which can be any 2D-to-3D lifting network, to encode the input into universal high-dimensional human representations h. Meanwhile, the poses from all four viewpoints are treated as positive pairs and inputted into $\hat{f}$. Specifically, four $\hat{f}$ are the momentum updated version of f for four distinct views, encoding these inputs into the topologic representations $\hat{h}$. Subsequently, we employ the InfoNCE loss [59] to minimize the distance between h and $\hat{s}$ representing the same topology, while maximizing the distance between features of different topologies:

$$\mathcal{L}_p = -\log \frac{\exp(\mathcal{S}(q,k^+)/\tau)}{\exp(\mathcal{S}(q,k^+)/\tau) + \sum_{i=0}^{K} \exp(\mathcal{S}(q,k_i^-)/\tau)}, \quad (11)$$

where q represents the feature vector, k^+ denotes the positive samples, k_i^- denotes the negative samples, $\mathcal{S}$ is the similarity function, τ is the temperature parameter (a tunable hyperparameter) used to control the scale of the similarity scores, and K indicates the number of negative samples. Notably, the positive samples are subsequently stored in a first-in-first-out memory bank, where they will serve as negative samples in future iterations.

2) Finetuning Stage: The pre-training stage emphasizes generic human body representations, while the finetuning stage focuses on representations more aligned with the 3D HPE task. Unlike previous networks with randomly initialized parameters, we load encoder parameters f from the pre-training stage, which benefits the network by: 1) Improving performance. Pre-training enables the model to capture high-dimensional human topological prior, further assisting in the comprehension of human-based inputs; 2) Accelerating convergence. The model does not require learning from scratch but refines upon pre-learned features, which substantially reduces the training time. In this phase, we introduce 3D pose labels and employ the simple Mean Per Joint Position Error (MPJPE) loss for supervised training:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| Y_j^t - \tilde{X}_j^t \right\|_2, \quad (12)$$

where X_j^t and Y_j^t are the predicted and ground truth 3D poses of j joint at t frame.

IV. Experiments

A. Datasets and Evaluation

Human3.6M [19] is the most widely used indoor dataset in the 3D HPE. There are 3.6 million images, 11 professional actors, and 15 different actions in it. Following [17], [29], [56], 5 subjects ($S1, S5, S6, S7, S8$) are used for training and 2 subjects ($S9, S11$) are used for testing. For the metrics, we adopt the mean per joint position error (MPJPE) and Procrustes MPJPE (P-MPJPE). The MPJPE (*i.e.*, Protocol 1) measures the Euclidean distance between the ground truth and estimated joints; the P-MPJPE (*i.e.*, Protocol 2) is the MPJPE after aligning the predicted results to the ground truth by translation, rotation, and scale.

MPI-INF-3DHP [69] is also a challenging dataset consisting of 1.3 million images from indoor and outdoor scenes. There are three different scenes in its test set: studio with a green screen (GS), studio without green screen (noGS), and outdoor scene (Outdoor). Following [17], [56], and [67],

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

TABLE II

QUANTITATIVE COMPARISONS WITH VIDEO-BASED METHODS ON HUMAN3.6M UNDER MPJPE AND P-MPJPE. FRAMES DENOTE THE NUMBER OF FRAMES INPUT. RF IS THE EFFECTIVE RECEPTIVE FIELD OF THE INPUT SEQUENCE.FLOPS (FLOATING POINT OPERATIONS PER SECOND) QUANTIFY COMPUTATIONAL COMPLEXITY. FPS (FRAME PER SECOND) IS IMPLEMENTED ON JETSON NANO

Method	Seq2Seq	Frames	RF	Param (M)	FLOPs (M)	FPS	MPJPE↓	P-MPJPE↓
PoseFormer [1]	✗	27	27	9.59	452	3.53	47.0	-
StridedFormer [55]	✗	27	27	4.01	128	5.36	46.9	-
StridedFormer [55]	✗	81	81	4.06	392	4.17	45.9	-
P-STMO [3]	✗	27	27	4.60	82	7.04	48.2	-
P-STMO [3]	✗	27	243	4.60	82	6.81	46.1	-
MHFormer [2]	✗	27	27	18.92	1031	2.43	45.9	36.1
MixSTE [18]	✓	3	3	33.65	3420	5.67	49.6	38.9
PoseFormerV2 [5]	✗	9	81	14.35	352	5.55	46.0	36.1
PoseFormerV2 [5]	✗	27	243	32.35	1055	2.30	45.2	35.6
STCFormer [4]	✓	27	27	4.75	4346	9.26	44.1	34.7
MotionAGFormer-XS [58]	✓	27	27	2.20	1937	19.2	45.1	36.9
HTNet-L (Conf. ver.) [20]	✓	27	27	10.1	631	5.12	46.1	36.0
NanoHTNet (Ours)	✓	9	243	**1.52**	**33**	54.81	44.5	35.2
NanoHTNet-L (Ours)	✓	27	243	4.40	156	**112.76**	**43.5**	**34.5**

we report metrics of MPJPE, Percentage of Correct Keypoint (PCK), and Area Under Curve (AUC).

B. Implementation Details

1) Pre-traing Stage: PoseCLR is pre-trained on the Human3.6M dataset using one NVIDIA RTX 4090 GPU. The 50Hz camera frequency of this dataset means that chronologically adjacent frames are highly correlated. If sampled sequentially, these frames would populate the FIFO memory bank with overly similar false negative samples, degrading the contrastive task. To mitigate this sampling bias, PoseCLR samples frames at equal intervals of $slice(s)$ to ensure the diversity and quality of negative samples in the memory bank. We employ a learning rate of 0.001 and train 200 epochs for effective extraction of human topological prior."

