Hybrid LSTM & Transformer for 3D Human Pose Estimation

Longjie Su*

School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

Abstract. 3D human pose estimation (3DHPE) has evolved into a sophisticated and pivotal technique, emerging as a prominent research focus in computer vision and robotics, based on the power of deep neural networks. Unlike conventional image-based methods, the integration of video data introduces rich temporal information, enabling the exploitation of spatiotemporal correlations to achieve more accurate and robust results. However, modeling both local motion continuity and long-range temporal dependencies remains a significant challenge, as recurrent architectures such as LSTMs excel at capturing local dynamics but struggle with long-term information retention, whereas Transformer-based models provide strong global reasoning yet are less sensitive to fine-grained motion details. In this research, a hybrid LSTM–Transformer network that effectively integrates the local feature sensitivity of LSTMs with the global coordination ability of Transformers was proposed to cover these deficiencies. Multiple fusion strategies were investigated to evaluate their impact on 3DHPE performance. Comprehensive experiments finished on benchmark datasets demonstrate that the proposed Single-branch LSTM+Transformer architecture achieves the most competitive results, yielding a mean per-joint position error (MPJPE) of 49.51 mm, which outperforms conventional Transformer-based-only and LSTM-based-only models. This hybrid framework provides a novel and effective paradigm for future research in video-based human motion understanding.

1 Introduction

The rapid advancement of intelligent robotics has made human-robot interaction (HRI) as a fundamental research domain within contemporary robotics and computer vision. HRI encompasses diverse modalities of communication, including not only explicit channels such as conventional speech interaction but also implicit, non-verbal channels conveyed through bodily movements. Consequently, human pose estimation (HPE) was regarded as a critical research direction for perceiving and interpreting human motion in intelligent systems.

HPE involves computational analysis of joint trajectories across sequential video frames to enable continuous articular tracking. Compared with earlier approaches, the advent of deep neural networks has substantially enhanced estimation accuracy [1]. However, most existing studies concentrate on two-dimensional (2D) pose estimation, which inherently lacks depth

* Corresponding author: 225040158@link.cuhk.edu.cn

information, thereby limiting its ability to capture human motion in real-world scenarios. Three-dimensional human pose estimation integrates spatial and depth information to generate enhanced realistic motion representations, possessing substantial theoretical and practical significance for applications in HRI, autonomous driving, virtual/augmented reality, and video surveillance [2].

Nevertheless, challenges including human body occlusion, motion complexity, and scarcity of large-scale annotated 3D datasets, necessitate developing architectures capable of robustly capturing both local and global spatiotemporal dynamics.

Contemporary research increasingly employs deep learning architectures to model spatiotemporal dependencies inherent in human motion data. Pavlakos et al. introduced a sophisticated approach that reformulates three-dimensional human joint coordinate regression into volumetric representation, subsequently refining predictions through a coarse-to-fine methodology. The approach splits three-dimensional space into volumetric bins based on joint locations, simplifying the problem by estimating the probability of each joint residing within specific bins. Multi-layer convolutional neural network was used to predict two-dimensional joint heatmaps, which are subsequently elevated to three-dimensional heatmaps for joint coordinate inference. Although this methodology provides an innovative pathway for transforming two-dimensional images to three-dimensional joint coordinates, accuracy limitations persist due to inherent depth prediction complexity and relatively high training computational requirements [3]. Building upon this research trajectory, Sungheon Park et al. proposed an end-to-end method using convolutional neural network architecture for three-dimensional human pose estimation. Their methodology initially learns two-dimensional joint positions from single RGB images, subsequently fusing learned two-dimensional pose information with original image features. This fusion enables integration of spatial location information cues with convolutional feature maps, effectively capturing semantic relationships between two-dimensional and three-dimensional joint coordinates while improving prediction accuracy [4]. Subsequent research in 3DHPE shifted toward video-based approaches. Incorporating temporal sequences introduces additional time-domain information, allowing models to better handle occlusion and motion ambiguities. For instance, Ce Zheng et al. exploited the self-attention mechanism of Transformer architectures to capture global spatiotemporal dependencies. By employing Transformers across both temporal and spatial dimensions, they effectively lifted 2D joint sequences to 3D coordinates, achieving further performance gains [5, 6].

