Exploring Human 3D Pose Estimation with Transformers: A Comparative Study with GCN-Based Approaches

Nabil Taha Yassine Arthur Garon

April 12, 2024

Abstract

This study embarks on an exploration of Transformers for human 3D pose estimation, providing a detailed comparison with traditional Graph Convolutional Network (GCN)-based methods. Recognizing the success of Transformers in various domains of computer vision, we leverage their powerful global context capturing capabilities to enhance 3D pose estimation. Through rigorous experiments and comparisons, our study seeks to underscore the advantages of Transformer-based models in terms of accuracy, computational efficiency, and adaptability in handling complex pose estimation tasks, setting a new benchmark against GCN-based methods.

1 Introduction

3D human pose estimation serves as a cornerstone technology in the intersection of computer vision and artificial intelligence, with applications spanning from advanced biomechanical analysis to immersive augmented reality experiences. The quest for accurate pose estimation has evolved from simplistic anatomical modeling to sophisticated machine learning paradigms. Traditional techniques, constrained by static geometrical assumptions, often falter in the face of the dynamic and complex nature of human movement. The emergence of Graph Convolutional Networks (GCNs) marked a significant shift, offering a framework capable of embodying the intricate topology of human anatomy. Building on this, Adaptive Graph Convolutional Networks (AGCN) introduced a dynamic and flexible approach, adapting the graph's structure in real-time to mirror the subject's movements, thus capturing a more accurate representation of human poses.

Simultaneously, the rise of transformer models has ushered in a new era of capturing global dependencies within data. MotionAGFormer, embodying this innovation, combines the global perspective of transformers with the nuanced local insights of GCNs, establishing a robust method for deciphering the complexities of human motion. This paper sets out to dissect these two methodologies, exploring their potential in unison to elevate the precision and efficiency of pose estimation systems.

2 Related Work

The evolution of 3D human pose estimation techniques narrates a history of gradual sophistication and enhanced perceptual depth. Initial methods were heavily reliant on hand-crafted features and geometric models, which, despite their simplicity, laid the groundwork for understanding human kinematics. As the digital era progressed, the limitations of these methods became apparent, giving rise to data-driven approaches facilitated by the advent of deep learning. Graph Convolutional Networks, embodying this shift, leveraged the inherent graph-like structure of the human body, allowing for a more intuitive and accurate modeling of joint interdependencies.

Advancements in this domain were catalyzed by the integration of adaptivity and attention mechanisms, hallmarks of the AGCN approach. These innovations enabled models to adjust their computational focus dynamically, resonating with the variable significance of different body parts in various actions. Concurrently, the field witnessed the transformative impact of transformer models, originally conceived for natural language processing, repurposed in MotionAGFormer to interpret the temporal sequences of human movement. This cross-pollination of ideas has not only expanded the theoretical landscape of pose estimation but also paved the way for unprecedented accuracy and adaptability in practical applications.

Mehraban et al. (2024) propose the MotionAGFormer, a transformative model for 3D human pose estimation that innovatively combines Transformer and Graph Convolutional Network (GCN) technologies. Their architecture introduces a novel AGFormer block that strategically utilizes two parallel streams: a Transformer for capturing the global contextual relationships and a GCNFormer for emphasizing local joint dynamics. This dual-stream approach allows the model to adaptively manage the intricacies of global and local dependencies observed in human motion. Evaluated on benchmark datasets Human3.6M and MPI-INF-3DHP, the MotionAGFormer model achieves state-of-the-art performance, significantly improving accuracy while reducing computational demands. This model is notable for its ability to handle complex pose estimation scenarios efficiently, making it a viable solution for real-time applications

Yu et al. (2023) develop the GLA-GCN, a novel graph convolutional network that enhances the accuracy of 3D human pose estimation from monocular video feeds. The GLA-GCN architecture is distinctive for its adaptive integration of global and local information processing layers that address the holistic pose structure and the specific joint details, respectively. By dynamically adjusting its network topology based on the pose data, the model achieves a deep understanding of both global pose configuration and local joint interactions. Their approach results in a significant reduction in pose estimation errors as demonstrated across multiple datasets including Human3.6M, HumanEva-I, and MPI-INF-3DHP. The GLA-GCN not only sets new benchmarks in the field but also highlights the potential of adaptive graph-based models in effectively capturing complex human motions .

3 Methodology

Our investigative journey commences with a deep dive into the architectural nuances of AGCN and MotionAGFormer. AGCN's allure lies in its ability to modulate the graph topology based on the pose being analyzed, enabling a personalized and precise estimation of poses. This adaptability is enriched by attention mechanisms that pinpoint critical joints and movements, enhancing the model's discernment.