2) Fintuning Stage: With pre-trained parameters, NanoHTNet converges at the optimal performance in 2 epochs, compared to 20 epochs when trained from scratch. The proposed NanoHTNet consists of 3 stacked spatial and temporal hierarchical mixers. Notably, the channel dimension must be a multiple of 48 (*e.g.*, 144, 240, 384) because there are eight heads in the MSA, three split channels in the spatial Hierarchical Mixers, and two split channels in the temporal Hierarchical Mixers. For a fair comparison, horizontal flip augmentation is adopted referring to [1], [18], and [55]. Refinement module is only used in imaged-based comparisons to explore the best performance following our previous work [20].

3) Edge Device Deployment: To accommodate deployment on edge devices, we evaluate our NanoHTNet on the Jetson Nano. Jetson Nano is equipped with a quad-core ARM Cortex-A57 CPU, 4GB memory, and NVIDIA Maxwell with 128 CUDA cores and 0.5 TFLOPs (FP16), offering a small and compact hardware setup ideal for integration into robots. However, this also imposes specific requirements on the parameters and computational complexity of the algorithms deployed on it.

C. Comparison With State-of-the-Art Results

1) Compared With Video-Based Methods: As shown in Tab. II, existing Seq2Frame and Seq2Seq approaches fail to reach real-time speed (30 FPS) on Jetson Nano. For the same video sequence input, Seq2Frame methods output only intermediate frames, achieving better performance but slower inference. Seq2Seq methods match the input length, offering higher efficiency but lower performance. To be competitive, Seq2Seq methods often require extensive computation (MixSTE, and STCFormer reach 3420 and 4346 MFLOPs, respectively). In this case, our NanoHTNet based on Seq2Seq pipeline leverages spatio-temporal priors to enhance structural understanding and significantly improves efficiency in processing skeleton data. Despite having only 1.52MB in parameters and 33 MFLOPS, NanaHTNet achieves impressive accuracies of 44.5*mm* MPJPE and 35.2*mm* P-MPJPE. By extending input frames (9→27) and channel dimension (240→384), NanoHTNet-L outperforms other State Of The Art (SOTA) methods in the speed-accuracy trade-off, achieving 43.5*mm* MPJPE, 34.5*mm* P-MPJPE with only 4.4MB parameters and 156 MFLOPs. Moreover, its inference speed on Jetson Nano further reach 112.76 FPS, satisfying real-time inference needs.

2) Compared With Image-base Methods: We compare our NanoHTNet with state-of-the-art image-based methods on Human3.6M. As shown in Tab. III (top part), the 2D keypoints from the 2D detector CPN [14] are used as input, and our NanoHTNet achieves the best results under both MPJPE (47.6*mm*) and P-MPJPE (38.6*mm*). Moreover, MPJPE can be further improved to 47.1*mm* under MPJPE and 38.0*mm* under P-MPJPE after introducing PoseCLR. In particular, we can observe that NanoHTNet with PoseCLR obtains the best performance in all kinds of action in MPJPE. As shown in Tab. III (bottom part), NanoHTNet with PoseCLR also reaches

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

TABLE III

QUANTITATIVE COMPARISONS WITH IMAGE-BASED METHODS ON HUMAN3.6M UNDER PROTOCOL #1 (MPJPE) AND PROTOCOL #2 (P-MPJPE). 2D KEYPOINTS DETECTED BY CPN ARE USED AS INPUTS IN THE TOP AND MIDDLE PART OF THE TABLE, AND THE GROUND TRUTH OF 2D POSES ARE USED AS INPUTS IN THE BOTTOM PART OF THE TABLE

