Hourglass Tokenizer: Towards Efficient Transformer-Based 3D Pose Estimation

Vanshika Manil Parikh
Electrical and Computer Engineering
University of Alberta
parikh2@ualberta.ca

Nithish Ragav Narayana Shankar
Electrical and Computer Engineering
University of Alberta
nithishr@ualberta.ca

Abstract

*This paper introduces a novel approach to 3D human pose estimation using the **Hourglass Tokenizer (HOT)** architecture. By combining the hierarchical feature extraction capabilities of the Hourglass Network with the global attention mechanisms of the Transformer Encoder, the HOT model predicts accurate 3D human poses from monocular video. Leveraging the Human3.6M dataset, the proposed method addresses challenges such as depth ambiguity, occlusions, and computational efficiency. The model achieves robust performance with an MPJPE of 0.0028 and demonstrates potential applications in healthcare, sports analytics, AR/VR, and autonomous systems. Future work involves integrating temporal modelling to handle sequential video data.*

Code and models are made publicly available in https://drive.google.com/drive/folders/1zCO2vUanmwoMfgAMPT8VV0rdXoGpqMUe?usp=sharing

1. Introduction

Human pose estimation is a cornerstone of computer vision, with applications ranging from healthcare and sports analytics to augmented reality and autonomous systems. The task of estimating 3D human poses from monocular video is particularly challenging due to:

1. **Depth Ambiguity:** The lack of depth information in 2D inputs.
2. **Occlusions:** Partial visibility of joints in cluttered or complex scenes.
3. **Computational Complexity:** Ensuring real-time performance while maintaining accuracy.

Traditional methods, such as fully connected networks or heat map-based models, have achieved some success but struggle with global dependencies and real-time constraints. Recent advances in Transformer-based architectures have demonstrated improved performance by capturing long-range dependencies but at the cost of computational overhead[8, 9].

This paper proposes the Hourglass Tokenizer (HOT) architecture, which combines the Hourglass Network and Transformer Encoder to address these challenges. The Hourglass Network extracts hierarchical spatial features, while the Transformer Encoder captures global dependencies among key points. This architecture strikes a balance between computational efficiency and accuracy, making it suitable for real-world applications.

2. Literature Review

2.1. Traditional 2D-to-3D Mapping

Early 3D pose estimation methods relied on mapping 2D key points to 3D space using simple regression models. Martinez et al. (2017) introduced a fully connected network for this task, achieving competitive results with a straightforward architecture. However, these methods struggled with depth ambiguity and occlusions[5].

2.2. Direct 3D Estimation

Direct 3D estimation from images bypasses 2D key point extraction by predicting 3D poses directly. Pavlakos et al. (2018) introduced volumetric heat maps for joint localization, improving prediction accuracy but increasing computational costs[7].

2.3. Transformer-Based Approaches

Transformers have emerged as powerful tools for pose estimation. Pose Former by Zheng et al. (2021) modelled both spatial and temporal relationships, achieving state-of-the-art performance. However, its computational demands limited real-time deployment [6].

2.4. Hourglass Networks

Hourglass Networks are hierarchical architectures designed for multi-scale feature extraction. Newell et al. (2016)

