TAHM-Cap: A Transformer-Based Approach for 3D Human Mesh Reconstruction from Monocular Video

Yangxu Yan
Beijing University of Posts and Telecommunications
yanyangxu@bupt.edu.cn

Anlong Ming*
Beijing University of Posts and Telecommunications
mal@bupt.edu.cn

Bing Bai
Beijing University of Posts and Telecommunications
fengxiaofanhua@bupt.edu.cn

Yongchang Zhang
Beijing University of Posts and Telecommunications
zhangyongchang@bupt.edu.cn

Weihong Yao
Shanghai Vision Era Co., Ltd
yaoweihong@xvgate.com

Huadong Ma
Beijing University of Posts and Telecommunications
mhd@bupt.edu.cn

ABSTRACT

This paper aims to capture more accurate and stable human motion from monocular video. By integrating various factors such as 2D keypoints, camera parameters, single-frame features, and video temporal information, our pipeline captures motion with both accuracy and naturalness. Our approach presents two advantages: stable poses and accurate global trajectories. For human pose, we first train a single-frame human mesh reconstruction network, then design a transformer-based temporal module to reduce jitter. For human trajectory reconstruction, after obtaining the initial trajectory, we apply a Kalman filter for smoothing, producing more realistic motion. This method is evaluated on the Minions dataset, where it demonstrated significant improvements in key metrics such as MPJPE and trajectory jitter. Our method achieved first place in the ACM MM'24 Multimodal Human Motion Capture Challenge (Track 1), proving its effectiveness in producing accurate, temporally coherent motion predictions.

ACM Reference Format:
Yangxu Yan, Anlong Ming, Bing Bai, Yongchang Zhang, Weihong Yao, and Huadong Ma. 2018. TAHM-Cap: A Transformer-Based Approach for 3D Human Mesh Reconstruction from Monocular Video. In *Proceedings of The 5th International Workshop on Human-centric Multimedia Analysis (MM'24)Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

3D human mesh reconstruction and motion trajectory estimation from monocular video are essential tasks in computer vision, with wide-ranging applications in areas such as augmented reality (AR), virtual reality (VR), biomechanics, and human-computer interaction. Accurately modeling human motion in 3D from 2D inputs is crucial for creating immersive virtual experiences, performing detailed motion analysis, and advancing various research and industrial applications. However, these tasks remain challenging due to the inherent limitations of monocular video, such as the loss of depth information and the difficulty in maintaining temporal consistency.

Existing monocular human mesh reconstruction approaches struggle with issues like inconsistent temporal coherence and inaccuracies in predicting motion trajectories. These limitations often result in unstable and unnatural motion sequences. Moreover, depth ambiguity in monocular videos leads to a significant challenge that has yet to be fully addressed.

In this work, we propose a novel method that addresses these challenges by combining advanced temporal modeling and trajectory smoothing techniques. Our approach leverages a fine-tuned version of HMR2.0 [1] and employs a Transformer network with Rotary Positional Encoding (ROPE [7]) to capture long-term temporal dependencies. This enables more accurate motion trajectory estimation, providing a significant improvement over existing methods in terms of both accuracy and temporal consistency. To further enhance the stability of the reconstructed motion, we apply Kalman filtering to smooth the predicted trajectories, minimizing noise and jitter while preserving natural motion (see Figure 1).

The primary advantages of our method are:

- Enhanced Temporal Consistency: By leveraging a Transformer with ROPE, we capture the temporal dynamics more effectively, reducing frame-to-frame inconsistency.
- Improved Motion Estimation: The inclusion of Kalman filtering ensures smoother, more coherent trajectory predictions, mitigating common issues like jitter and unnatural motion transitions.

2 METHOD

2.1 Overview

Monocular video-based 3D human mesh reconstruction and motion trajectory estimation pose unique challenges due to the absence of depth information and the need for temporal consistency. By integrating state-of-the-art image encoding, advanced temporal modeling, and post-optimization techniques, our method forms a robust and efficient pipeline that accurately captures 3D human motion from video inputs.