Protocol #1 (CPN)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez *et al.*. [9]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Fang *et al.* [60]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
LCN [61]	46.8	52.3	44.7	50.4	52.9	68.9	49.6	46.4	60.2	78.9	51.2	50.0	54.8	40.4	43.3	52.7
ST-GCN [33]	46.5	48.8	47.6	50.9	52.9	61.3	48.3	45.8	59.2	64.4	51.2	48.4	53.5	39.2	41.2	50.6
UGRN [62]	47.9	50.0	47.1	51.3	51.2	59.5	48.7	46.9	56.0	61.9	51.1	48.9	54.3	40.0	42.9	50.5
SRNet [29]	44.5	48.2	47.1	47.8	51.2	56.8	50.1	45.6	59.9	66.4	52.1	45.3	54.2	39.1	40.3	49.9
MLP-JCG [63]	43.8	**46.7**	46.9	48.9	50.3	60.1	**45.7**	43.9	56.0	73.7	48.9	48.2	50.9	39.9	41.5	49.7
Gong *et al.* [38]	42.8	49.1	45.2	48.7	52.1	63.5	46.3	45.2	58.6	66.3	50.4	47.6	52.0	37.6	40.2	49.7
PGFormer [64]	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
MGCN [56]	45.4	49.2	45.7	49.4	50.4	58.2	47.9	46.0	57.5	63.0	49.7	46.6	52.2	38.9	40.8	49.4
RS-Net [65]	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
HopFIR [57]	43.9	47.6	45.5	48.9	50.1	58.0	46.2	44.5	55.7	62.9	49.0	45.8	51.8	38.0	39.9	48.5
GraphMLP [66]	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
NanoHTNet (Ours)	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
NanoHTNet + PoseCLR (Ours)	**42.2**	47.4	**43.6**	**47.1**	**48.7**	**55.1**	45.9	**43.5**	**55.0**	61.0	**46.9**	**45.0**	**50.0**	37.5	**38.4**	**47.1**
Protocol #2 (CPN)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Fang *et al.* [60]	38.2	41.7	43.7	44.9	48.5	55.3	40.2	38.2	54.5	64.4	47.2	44.3	47.3	36.7	41.7	45.7
LCN [61]	36.9	41.6	38.0	41.0	41.9	51.1	38.2	37.6	49.1	62.1	43.1	39.9	43.5	32.2	37.0	42.2
Liu *et al.* [67]	35.9	40.0	38.0	41.5	42.5	51.4	37.8	36.0	48.6	56.6	41.8	38.3	42.7	31.7	36.2	41.2
ST-GCN [33]	36.8	38.7	38.2	41.7	40.7	46.8	37.9	35.6	47.6	51.7	41.3	36.8	42.7	31.0	34.7	40.2
SRNet [29]	35.8	39.2	36.6	**36.9**	39.8	45.1	38.4	36.9	47.7	54.4	**38.6**	36.3	**39.4**	30.3	35.4	39.4
MLP-JCG [63]	**33.7**	37.4	37.3	39.6	39.8	47.1	**33.7**	**33.8**	45.7	60.5	39.7	37.7	40.1	30.1	33.8	39.3
MGCN [56]	35.7	38.6	36.3	40.5	39.2	44.5	37.0	35.4	46.4	51.2	40.5	35.6	41.7	30.7	33.9	39.1
RS-Net [65]	35.5	38.3	**36.1**	40.5	39.2	44.8	37.1	34.9	45.0	49.1	40.2	35.4	41.5	31.0	34.3	38.9
GraphMLP [66]	35.1	38.2	36.5	39.8	39.8	43.5	35.7	34.0	45.6	**47.6**	39.8	35.1	41.1	**30.0**	**33.4**	38.4
NanoHTNet (Ours)	35.0	38.4	36.2	39.7	39.3	43.7	36.3	34.5	44.9	51.3	39.6	35.0	41.3	30.9	33.4	38.6
NanoHTNet + PoseCLR (Ours)	34.7	**37.2**	36.2	38.9	**38.4**	**42.5**	36.1	34.2	**43.9**	50.0	39.3	**34.8**	40.4	30.8	**33.4**	**38.0**
Protocol #1 (GT)	Dir.	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
SemGCN [32]	37.8	49.4	37.6	40.9	45.1	41.4	40.1	48.3	50.1	42.2	53.5	44.3	40.5	47.3	39.0	43.8
ST-GCN [33]	33.4	39.0	33.8	37.0	38.1	47.3	39.5	37.3	43.2	46.2	37.7	38.0	38.6	30.4	32.1	38.1
Liu *et al.* [67]	36.8	40.3	33.0	36.3	37.5	45.0	39.7	34.9	40.3	47.7	37.4	38.5	38.6	29.6	32.0	37.8
GraFormer [17]	35.2	38.0	30.4	34.4	34.7	43.3	35.2	31.4	38.0	46.2	34.2	35.7	36.1	27.4	30.6	35.2
MLP-JCG [63]	**29.1**	36.0	30.4	33.8	35.5	46.5	35.3	31.2	39.2	48.8	33.9	35.2	35.8	26.9	29.4	35.1
PHGANet [68]	32.4	36.5	30.1	33.3	36.3	43.5	36.1	30.5	37.5	45.3	33.8	35.1	35.3	27.5	30.2	34.9
PGFormer [64]	32.3	36.2	30.5	32.6	33.8	42.4	36.3	29.1	36.0	43.9	33.9	34.5	35.1	26.6	29.6	34.2
GraphMLP [66]	32.2	38.2	29.3	33.4	33.5	38.1	38.2	31.7	37.3	38.5	34.2	36.1	35.5	28.0	29.3	34.2
HopFIR [57]	31.3	**34.0**	28.0	32.0	33.1	42.1	34.1	**28.1**	33.6	39.8	31.7	**32.9**	33.8	**26.7**	**28.9**	32.7
Azizi *et al.* [12]	30.1	35.3	30.6	**27.6**	36.2	**38.4**	**30.7**	30.3	35.9	40.7	32.9	34.9	35.2	27.2	32.0	32.7
NanoHTNet(Ours)	33.5	36.0	27.8	32.4	31.8	40.1	34.6	28.8	33.9	35.1	31.7	34.2	32.8	27.6	29.1	31.9
NanoHTNet + PoseCLR (Ours)	33.3	35.3	**27.6**	31.9	**31.7**	39.6	33.1	28.7	**33.3**	**35.0**	**31.4**	34.0	**32.5**	27.6	**28.9**	**31.6**

the best performance under MPJPE (31.9*mm*) using ground truth 2D joints, outperforming [57] by 1.1*mm*.

Methods with PoseCLR. To validate the generality and effectiveness of PoseCLR as a pre-training approach in 3D HPE task, we employ three image-based methods (SemGCN [32], MGCN [56], GraFormer [17]), three video-based methods (MixSTE [18], PoseFormerV2 [5], STCFormer [4]), and our NanoHTNet as the encoder for training and testing. As indicated in Table IV, the introduction of PoseCLR leads to further improvements in the MPJPE metric across all methods (*e.g.*, 3.67% increase in the MPJPE metric for GraFormer after incorporating PoseCLR). Furthermore, the training time of finetuning is approximately one-tenth of that required for training from scratch. This indicates that pre-training with PoseCLR enables better initialization parameters embedded with inherent priors of human body topology, which significantly enhance the accuracy and robustness of understanding the human body for 3D HPE models.