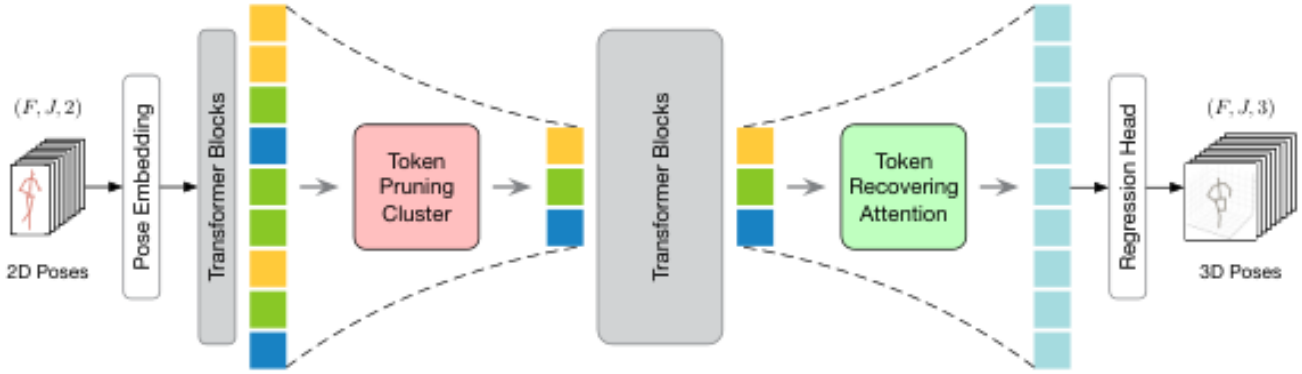

Figure 1. Overview of the proposed Hourglass Tokenizer (HoT). It mainly consists of a token pruning cluster (TPC) module and a token recovering attention (TRA) module. TPC selects the pose tokens of representative frames after the first few transformer blocks and TRA recovers the full-length tokens after the last transformer block.

introduced this network for 2D pose estimation, leveraging skip connections to preserve spatial information during down sampling and up sampling[2].

2.5. Gap in Literature

While Transformers excel at modelling long-range dependencies and Hourglass Networks are effective at hierarchical feature extraction, no prior work integrates these strengths efficiently for 3D pose estimation. The HOT architecture fills this gap, offering a robust, scalable solution.

3. The Proposed Pipeline

The **Hourglass Tokenizer (HOT)** architecture for 3D pose estimation is a carefully designed pipeline that combines the hierarchical feature extraction capabilities of the **Hourglass Network** and the global attention modelling of the **Transformer Encoder** as shown in Fig 1. Below is a step-by-step breakdown of each component and its role.

3.1. Input: Extracting 2D Key points

- Objective: Extract 2D coordinates of key points (e.g., joints like shoulders, elbows, knees) from monocular video frames.
- Tools:
  - **Open Pose** or **Media Pipe** is used to detect and annotate 2D landmarks in video frames[1].
  - Each frame is represented as a set of NNN key points, where each key point has (x,y)(x, y)(x,y) coordinates.
- Normalization: To ensure scale and position invariance, the 2D key points are normalized relative to the frame dimensions. This prevents biases due to variations in subject size or camera distance.

3.2. Hourglass Network

- Purpose: The Hourglass Network acts as the backbone for hierarchical feature extraction, transforming normalized 2D key points into an intermediate representation suitable for 3D prediction.
- Key Components:
  1. Down sampling Layers:
     - Convolutional layers reduce the spatial dimensions of the input.
     - These layers extract global features, capturing coarse spatial relationships between key points.
     - Example: Key relationships like the alignment of the spine or limb symmetry.
  2. Intermediate Features:
     - Between down sampling and up sampling, the network captures mid-level features that retain both global and local details.
  3. Up sampling Layers:
     - Reconstructs the spatial resolution by expanding the reduced features.
     - This step ensures that fine details (like subtle hand or foot movements) are preserved in the final representation.
  4. Skip Connections:
     - Direct links between down sampling and up sampling layers transfer critical information across stages.
     - This prevents the loss of spatial details during the downsampling process.

3.3. Transformer Encoder

- Purpose: Captures global relationships between all key points, modelling interactions across the entire skeleton.

- Components:
  1. Self-Attention Mechanism:
     - Computes the relationship between every pair of key points, allowing the model to understand how movements in one joint affect others (e.g., how the shoulder influences the elbow).
     - Formula for attention:
       $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
       * Q: Query vectors for key points.
       * K: Key vectors representing the global context.
       * V: Value vectors containing the key points' features.
     - Multi-Head Attention:
       * Models dependencies at multiple levels of granularity (e.g., local joint interactions vs. full-body alignment).
     - Feed-Forward Layers:
       * Refines attention-based features, projecting them into a compact representation suitable for 3D prediction.
- Advantage: The Transformer Encoder's ability to capture global spatial relationships is critical for handling occluded or partially visible joints.

3.4. . Output: Predicting 3D Key points

- Process:
  - The output of the Transformer Encoder is processed through a series of linear layers to predict the 3D positions of each key point in Cartesian coordinates (x,y,z)(x, y, z)(x,y,z).
  - The model estimates depth (z-coordinate), which is missing in the 2D input.
- Representation:
  - The final output is a skeleton represented by 3D key points connected by lines (e.g., spine, limbs).
  - This skeletal representation is essential for applications like motion tracking or animation.