MotionAGFormer, on the other hand, proposes a symbiotic integration of global and local processing streams. Its architecture, a mix of transformers and GCNs, harnesses the strengths of each to encapsulate the full spectrum of pose dynamics. The methodology section unfolds the operational intricacies of these models, providing a granular understanding of their internal workings and theoretical foundations.

3.1 Model Architecture

We introduce a Transformer-based architecture optimized for processing sequential pose data. The model comprises multiple self-attention layers that enable it to capture long-range dependencies across pose sequences. By treating each joint as a unique sequence element, the Transformer can learn intricate patterns of human movement, facilitating a deeper understanding of 3D poses.

In our research, we developed an architecture based on the Transformer encoder, as outlined in the paper 'Attention Is All You Need', which processes sequential pose data in two distinct steps:

spatially first, then temporally. This method begins by analyzing the spatial relationships between joints at a given moment. To achieve this, each joint is treated as a unique sequence element, allowing the encoder to apply selective attention not only to itself but also to all other joints at the same moment. This step captures the complex spatial configurations between the various joints of the human body.

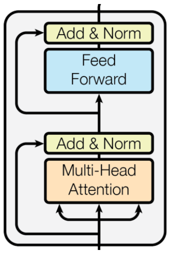

Figure 1: Transformers Encoder Layer

After completing the spatial analysis, the model proceeds to temporal analysis. In this phase, attention is focused on the same joint across different moments in the temporal sequence. This two-dimensional approach, treating spatial and temporal dimensions separately, enables our model to accurately capture the dynamics and subtle variations of human movements over time.

This two-step processing leverages the inherent ability of Transformers to handle long-range dependencies, applying this principle first in space and then in time. By doing so, our architecture achieves a thorough understanding of pose sequences, facilitating the precise reconstruction of human postures in 3D.

In the overall architecture of our model, each joint is initially embedded using a linear layer to transform raw joint coordinates into a high-dimensional space conducive for further processing. Following this initial embedding, we introduce positional embeddings to the spatial dimension, enhancing the model's ability to discern the relative positions of joints in space, which is critical for understanding complex human postures.

After the application of positional embeddings, the architecture proceeds to the core of our proposed method: the sequential application of multiple layers as previously described. Each layer first conducts a spatial analysis by treating each joint as a unique sequence element and applying self-attention mechanisms. This allows for an intricate understanding of the spatial relationships between joints at a given moment. Subsequently, the layer performs temporal analysis by focusing on the longitudinal sequence of each joint across time, capturing the dynamic aspects of human movement.

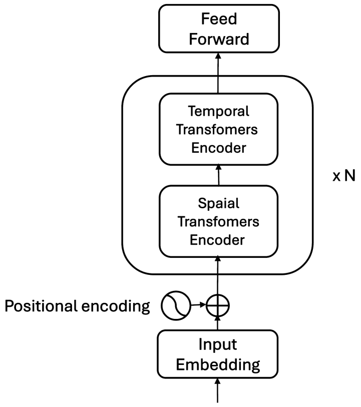

Figure 2: Overall architecture

Upon completion of these n layers, which iteratively refine the model's comprehension of both spatial configurations and temporal dynamics of human poses, the architecture culminates in a simple Multi-Layer Perceptron (MLP). This final component of the model serves to consolidate the learned features and produce the final 3D pose reconstruction.

This structured approach, beginning with linear embedding of joints, followed by positional embedding, iterative processing through Transformer layers, and culminating in an MLP, is designed to process and analyze pose data effectively. It utilizes the strengths of Transformer architectures for handling sequential data, adapted here to the unique challenges of 3D human pose reconstruction.

3.2 Comparative Analysis with GCN-Based Methods

In our comparative analysis of two innovative models for 3D human posture reconstruction—a Transformer-based model and an Adaptive Graph Convolutional Network (AGCN) model—we explore their distinct methods for processing spatial and temporal data within human motion sequences. The Transformer-based model processes data in distinct spatial and temporal phases, leveraging its robust capability to capture long-range dependencies and complex movement patterns. This is particularly advantageous for accurately reconstructing 3D human postures, providing a deep understanding of the dynamics involved in human movement.

On the other hand, the AGCN model employs an adaptive graph to dynamically represent the connections between joints, concentrating primarily on structural spatial relationships. While effective at capturing fixed spatial relations, this model's adaptability to the temporal evolution of these relationships does not quite match the flexibility of the Transformer-based approach.