Comparison on MPI-INF-3DHP. To evaluate the robustness of our NanoHTNet and PoseCLR, we only train the model on the Human3.6M and test it on more challenging MPI-INF-3DHP directly, without any extra training and refinement. As shown in Tab. V, our NanoHTNet achieves the best performance on all various scenes. Additionally, loading pre-trained parameters from Human3.6M into models (MGCN and NanoHTNet) facilitates further performance enhancements. This verifies the strong generalization of proposed NanoHTNet and PoseCLR in unseen environments.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

QUANTITATIVE COMPARISON OF SOTA METHODS WITH POSECLR ON HUMAN3.6M UNDER MPJPE AND P-MPJPE

	Method	MPJPE (↓)	P-MPJPE(↓)
Image input	SemGCN [32]	60.8	-
	SemGCN + PoseCLR	58.7 (↓1.1)	-
	MGCN [56]	49.4	39.1
	MGCN + PoseCLR	48.5 (↓0.9)	38.4 (↓0.7)
	GraFormer [17]	51.8	-
	GraFormer + PoseCLR	49.9 (↓1.9)	-
Video input	MixSTE(T=3) [18]	49.6	38.9
	MixSTE + PoseCLR	48.7 (↓0.9)	38.3 (↓0.6)
	PoseFormerV2 (T=27, RF=243) [5]	45.2	35.6
	PoseFormerV2 + PoseCLR	44.8 (↓0.4)	35.1 (↓0.5)
	STCFormer (T=27) [4]	44.1	34.7
	STCFormer + PoseCLR	43.5 (↓0.6)	34.3 (↓0.4)
	NanoHTNet (Ours, T=27, RF=243)	43.5	34.5
	NanoHTNet + PoseCLR	43.0 (↓0.5)	34.1 (↓0.4)

TABLE V

QUANTITATIVE COMPARISON ON MPI-INF-3DHP WITH FIVE METRICS. HERE, ↑ INDICATES THE HIGHER, THE BETTER

Methods	PCK				AUC
	GS	no GS	Outdoor	All	
Martinez *et al.* [9]	49.8	42.5	31.2	42.5	17.0
LCN [61]	74.8	70.8	77.3	74.0	36.7
Zhou *et al.* [70]	75.6	71.3	80.3	75.3	38.0
SRNet [29]	-	-	80.3	77.6	43.8
Liu *et al.* [67]	77.6	80.5	80.1	79.3	45.8
Zeng *et al.* [71]	-	-	84.6	82.1	46.2
Xu *et al.* [72]	81.5	81.7	75.2	80.1	45.8
PGFormer [64]	84.4	84.5	77.4	83.9	52.3
UGRN [62]	86.2	84.7	81.9	84.1	53.7
RS-Net [65]	-	-	-	85.6	53.2
MGCN [56]	86.4	86.0	85.7	86.1	53.7
MGCN + PoseCLR	87.1	86.6	86.7	86.4	54.4
GraphMLP [66]	87.3	87.1	86.3	87.0	54.3
HopFIR [57]	89.1	85.9	85.9	87.2	**57.0**
NanoHTNet(Ours)	88.9	86.2	85.9	86.7	56.1
NanoHTNet + PoseCLR	**89.4**	**86.6**	**86.1**	**88.1**	**57.0**

D. Ablation Study

1) Impact of Model Components: Given the similar structure of the Spatial and Temporal Hierarchical Mixers, Spatial Hierarchical Mixer is selected as a representative to investigate the impact of model components. As shown in the top part of Tab. VI, NanoHTNet with only one single-level module cannot achieve satisfactory performance, while the combination of all proposed modules performs the best via learning representations from various levels. Notably, IPC cannot work individually due to no operations on 0-LDoF and 1-LDoF joints. To further investigate the influence of IPC, we categorize the limb joints by LDoFs and calculate the average MPJPE of joints in each category in Fig. 7. It is obvious to find that the primary reduction in estimation errors originates from distal limb joints (MPJPE on 1-LDoF, 2-LDoF and 3-LDoF joints noticeably reduces by 2.3%, 8.7% and 12.3%, respectively), which can be attributed to the increased number of topological constraints that these joints receive as detailed in Sec. III-B: *(i)* 3-LDoF joints are constrained by both 1-LDoF and 2-LDoF joints; *(ii)* 2-LDoF joints are only constrained by 1-LDoF joints; *(iii)* 1-LDoF joints are not subject to any constraints. Moreover, we observe that model parameters significantly reduce via the introduction of the ETST (31.75M → 1.52M) while maintaining performance, which is highly beneficial for deployment on edge devices. In summary, the efficiency of NanoHTNet is attributed to two key factors: (1) ETST with efficient data mapping significantly reduces parameters and computational complexity; (2) The hierarchical design could capture multi-level spatio-temporal priors, enabling effective understanding of the human body and superior estimation with smaller model size.

2) Impact of Model Structures: As for the structures (middle part of Tab. VI), the serial structure exhibits competitive

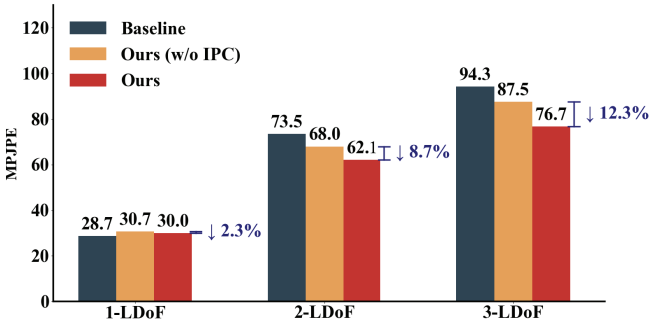


Fig. 7. Ablation study for the effect of IPC module.