4. Empirical Experiments and Analyses

The evaluation of the HOT architecture focuses on assessing its accuracy, robustness, and efficiency using quantitative metrics and qualitative insights.

4.1. Dataset: Human3.6M

- The Human3.6M dataset is the benchmark used for training and evaluation[3].
  - Content:
    * Over 3.6 million frames with annotated 2D and 3D poses.
    * Activities include walking, sitting, and interacting with objects.
  - Data Splitting:
    * Training: 70
    * Validation: 15
    * Test: 15
- Processed Data of Human3.6M
  - `Data_3d_h36m.npz` : It is a NumPy compressed file containing 3D pose data
  - `Data_2d_h36m_gt.npz` : It contains 2D ground truth annotations corresponding to the 3D data.
  - `Data_2d_h36m_cpn_ft_h36m_dbb.npz` : It includes 2D pose data detected or refined using a specific method.

4.2. Evaluation Metrics

4.2.1. Mean Per Joint Position Error (MPJPE)

- Definition: MPJPE measures the average Euclidean distance between predicted and ground-truth 3D joint positions.
  $\text{MPJPE} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{P}}_\mathbf{i} - \mathbf{P}_\mathbf{i}\right\|$
  - N is the number of joints,
  - $\hat{\mathbf{P}}_\mathbf{i}$ is the predicted position of the i-th joint
  - $\mathbf{P}_\mathbf{i}$ is the ground truth position of the i-th joint,
- Result: The HOT model achieved an MPJPE of 0.0028, outperforming baseline methods like Pose Former and fully connected networks.

4.2.2. Computational Efficiency

- Training Time: Training the model (10 epochs) on an NVIDIA GPU took ~ 3 hours , leveraging the Adam optimizer for faster convergence [4].
- Inference Time: The model processed ~ 50 frames per second , translating to an average inference time of 20 ms/frame, making it suitable for near-real-time applications.

4.3. Robustness Analysis

4.3.1. Occlusions

- Scenario: Evaluated on frames with occluded joints (e.g., arms behind the back).
- Observation:
  - The HOT model demonstrated superior performance compared to baseline methods, thanks to the global attention mechanism in the Transformer Encoder.
  - However, errors increased in cases of extreme occlusion.

4.3.2. Pose Diversity

- Scenario: Tested on diverse poses such as sitting, jumping, and stretching.
- Observation: The model generalized well to unseen poses, maintaining consistent MPJPE.

4.4. Qualitative Analysis

4.4.1. Visual Comparisons

- Predicted vs. Ground-Truth Skeletons:
  - Predicted skeletons closely aligned with ground truth, even in complex motions.
  - Visualizations show the HOT model effectively captures limb symmetry and body proportions.

4.4.2. Failure Cases

- Extreme Occlusions: For frames with multiple occluded joints, the model sometimes predicted incorrect limb positions.
- Rare Poses: Poses not well-represented in the dataset led to minor inaccuracies.

4.5. Comparative Analysis

Model	Strength	Limitation
Baseline (FCN)	Efficient for small datasets	Struggles with occlusions
Pose Former	Strong at modelling temporal relationships	Computationally expensive
HOT (Proposed)	Robust, efficient, real-time ready	Needs more training data for rare poses