Motivated by the aforementioned studies, this work aims to further enhance the extraction of temporal-scale information from video sequences. Long Short-Term Memory (LSTM) networks, with their inherent multi-gate mechanisms, are well-suited for capturing local temporal dependencies. However, a single LSTM tends to suffer from information loss when dealing with long-range temporal relationships. In contrast, Transformer architectures exhibit strong global reasoning capabilities but are comparatively less sensitive to local details, which can also result in the omission of fine-grained information [7]. To address these complementary limitations, this study proposes a hybrid approach that integrates LSTMs with Transformers. The objective is to amplify the receptive field of LSTMs while preserving local feature sensitivity. Given that hybrid model integration can be achieved through various strategies—such as serial, parallel, or interleaved configurations—this paper introduce two distinct LSTM–Transformer architectures and, for the first time, apply such a fused framework to the 3DHPE task. On this basis, this paper investigate which combination strategy yields the most effective performance for modeling spatiotemporal dynamics in 3DHPE.

2 Datasets

The research was conducted on two publicly available benchmark datasets: Human3.6M and MPI-INF-3DHP, while the majority of the evaluations were performed on Human3.6M [8, 9]. The Human3.6M dataset contains approximately 3.6 million video frames. A synchronized multi-camera system combined with high-precision motion capture technology was used as the tool to form the dataset. It features motion sequences performed by five female and six male actors in various indoor scenarios and recorded from multiple viewpoints. Within the eleven subjects included in the dataset, seven are annotated with 3D human pose ground truth. In accordance with the standard evaluation protocol proposed in the original work, this paper utilize subjects S1, S5, S6, S7, and S8 for training sets, while subjects S9 and S11 are reserved for testing sets.

Early 3D human pose estimation datasets, such as Human3.6M, were predominantly collected in highly controlled indoor environments with limited variability in lighting, actions, and background, which constrained the generalization performance of models to real-world scenarios. To address this limitation, the MPI-INF-3DHP dataset was introduced [10], employing a combination of video recording and a marker-less motion capture system to enable more natural motion acquisition. Furthermore, the dataset incorporates foreground and background augmentation techniques and includes a diverse set of actions performed by eight subjects across both indoor and outdoor scenes. This design significantly enhances the dataset's variability, contributing to improved model robustness and generalization to more complicated conditions.

During the data preprocessing stage, this research first loads the 3D joint location annotations and performs normalization based on the corresponding camera parameters, including both intrinsic and extrinsic calibrations. Subsequently, 2D keypoints are extracted using a pre-trained 2D pose detector. Researchers then temporally align and match the 2D and 3D pose sequences on a frame-by-frame basis to ensure consistency across modalities. After this initial preprocessing, the data are partitioned into training and testing sets following the dataset split protocol described previously, and data loaders are constructed accordingly.

In the experimental evaluation, the Mean Per Joint Position Error (MPJPE) serves as the primary assessment metric, which quantifies the average Euclidean distance between predicted joint coordinates and their corresponding ground truth annotations.

3 Solution Methodology

This study adopts a hybrid modeling approach that integrates two well-established deep neural network architectures: the Long Short-Term Memory (LSTM) network and the Transformer model.

The LSTM component addresses fundamental limitations of traditional Recurrent Neural Networks (RNNs) through sophisticated gating mechanisms that effectively mitigate gradient vanishing and explosion phenomena. This enhancement allows LSTM networks to maintain a more effective balance between the retention of historical information and the incorporation of new input, thereby enabling more robust processing of extended temporal sequences compared to conventional RNNs.

LSTM networks operate through three critical gating mechanisms: the input gate controls incorporation of new information, the output gate modulates current cell state contributions to network output, and the forget gate governs retention or elimination of previous cell state information C_(t-1) through outputs constrained to [0,1] as defined in Equation 1.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \in (0,1) \quad (1)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2}$$

The cell state update mechanism, formulated in Equation (2), demonstrates how these gating operations collectively enable selective information propagation across time steps, where C_t denotes the current memory state, f_t represents the output of the forget gate, i_t corresponds to the input gate activation, and $\tilde{C}_t$ is the candidate memory content derived from the current input.

Subsequently, inter-gate interactions within the LSTM framework are disregarded, with analysis focusing exclusively on historical information propagation mechanisms. By isolating forget gate influence and excluding input gate contributions, Equation (2) simplifies to Equation (3), wherein cell state evolution depends exclusively on recursive forget gate operations.

$$C_t = f_t * C_{t-1} \tag{3}$$

Recursive expansion across time steps yields Equation (4), expressing current cell state C_t as a function of cumulative multiplicative forget gate effects across preceding intervals. This reveals the fundamental mathematical structure of LSTM memory propagation: extended sequences require successive multiplication of forget gate activations within [0,1], creating exponential decay that systematically reduces temporally distant information influence.

$$C_t = \left(\prod_{j=1}^{k} f_{t-j+1} \right) \cdot C_{t-k} \tag{4}$$

Consequently, LSTM networks exhibit pronounced temporal locality bias, where recent inputs maintain substantially greater influence over internal representations compared to long-range dependencies. This characteristic limits the network's capacity for modeling complex temporal relationships spanning extended time horizons, necessitating architectural enhancements for comprehensive sequence analysis.