TABLE VI

ABLATION STUDY FOR COMPONENTS AND STRUCTURES. THE EVALUATION IS PERFORMED ON HUMAN3.6M. WE SET NANOHTNET WITH ONLY GBI SUB-MODULE AND ETST AS BASELINE

	LJC	IPC	GBI	ETST	Params (M)	FLOPS (M)	MPJPE
Baseline			✓		28.62	762.2	47.4
			✓	✓	1.25	52.1	47.1
	✓			✓	1.47	39.7	46.2
	✓		✓	✓	1.43	45.9	45.9
		✓	✓	✓	1.74	39.6	46.5
	✓	✓		✓	1.92	30.1	45.0
	✓	✓	✓		31.75	501.2	44.1
Parallel	✓	✓	✓	✓	1.52	33.4	46.0
Parallel	✓	✓	✓		31.75	501.2	45.2
Serial	✓	✓	✓	✓	4.07	107.7	44.7
Serial	✓	✓	✓		39.58	612.9	44.3
NanoHTNet	✓	✓	✓	✓	1.52	33.4	44.5

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

ABLATION STUDY ON RF (RECEPTIVE FIELD, INPUT FRAMES BEFORE DCT AND LF) AND FRAMES (INPUT FRAMES AFTER DCT AND LF). THE EVALUATION IS IMPLEMENTED ON HUMAN3.6M

Frames (t)	RF (n)	Params (M)	FLOPs (M)	MPJPE
1	9	1.39	19.6	47.3
3	27	1.40	22.4	46.8
9	27	1.42	29.9	46.4
9	81	1.44	30.7	45.2
27	81	1.48	53.2	45.0
9	243	1.52	33.4	44.5
27	243	1.56	55.8	44.5
81	243	1.66	123.2	45.1

performance (46.0*mm*) due to the local-to-global learning; the parallel structure can reduce parameters from 4.07M to 1.52M due to the channel-split structure. The channel-split progressive structure of NanoHTNet adopts residual connections and combines the advantages of these two designs, which performs the best (44.5*mm*) and maintains a small model size (1.52M).

3) Impact of Frames and Receptive Field: As illustrated in Fig. 7, the receptive field significantly affects the overall performance by capturing extensive spatio-temporal context, while input frames determine the parameters and FLOPs. In this work, the introduction of ETST enables NanoHTNet to maintain a tiny model size, and frequency low-pass filtering prevents size boosting as the receptive field expands. For example, when the receptive field increases (27→81) with the same input length (9 frames), NanoHTNet achieves 1.2*mm* improvement under MPJPE while the parameters and FLOPs merely increase 0.02M (1.42M→1.44M) and 0.8M (29.9M→30.7M), respectively. Notably, with larger receptive fields and more input frames, the performance bottleneck shifts towards the model size, prompting further exploration into how these hyperparameters impact performance.

4) Impact of Hyperparameters in PoseCLR: We evaluate different hyperparameter settings of PoseCLR in VIII. We observe that: (i) More views (s = 4) are crucial for pre-training; (ii) The increase of memory bank β brings limited improvement to the MPJPE metric, thus β is set to 32,768 to balance memory usage and performance; (iii) Sampling interval s effectively mitigates the noise from adjacent frames as negative samples, yet a large interval reduces the training database. PoseCLR performs optimally at s = 3; (iv) The momentum update decay rate τ is optimal at 0.999, as high or low values degrade performance by disrupting learning stability.

5) Impact of Hyperparameters in NanoHTNet: As shown in Tab. IX, we explore the hyperparameter settings of NanoHTNet. It can be seen that the channel dimension C and layers L of the hierarchical mixing block have a significant impact on the performance, and the combination of C = 240 and L = 3 works best. The reason for this is the inadequacy of parameters in NanoHTNet for capturing the complex topological features of the human body, leading to underfitting, while an overly large model tends to overfit the training data, focusing on noise instead of achieving generalization.

TABLE VIII

ABLATION STUDY ON HYPERPARAMETERS OF POSECLR. VIEWS (α) IS THE NUMBER OF CAMERAS. MEMORY (β) IS THE SIZE OF THE MEMORY BANK. SLICE (s) IS THE SAMPLING INTERVAL. DECAY (τ) IS THE RATE OF MOMENTUM UPDATE

Views(α)	Memory Bank(β)	Slice(s)	Decay(τ)	MPJPE (↓)
4	32768	3	0.999	43.7
3	32768	3	0.999	44.1
2	32768	3	0.999	44.5
4	65536	3	0.999	43.7
4	16384	3	0.999	43.8
4	8192	3	0.999	44.0
4	32768	1	0.999	45.1
4	32768	2	0.999	43.9
4	32768	4	0.999	44.2
4	32768	3	1.0	45.3
4	32768	3	0.99	43.8
4	32768	3	0.9	44.3

TABLE IX

ABLATION STUDY ON HYPERPARAMETERS OF NANOHTNET. LAYERS REFER TO THE NUMBER OF STACKED MIXERS. DIMENSIONS REFER TO THE INPUT CHANNEL. THE EVALUATION IS IMPLEMENTED ON HUMAN3.6M