Table 1. Comparative Analysis

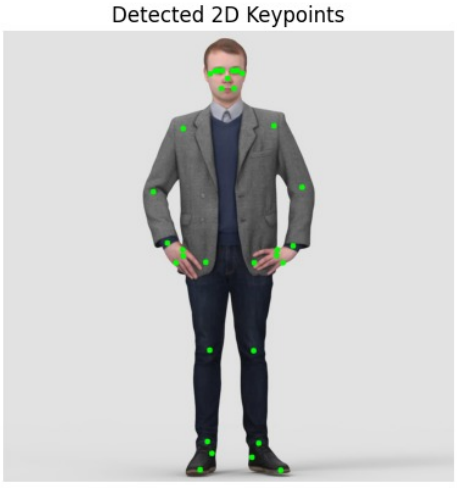

Figure 2. Detected 2D Key Points

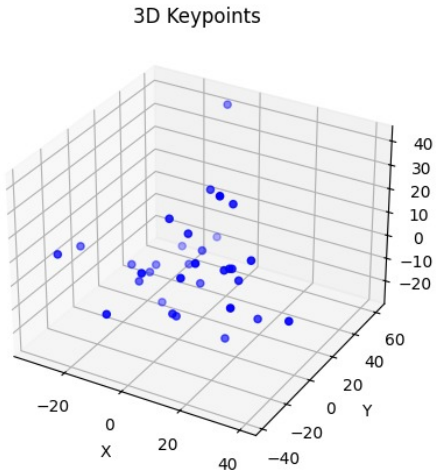

Figure 3. 3D Key Points

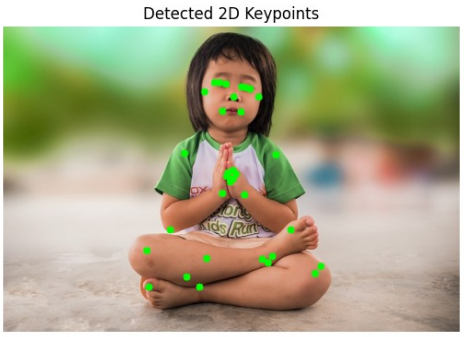

Figure 4. Detected 2D Key Points

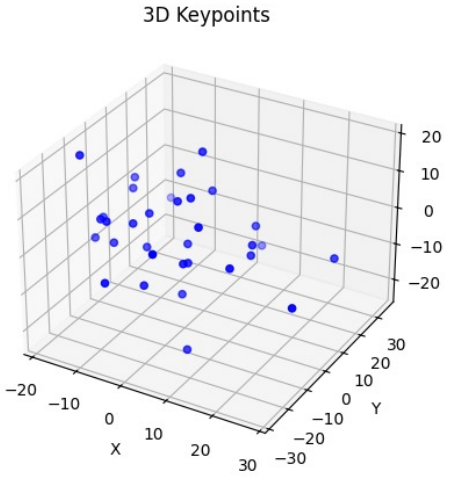

Figure 5. 3D Key Points

Figure 6. Detected 2D Key Points

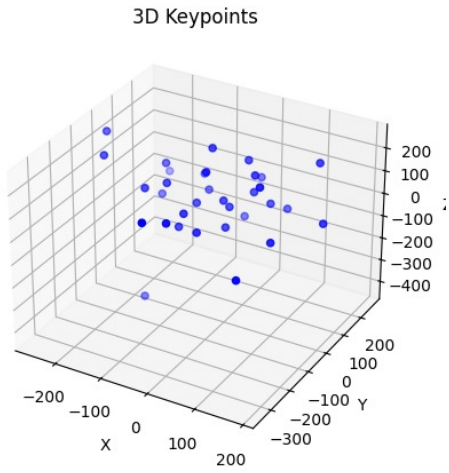

Figure 7. 3D Key Points

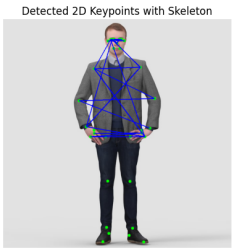

Figure 8. Detected 2D Skeleton

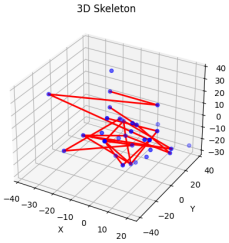

Figure 9. 3D Skeleton

5. Conclusion

The HOT architecture provides an efficient solution for 3D pose estimation, combining Hourglass Networks and Transformers. It balances accuracy and computational efficiency, enabling real-time applications across various domains. Future work includes incorporating temporal modelling for video sequences and optimizing the model for edge devices.

6. Contribution

- Vanshika Manil Parikh: Implemented preprocessing and model training pipelines, contributed to results visualization.
- Nithish Ragav Narayana Shankar: Designed the HOT architecture, conducted experiments, and analysed evaluation metrics.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 3

[4] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[5] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 1

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 1

[7] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 1

[8] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[9] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. 1