The Transformer architecture represents a fundamental advancement in sequence modeling by eliminating recurrent processing in favor of self-attention mechanisms combined with positional encoding. The multi-head self-attention mechanism enables dynamic computation of contextualized token representations through simultaneous attention to all sequence elements. Operating through Query (Q), Key (K), and Value (V) parameter matrices, this mechanism captures both local and global dependencies regardless of positional distance while generating context-aware encodings conditioned on complete sequence information. To preserve temporal structure without sequential processing, the Transformer incorporates positional encodings into input embeddings through sinusoidal functions or learned vectors. This design provides explicit positional information enabling the model to distinguish relative positional relationships.

To synthesize the complementary characteristics of these architectures, two distinct LSTM+Transformer network configurations are proposed. Within this integrated framework, the LSTM module serves as the foundational component responsible for capturing fine-grained motion variations between adjacent video frames, with particular emphasis on modeling short-term joint adjustments and local temporal dynamics. The Transformer module operates at a higher abstraction level to integrate global movement coherence, facilitating comprehensive modeling of whole-body coordination throughout complete motion sequences. This architectural fusion enables the model to simultaneously process

local temporal dependencies and global spatial-temporal relationships, thereby optimizing three-dimensional human pose estimation performance, which has not been used in 3D HPE.

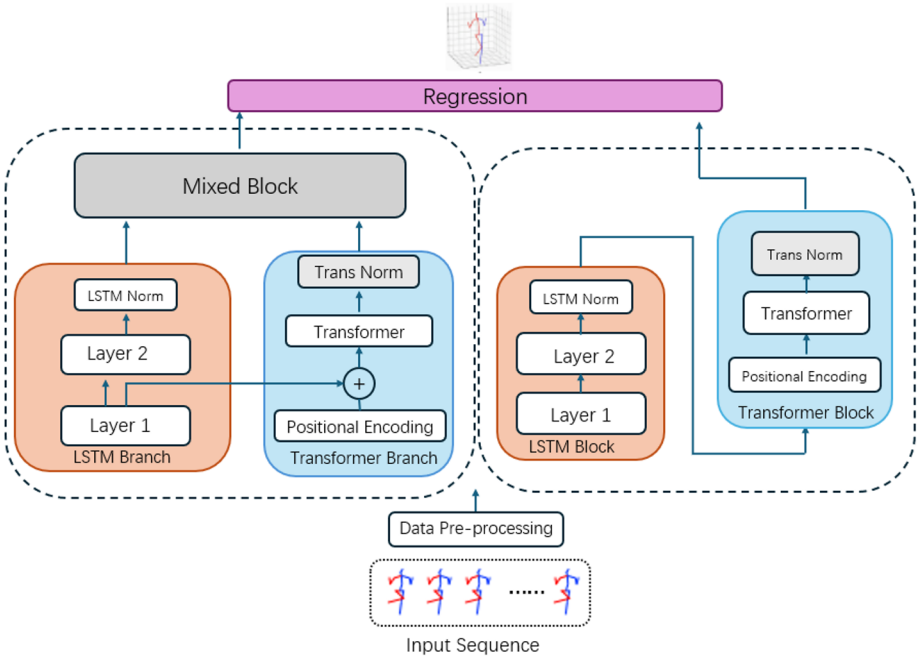

Fig. 1. Two LSTM+Transformer network architectures: the one on the left is a dual-branch network architecture, and the one on the right is a single-branch serial network architecture. (Picture credit: Original)

3.1 Architecture Ⅰ single branch

The network architecture is presented in the right panel of Figure 1. This architecture initiates processing through a two-layer LSTM module that directly processes preprocessed sequential input data to extract local temporal feature representations. This initial LSTM component specializes in capturing fine-grained short-term motion dynamics across adjacent temporal frames. However, given the inherent temporal locality bias characteristics previously examined, exclusive reliance on LSTM architecture proves inadequate for modeling the long-range temporal dependencies that characterize extended motion sequences.

To mitigate this architectural limitation, the framework integrates a Transformer-based module employing multi-head self-attention mechanisms. The two-layer LSTM output undergoes normalization processing before serving as input to the Transformer encoder block. Positional encoding is incorporated into the normalized sequence prior to attention computation to preserve temporal ordering information. The Transformer encoder utilizes eight self-attention heads, enabling comprehensive capture of global temporal relationships and inter-joint interactions throughout the complete sequence duration.