Layers (L)	Dimensions (C)	Params (M)	FLOPs (M)	MPJPE
3	240	1.52	33.4	44.5
2	240	1.07	24.0	45.1
4	240	1.97	44.7	44.8
3	144	0.83	18.4	45.6
3	384	2.76	60.6	44.7

E. Visualization

1) Attention Visualization: In order to illustrate the effectiveness of Hierarchical Mixers, we visualize the adjacency matrix from LJC and IME, along with multi-head self-attention maps from GBI and GCP. It can be observed in Fig. 8(a) that the joints (*e.g.*, Joint 3, Right Ankle) primarily pay attention to themselves and their physically connected ones. This demonstrates that LJC in NanoHTNet mainly focuses on the most related local region. In addition, Fig. 8(b) shows that the limb joints (*e.g.*, Joint 3, Right Ankle) concentrate on the other limb joints across limbs, which demonstrates the GBI mainly learns the less adjacent global context among joints. Similarly, Fig. 8(c) demonstrates that in the IME module, each frame focuses on the local instantaneous features of adjacent frames (*e.g.*, the attention in frame 7 is primarily on frames 8 and 9). Fig. 8(d) illustrates that each frame pays more attention to the global coherence and coordination of actions in the GCP module (*e.g.*, the attention in frame 15 is on frame 25 due to

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

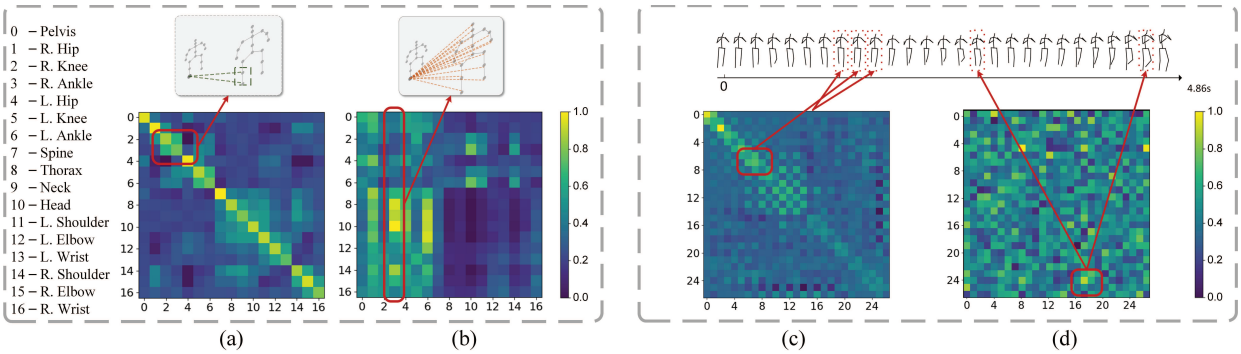


Fig. 8. (a) (b) Visualization of adjacency matrix from LJC and self-attention maps from GBI. The x-axis and y-axis correspond to the query of 17 joints. (c) (d) Visualization of adjacency matrix from IME and self-attention maps from GCP. The x-axis and y-axis correspond to the query of 27 frames. Yellow indicates stronger attention. Attention values are normalized from 0 to 1.

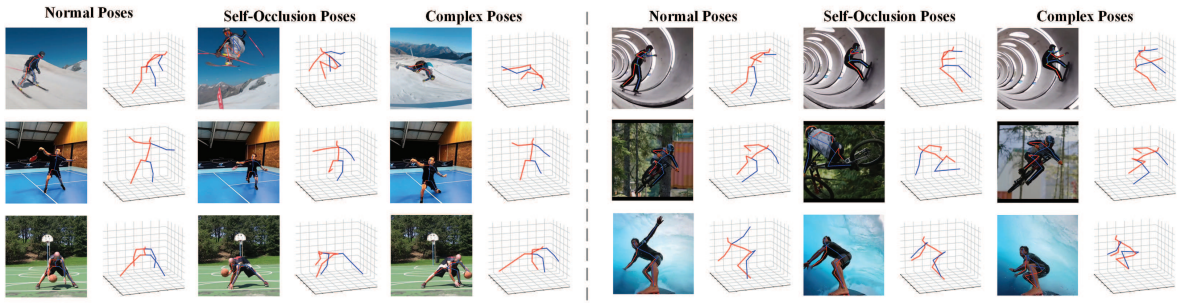


Fig. 9. Qualitative results of NanoHTNet in wild scenes.

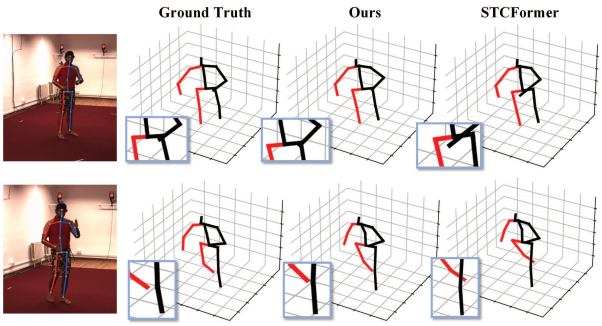


Fig. 10. Qualitative comparison between our method, previous SOTA method, and ground truth on Human3.6M.

the cyclical repetition of similar actions). Overall, these results prove that our Hierarchical Mixers can effectively extract multiple semantic level features according to the spatio-temporal hierarchy of the human topology. Additionally, different submodules sequentially extract features at specific target levels, which aligns well with our original design intent.