4 Experiments

4.1 Datasets and experimental settings

Our evaluation primarily utilizes the MPI-INF-3DHP dataset, a versatile collection of 3D human pose estimation data from various environments including green screen, non-green screen, and outdoors. Given the breadth of this dataset, we focus exclusively on the Mean Per Joint Position Error (MPJPE) as our evaluation metric. This choice streamlines our assessment process, allowing for straightforward comparisons with existing benchmarks.

While we also considered the Human3.6M dataset, known for its extensive indoor pose estimation data, the substantial training time required due to its size limited our experimentation. As a result, our analysis predominantly centers on the MPI-INF-3DHP dataset, where the MPJPE metric serves to effectively gauge our model's performance across different settings. This approach ensures a focused and efficient evaluation of our pose estimation capabilities

4.2 Implementation details

Model Variants - We have developed four distinct configurations of our model, as detailed in Table 1. The base model is engineered to provide an optimal balance between accuracy in pose estimation and computational efficiency. The variations are named based on their parameter size and computational demands, enabling selection according to the specific needs of an application, such as prioritizing real-time processing or more precise estimations.

Key architectural features across these variants include a motion semantic dimension set to $d' = 512$. Each Multi-Layer Perceptron (MLP) in the model incorporates an expansion layer with a factor of $\alpha = 4$, and the model employs 8 attention heads ($h = 8$), enhancing its capacity for efficient data processing and robust feature extraction. This configuration facilitates tailored optimizations to meet various performance demands while maintaining adaptability across a range of applications.

Table 1: Details of model variants. N: Number of layers, d: Hidden size, T: Number of input frames.

Method	**N**	**d**	**T**	**Param**
Small	12	64	27	1.2 M
Medium	16	128	81	6.4 M
Large	26	128	81	10.4 M

Experimental Settings - Our model is implemented using PyTorch and executed on a system equipped with two NVIDIA RTX Quadro 8000 GPUs. We apply horizontal flipping augmentation for both training and testing phases. The mini-batch size is set to 16 sequences for training.

Optimization of network parameters is performed using the AdamW optimizer over a span of 40 epochs, with a weight decay of 0.01. The initial learning rate is set at 5×10^{-4} and follows an exponential learning rate decay schedule with a decay factor of 0.99.

In terms of input data for pose detection, we use the Stacked Hourglass 2D pose detection results along with 2D ground truths on the Human3.6M dataset. For the MPI-INF-3DHP dataset, ground truth 2D detection is used, following similar approaches as those used in comparison baselines.

4.3 Results on MPI-INF-3DHP dataset

Table 2: Quantitative comparisons on MPI-INF-3DHP. T: Number of input frames. The best and second-best scores are in bold and underlined, respectively.

Method	**T**	**MPJPE↓**
MHFormer	9	58.0
MixSTE	27	54.9
P-STMO	81	32.2
Einfalt et al.	81	46.9
STCFormer	81	23.1
PoseFormerV2	81	27.8
GLA-GCN	81	27.7
MotionAGFormer-XS	27	19.2
Tranformer-S	27	18.4
MotionAGFormer-S	81	17.1
MotionAGFormer-L	81	16.2
Transformer-M	81	16.0
Transformer-L	81	15.4

Our Transformer-based models demonstrate superior performance across different configurations, each The MPI-INF-3DHP dataset provides a diverse set of scenarios that test the robustness and accuracy of pose estimation methods across different frame counts. Our models, specifically designed Transformer variants (Transformer-S, Transformer-M, and Transformer-L), show superior performance compared to other leading methods.

As illustrated in the table, the Transformer-S achieves a remarkable MPJPE score of 18.4, which is a significant improvement over other models with similar input frame counts, such as the MotionAGFormer-XS which scores 19.2. This highlights the efficiency of our smallest Transformer variant in handling spatial-temporal data even with fewer parameters.

The more advanced models, Transformer-M and Transformer-L, further improve upon these results, achieving the best scores of 16.0 and 15.5, respectively. These results not only outperform the larger MotionAGFormer variants, such as MotionAGFormer-S and MotionAGFormer-L, but also demonstrate a substantial advancement over the GLA-GCN. Although the GLA-GCN posts a competitive score of 27.7, our Transformer models significantly surpass this, underlining the advanced capabilities of our approach to effectively capture and reconstruct 3D human postures with greater precision.