Subsequent to Transformer encoding, Layer Normalization is applied to make output representations stablized. The architecture concludes with a regression module comprising convolutional layers and linear projection components that map the learned temporal-spatial feature representations to final pose estimation outputs.

3.2 Architecture Ⅱ parallel branch

In distinction to the single-branch configuration, the proposed parallel dual-branch architecture implements a comprehensive framework for temporal feature extraction and representation learning. As depicted in the left panel of Figure 1, input data undergoes concurrent processing through two distinct yet complementary computational pathways.

The first processing pathway directly makes input data through the initial hidden layer of a two-layer LSTM network, which specializes in capturing local temporal dynamics and short-term dependency patterns. Simultaneously, the second pathway applies positional encoding to inject explicit temporal ordering information into the input sequence, preparing it for subsequent Transformer-based processing.

The output from the first LSTM hidden layer is partitioned into two discrete subcomponents. The first partition is concatenated with the corresponding output from the Transformer's positional encoding module, with this combined representation serving as input to a Transformer encoder layer configured with eight self-attention heads. This encoder output passes through a Layer Normalization module to stabilize and refine the learned representations, with the final result designated as the Transformer branch output.

Meanwhile, the second partition of the LSTM hidden state continues processing through the second LSTM hidden layer to maintain sequential information extraction. This output undergoes Layer Normalization processing, with the resulting representation serving as the LSTM branch output.

The network's final output emerges through learnable weighted fusion of the Transformer and LSTM branch outputs, integrating both local and global temporal information. The regression layer maintains an identical configuration to the first architecture. This fusion methodology enables the architecture to simultaneously leverage the LSTM's fine-grained motion pattern modeling capabilities and the Transformer's long-range temporal dependency capture capacity, thereby enhancing overall performance in complex three-dimensional human pose estimation applications.

4 Experiments

4.1 Configurations

All experimental procedures were conducted using a single NVIDIA RTX 3090 Ti graphics processing unit operating under a Linux environment. The implementation utilized the PyTorch deep learning framework (version 1.13.0) with CUDA 11.7.0 for GPU acceleration and Python 3.8 as the primary programming environment shown in Table 1.

Table 1. Experiment configurations.

GPU	Env	Input Sequence	Optimizer	Batch Size
NVIDIA RTX3090Ti x1	Python 3.8 PyTorch 1.13.0 CUDA 11.7.0	81 frames	Adam	512

The training protocol encompassed 80 epochs with a batch size configuration of 512. Input sequence length was standardized at 81 video frames to ensure consistent temporal context across all experimental samples. The Adam optimizer was used to facilitate efficient model convergence, incorporating a weight decay coefficient of 0.01 for regularization purposes.

The initial learning rate was established at 0.0001, with a linear learning rate decay schedule applied throughout the training process, as demonstrated in Figure 2. This scheduling methodology enabled accelerated learning while promoting gradual parameter refinement in later epochs, thereby enhancing both training stability and convergence characteristics. The training time of the single epoch for both LSTM+Transformer models is about 25-45 minutes. The baseline model of the experiment is the LSTM model.

4.2 Results and Analysis

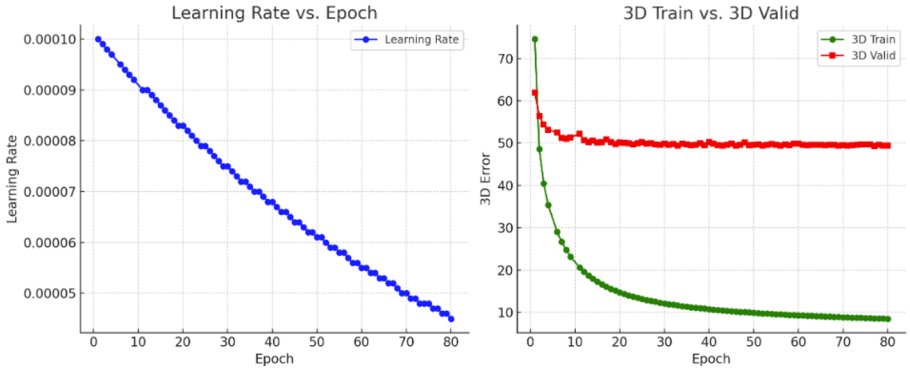

Fig. 2. Learning Rate Decay and Training Convergence of the optimal model. (Picture credit: Original)

Table 2. Experimental results of different architectures on the Human 3.6m dataset.