2) Visualization in Wild Scenes: As illustrated in Fig. 9, the following indoor and outdoor sports are selected from online video websites for evaluation: alpine skiing, badminton, basketball, skateboarding, mountain biking, and surfing. Although some complex and self-occlusion poses in these images are unseen or rare on Human3.6M, our NanoHTNet can produce accurate and plausible 3D poses owning to the comprehensive understanding of human spatial-temporal topology.

3) Qualitative Comparison of Visualization: Fig. 10 provides qualitative comparisons on Human3.6M between NanoHTNet and the state-of-the-art approach, *i.e.*, STCFormer [4]. It can be seen that our method produces more precise 3D keypoints at the ends of limbs, demonstrating its effectiveness.

V. Conclusion

In this work, we leverage spatio-temporal priors of human body to improve model efficiency in 3D human pose estimation. First, we propose a nano network namely NanoHTNet based on the explicit human features, which captures spatial hierarchical features and temporal multi-scale kinematics of the human topology. Furthermore, we integrate Efficient Temporal-Spatial Tokenization and Low-pass Filtering to achieve more efficient data processing and computation. Second, further based on cross-view spatio-temporal topological consistency, a contrastive learning approach PoseCLR is proposed to capture implicit features of the human body. Extensive experiments validate that NanoHTNet achieves outstanding performance and efficiency, enabling real-time inference on computationally constrained devices like Jetson Nano. Moreover, as a general pre-training approach for the 3D HPE, PoseCLR effectively optimizes the initialization parameters for various encoders, further improving performance.

References

[1] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11636–11645.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

[2] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "MHFormer: Multi-hypothesis transformer for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13137–13146.

[3] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao, "P-STMO: Pre-trained spatial temporal many-to-one model for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 461–478.

[4] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, "3D human pose estimation with spatio-temporal criss-cross attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4790–4799.

[5] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8877–8886.

[6] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas, "Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1106–1118, Jan. 2023.

[7] W. Xu, D. Xiang, G. Wang, R. Liao, M. Shao, and K. Li, "Multiview video-based 3-D pose estimation of patients in computer-assisted rehabilitation environment (CAREN)," *IEEE Trans. Hum.-Mach. Syst.*, vol. 52, no. 2, pp. 196–206, Apr. 2022.

[8] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4316–4327.

[9] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2640–2649.

[10] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.

[11] J. Lin and G. H. Lee, "Trajectory space factorization for deep video-based 3D human pose estimation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 101–114.

[12] N. Azizi, M. Fayyaz, and H. Bischof, "Occlusion handling in 3D human pose estimation with perturbed positional encoding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 441–458.

[13] F. Zhou, J. Yin, and P. Li, "Lifting by image—Leveraging image cues for accurate 3D human pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2024, vol. 38, no. 7, pp. 7632–7640.

[14] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[16] T. Jiang et al., "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.

[17] W. Zhao, W. Wang, and Y. Tian, "GraFormer: Graph-oriented transformer for 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20438–20447.

[18] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13232–13242.

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[20] J. Cai, H. Liu, R. Ding, W. Li, J. Wu, and M. Ban, "HTNet: Human topology aware network for 3D human pose estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[21] A. Zeng et al., "DeciWatch: A simple baseline for 10× efficient 2D and 3D pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 607–624.

[22] W. Li, M. Liu, H. Liu, P. Wang, J. Cai, and N. Sebe, "Hourglass tokenizer for efficient transformer-based 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 604–613.

[23] S. Choi, S. Choi, and C. Kim, "MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2328–2338.

[24] Y. Sun, A. W. Dougherty, Z. Zhang, Y. K. Choi, and C. Wu, "MixSynthFormer: A transformer encoder-like structure with mixed synthetic self-attention for efficient human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 14838–14847.

[25] F. Zhong, M. Li, K. Zhang, J. Hu, and L. Liu, "DSPNet: A low computational-cost network for human pose estimation," *Neurocomputing*, vol. 423, pp. 327–335, Jan. 2021.

[26] S. Du, Z. Yuan, and T. Ikenaga, "Kinematics-aware spatial–temporal feature transform for 3D human pose estimation," *Pattern Recognit.*, vol. 150, Jun. 2024, Art. no. 110316.

[27] Y. Dang, J. Yin, and S. Zhang, "Relation-based associative joint location for human pose estimation in videos," *IEEE Trans. Image Process.*, vol. 31, pp. 3973–3986, 2022.

[28] X. Wang, W. Zhang, C. Wang, Y. Gao, and M. Liu, "Dynamic dense graph convolutional network for skeleton-based human motion prediction," *IEEE Trans. Image Process.*, vol. 33, pp. 1–15, 2024.

[29] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin, "SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach," in *Proc. Conf. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 507–523.

[30] J. Wang, S. Huang, X. Wang, and D. Tao, "Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7771–7780.

[31] L. Wu, Z. Yu, Y. Liu, and Q. Liu, "Limb pose aware networks for monocular 3D pose estimation," *IEEE Trans. Image Process.*, vol. 31, pp. 906–917, 2022.

[32] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3D human pose regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3425–3435.

[33] Y. Cai et al., "Exploiting spatial–temporal relationships for 3D pose estimation via graph convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2272–2281.

[34] B. X. B. Yu, Z. Zhang, Y. Liu, S.-H. Zhong, Y. Liu, and C. W. Chen, "GLA-GCN: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8784–8795.

[35] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–19.

[36] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[37] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15039–15053.

[38] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward more reliable 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13041–13051.

[39] K. Holmquist and B. Wandt, "DiffPose: Multi-hypothesis human pose estimation using diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15931–15941.