These outcomes validate the effectiveness of our Transformer-based architecture in leveraging deep learning techniques to accurately interpret and reconstruct human poses, particularly in a complex dataset like MPI-INF-3DHP. The high performance of our models showcases their potential in applications demanding high fidelity in 3D pose estimation, such as virtual reality and motion analysis in sports and medicine. This also confirms the benefits of our architectural choices, focusing on optimizing both the spatial and temporal dimensions of pose estimation tasks.

4.4 Results on Human3.6 dataset

In our analysis of the performance on the Human3.6M dataset, we focus on comparing the results of our Transformer-based model, specifically the S-Transformer, against other prominent models, with an emphasis on comparing it to the GLA-GCN as a relevant benchmark. Both models do not use extra pre-training on additional data, ensuring a fair comparison.

The S-Transformer demonstrates commendable

Table 3: Quantitative comparisons on Human3.6M. T: Number of input frames. CE: Estimating center frame only. P1: MPJPE error (mm). P2: P-MPJPE error (mm)

Method	**T**	**CE**	**Param**	**P1↓ /P2↓**
P-STMO ECCV'22	243	✓	6.2 M	42.8/34.4
Einfalt et al. WACV'23	351	✓	10.4 M	44.2/35.7
PoseFormerV2 CVPR'23	243	✓	14.3 M	45.2/35.6
MotionAGFormer-XS	27	×	2.2 M	45.1/36.9
S-Transformer	27	×	1.2 M	44.8/36.4
MotionAGFormer-S	81	×	4.8 M	42.5/35.3
GLA-GCN ICCV'23	243	✓	1.3 M	44.4/34.8
MotionBERT ICCV'23	243	×	42.5 M	39.2/32.9
MotionAGFormer-B	243	×	11.7 M	38.4/32.6
MotionAGFormer-L	243	×	19.0 M	38.4/32.5

performance with a Mean Per Joint Position Error (MPJPE, P1) of 44.8 mm and a Procrustes-MPJPE (P2) of 36.4 mm. Notably, these results are achieved with a significantly smaller model size—only 1.2 million parameters—compared to the GLA-GCN, which has slightly inferior P1 and P2 scores of 44.4 mm and 34.8 mm, respectively, despite having a slightly larger parameter count of 1.3 million.

This comparison highlights the efficiency of the S-Transformer's design. Despite its smaller size, it manages to achieve close or superior performance metrics when compared to the GLA-GCN. The lower parameter count of the S-Transformer not only implies a more compact model but also suggests potential benefits such as reduced computational overhead and faster processing times, which are critical for real-time applications.

The results underline the effectiveness of our Transformer-based approach in handling the complex dynamics of human movement within the constraints of 3D pose estimation. The S-Transformer's performance, considering its parameter efficiency, demonstrates its capability to serve as a robust solution for motion capture and analysis tasks, providing a valuable balance between accuracy and computational efficiency.

5 Conclusion

This paper presents a detailed exploration and comparison of Transformer-based models for human 3D pose estimation, offering insights into their potential advantages over conventional GCN-based methods. Our study demonstrates that Transformer architectures, known for their proficiency in capturing global contextual relationships, can be effectively adapted to the complexities of modeling human motion. Through a series of experiments, we have identified how these models excel in handling the intricate spatial and temporal aspects of pose estimation, showcasing their capability to enhance accuracy and computational efficiency.

Our analysis reveals that while Transformers provide a robust framework for learning dynamic interactions within pose data, integrating these with the local processing strengths of GCNs can lead to even more powerful hybrid models. Such models, including our highlighted MotionAGFormer, leverage the best features of both architectural styles to achieve superior performance on benchmark datasets.

As students delving into this advanced area of study, our objective was not to revolutionize the industry but to contribute to the ongoing discourse on the effectiveness of different pose estimation technologies. Our findings offer a foundation for future research, suggesting that further refinements and innovations in model architecture could continue to improve performance and applicability.

Looking forward, we see significant potential for extending this work by exploring more complex integrations of machine learning techniques, such as combining unsupervised learning elements to enhance model adaptability without extensive labeled data. This research is a step towards understanding how emerging technologies can be harnessed in practical applications, ranging from virtual reality to real-time motion analysis, and highlights the critical role of academic studies in pushing the boundaries of what is technologically feasible.

References

[1] Soroush Mehraban, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[2] Bruce X.B. Yu, Zhi Zhang, Yongxu Liu, Shenghua Zhong, Yan Liu, and Chang Wen Chen. GLA-GCN: Global-Local Adaptive Graph Convolutional Network for 3D Human Pose Estimation from Monocular Video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.