Methods	MPJPE(mm)
LSTM	55.55
Dual-Branch LSTM+Transformer	51.57
PoseformerV2	50.10
Single-Branch LSTM+Transformer	49.51

Table 2 presents a comprehensive summary of all experimental architectures and their corresponding performance on the Human3.6M dataset. As shown in the results, this paper proposed single-branch LSTM+Transformer architecture that achieves the best overall performance, yielding a Mean Per Joint Position Error (MPJPE) of 49.51 mm, outperforming all compared models.

The alternative dual-branch network architecture exhibits competitive performance characteristics, attaining an MPJPE of 51.57 mm. This result represents only marginal deviation from the PoseFormerV2 baseline performance of 50.10 mm, indicating substantial potential for performance optimization through further architectural refinement. The LSTM-only baseline model achieved an MPJPE of 55.55 mm, thereby demonstrating the significant performance enhancement attributable to Transformer integration within the hybrid architecture..

Furthermore, as illustrated in Figure 2, the training and validation curves of the models exhibit stable convergence, reinforcing the reliability of the experimental outcomes. These

results collectively validate the effectiveness of the proposed architectures in capturing both short- and long-term temporal dependencies, leading to superior performance in 3D HPE.

To assess the generalization capabilities of the optimal model configuration, additional evaluation was conducted using the MPI-INF-3DHP dataset for both training and testing procedures. As previously established, the MPI-INF-3DHP dataset presents significantly greater challenges relative to Human3.6M due to enhanced environmental diversity and the introduction of more noise. The model can achieve better results compared with the Human3.6m, as presented in Table 3.

Table 3. Performance of the optimal architecture on different datasets.

Model	Human3.6m	MPI-INF-3DHP
Single-Branch LSTM+Transformer	49.51	29.34

5 Conclusions

Building upon contemporary 3DHPE paradigms, this investigation introduces a novel LSTM+Transformer hybrid architecture for three-dimensional human pose estimation. The proposed framework enhances local and global feature diversity while addressing limitations inherent in single-model architectures. Two architectural variants are developed—a single-branch design and a parallel dual-branch design—to extract complementary representations from sequential input data.

Experimental evaluation demonstrates that the single-branch LSTM+Transformer architecture outperforms PoseFormerV2, establishing superior performance within the evaluation framework, while the parallel dual-branch architecture achieves comparable accuracy, indicating substantial potential for further optimization. Additional validation conducted on MPI-INF-3DHP confirms the proposed model's effectiveness. The results affirm the robustness of the optimal architecture while empirically demonstrating that MPI-INF-3DHP provides enhanced diversity and multi-scene stability compared to Human3.6M, reflecting superior suitability for real-world pose estimation applications.

Future research directions encompass three primary areas. First, given inherent differences between Human3.6M and MPI-INF-3DHP regarding environmental conditions, camera perspectives, and scene diversity, cross-dataset joint training strategies will be investigated to enhance generalization in complex and unconstrained environments. Second, the parallel dual-branch architecture requires further enhancement through incorporation of attention mechanisms within the LSTM branch and exploration of multi-layer feature fusion between LSTM and Transformer blocks, thereby unlocking latent representational capacity. Third, implementation of robust learning rate scheduling strategies, such as cosine annealing with warm restarts, may improve optimization stability and accelerate convergence, ultimately enhancing model reliability in practical deployment scenarios.

References

1. Q. Zhao, C. Zheng, M. Liu, P. Wang, C. Chen, 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, 8877–8886 (2023)
2. J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, L. Shao, Deep 3D human pose estimation: A review. Comput. Vis. Image Underst. 210, 103225 (2021)

3. G. Pavlakos, X. Zhou, K.G. Derpanis, K. Daniilidis, Coarse-to-fine volumetric prediction for single-image 3D human pose, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7025–7034 (2017)
4. S. Park, J. Hwang, N. Kwak, 3D human pose estimation using convolutional neural networks with 2D pose information, in Proceedings of the European Conference on Computer Vision, 156–169 (2016)
5. C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, M. Shah, Deep learning-based human pose estimation: A survey. ACM Comput. Surv. 56(1), 1–37 (2023)
6. C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, Z. Ding, 3D human pose estimation with spatial and temporal transformers, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 11656–11665 (2021)
7. S. Tang, C. Li, P. Zhang, R.N. Tang, SwinLSTM: Improving spatiotemporal prediction accuracy using Swin Transformer and LSTM, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 13470–13479 (2023)
8. C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. 36(7), 1325–1339 (2013)
9. D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 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, 7753–7762 (2019)
10. D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3D human pose estimation in the wild using improved CNN supervision, in Proceedings of the International Conference on 3D Vision (3DV), 506–516 (2017)