[40] J. Xu, Y. Guo, and Y. Peng, "FinePOSE: Fine-grained prompt-driven 3D human pose estimation via diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 561–570.

[41] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua, "Direct prediction of 3D body poses from motion compensated sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 991–1000.

[42] J. Peng, Y. Zhou, and P. Y. Mok, "KTPFormer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1123–1132.

[43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[45] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 110, pp. 3371–3408, 2010.

[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.

[47] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[48] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.
[49] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges, "Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11230–11239.
[50] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4739–4748.
[51] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 762–770.
[52] Z. Jiang, W. Chai, L. Li, Z. Zhou, C.-Y. Yang, and J.-N. Hwang, "UniHPE: Towards unified human pose estimation via contrastive learning," 2023, *arXiv:2311.16477*.
[53] J. Wang et al., "OCR-pose: Occlusion-aware contrastive representation for unsupervised 3D human pose estimation," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5477–5485.
[54] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.
[55] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang, "Exploiting temporal contexts with strided transformer for 3D human pose estimation," *IEEE Trans. Multimedia*, vol. 25, pp. 1282–1293, 2023.
[56] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11477–11487.
[57] K. Zhai, Q. Nie, B. Ouyang, X. Li, and S. Yang, "HopFIR: Hop-wise GraphFormer with intragroup joint refinement for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 14939–14949.
[58] S. Mehraban, V. Adeli, and B. Taati, "MotionAGFormer: Enhancing 3D human pose estimation with a transformer-GCNFormer network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6920–6930.
[59] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[60] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6821–6828.
[61] H. Ci, C. Wang, X. Ma, and Y. Wang, "Optimizing network structure for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2262–2271.
[62] H. Li et al., "Pose-oriented transformer with uncertainty-guided refinement for 2D-to-3D human pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 1296–1304.
[63] Z. Tang, J. Li, Y. Hao, and R. Hong, "MLP-JCG: Multi-layer perceptron with joint-coordinate gating for efficient 3D human pose estimation," *IEEE Trans. Multimedia*, vol. 25, pp. 8712–8724, 2023.
[64] M. Wei, X. Xie, Y. Zhong, and G. Shi, "Learning pyramid-structured long-range dependencies for 3D human pose estimation," *IEEE Trans. Multimedia*, vol. 27, pp. 4684–4697, 2025.
[65] M. T. Hassan and A. Ben Hamza, "Regular splitting graph network for 3D human pose estimation," *IEEE Trans. Image Process.*, vol. 32, pp. 4212–4222, 2023.
[66] W. Li et al., "GraphMLP: A graph MLP-like architecture for 3D human pose estimation," *Pattern Recognit.*, vol. 158, Feb. 2025, Art. no. 110925.
[67] K. Liu, R. Ding, Z. Zou, L. Wang, and W. Tang, "A comprehensive study of weight sharing in graph networks for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 318–334.
[68] S. Zhang, C. Wang, L. Nie, H. Yao, Q. Huang, and Q. Tian, "Learning enriched hop-aware correlation for robust 3D human pose estimation," *Int. J. Comput. Vis.*, vol. 131, no. 6, pp. 1566–1583, Jun. 2023.
[69] D. Mehta et al., "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.
[70] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.
[71] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, and Q. Xu, "Learning skeletal graph neural networks for hard 3D pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11436–11445.
[72] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16105–16114.



Jialun Cai received the B.Eng. degree from Central South University in 2021 and the degree from the Computer Application Technology, School of Electronic and Computer Engineering, Peking University (PKU), China, under the supervision of Prof. Hong Liu. He is currently an Algorithm Engineer with Ant Group. His research interests include 3D human pose estimation, deep learning, and their applications and deployments to computer vision.



Mengyuan Liu received the Ph.D. degree from the School of EE&CS, Peking University (PKU), China, in 2017, under the supervision of Prof. H. Liu. He is currently an Assistant Professor with PKU. His research interests include human action recognition, human motion prediction, and human motion generation using RGB, depth, and skeleton data. Related methods have been published in IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, PR, CVPR, ECCV, ACM MM, and AAAI. He has been invited to be a Technical Program Committee (TPC) Member for ACPR, ACM MM, and AAAI.



Hong Liu (Member, IEEE) is currently a Full Professor with Peking University; the Co-Director of the Center of Embodied Intelligence and Robotics, Institute of AI, Peking University; and the Founding Dean of Chongqing Liangjiang School of Artificial Intelligence. He has engaged in the research of artificial intelligence, robotics, machine learning, and intelligent human–robot interaction, for more than twenty years. He has published over 300 articles in international journals and proceedings in the above areas and these papers have been cited over 14500 times. He earned the National Aerospace Science and Technology Progress Award, Geneva International Invention Award, the First Prize in Natural Science of Wu Wenjun Artificial Intelligence Science and Technology Award, and the First Prize in Natural Science of Shenzhen Science and Technology Award. He serves as the Vice President for Chinese Association for Artificial Intelligence (CAAI) and the Founding Editor-in-Chief for the top international journal of *CAAI Transactions on Intelligence Technology*.



Shuheng Zhou received the master's degree from East China University of Science and Technology. He is currently a Senior Algorithm Expert with Ant Group, with extensive experience in the application and deployment of deep learning algorithms. His research interests include the inference acceleration and deployment of large language models (LLMs).



Wenhao Li received the Ph.D. degree from the School of Computer Science, Peking University, China. He is currently a Postdoctoral Researcher with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include deep learning, machine learning, and their applications to computer vision.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 06:26:18 UTC from IEEE Xplore. Restrictions apply.