*Corresponding author

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

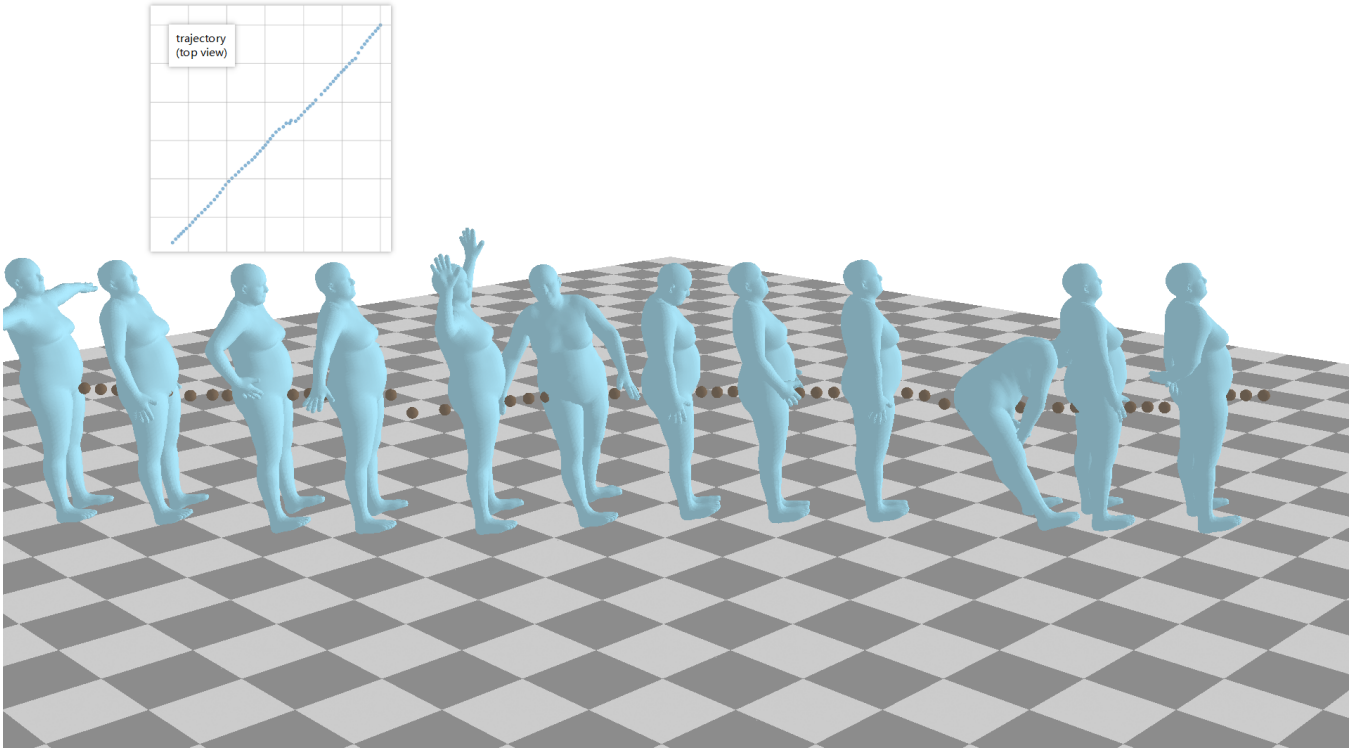

Figure 1: Detected Human and Recovered 3D Trajectory. The figure shows the detected human and corresponding 3D motion trajectory from a monocular video, including a top-view visualization of the global trajectory, demonstrating the accuracy of the reconstructed motion path.

At the key of our approach is a carefully designed pipeline that processes a monocular video sequence with known camera intrinsic parameters (K) to predict the SMPL [4] parameters (pose, shape, and translation) for each frame in the camera coordinate system. The method leverages a fine-tuned Vision Transformer (ViT [1]) as the image encoder, capable of extracting rich, global features from each video frame. By utilizing a Transformer-based temporal network with Rotary Positional Encoding (ROPE[7]), we effectively model the long-range dependencies between frames, ensuring accurate and temporally coherent motion predictions.

To address the common issue of jittery motion in monocular video-based reconstructions, we integrate Kalman filtering in the post-optimization stage. This ensures that the estimated motion trajectories are smooth and realistic, reducing the noise and instability often found in raw network outputs. Through this multi-step approach, our method not only improves accuracy in mesh reconstruction but also enhances the naturalness and coherence of the estimated motion, providing a significant advantage over existing solutions (see Figure 2).

2.2 Temporal Network

In monocular video-based human mesh reconstruction, maintaining temporal consistency across frames is a major challenge. Without a mechanism to account for the motion dynamics over time, results are prone to inconsistency, leading to unrealistic reconstructions. Our temporal network addresses this challenge by employing a Transformer architecture, designed to model temporal dependencies effectively.

Inspired by recent advances[3, 5, 6, 9], our temporal network takes the fused inputs of human bounding boxes, 2D keypoints, and camera intrinsic parameters, and processes them to generate temporally coherent predictions. To achieve this, we use a Transformer with Rotary Positional Encoding (ROPE) [5], which is particularly effective at capturing long-range temporal relationships. This allows the network to understand the continuity of motion across frames, preventing abrupt changes in the predicted mesh.

The process begins by first encoding the bounding boxes and camera intrinsic parameters into a unified representation, following the approach used in CLIFF. This unified information, along with 2D keypoints and image features from the ViT encoder, is mapped to a common space through a multi-layer perceptron (MLP), preparing the inputs for feature fusion. The fused features are then passed into the Transformer network, where the self-attention mechanism dynamically models the dependencies across frames.

The outcome is a temporally coherent sequence of SMPL parameters (pose, shape, and translation) for each frame. This ensures that the predicted human motion not only aligns with the static input features but also remains consistent across the entire sequence, providing more realistic and fluid motion trajectories.

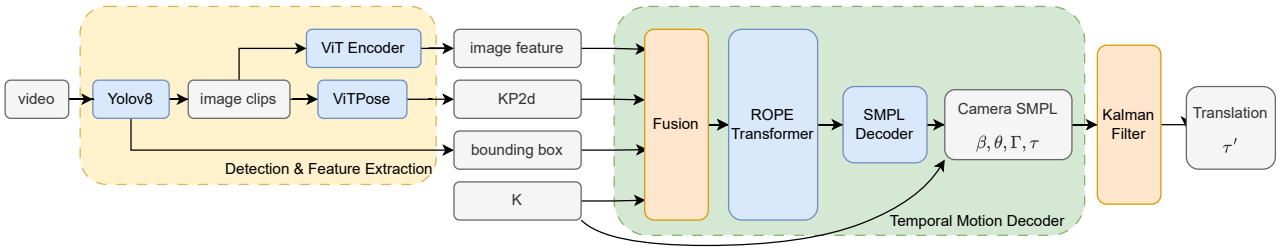

Figure 2: Overview of the Proposed Pipeline. Given a monocular video input, our method first preprocesses the frames by detecting human bounding boxes and 2D keypoints using YOLOv8 [8] and ViTPose [10]. These features are fed into an image encoder to extract static representations. The temporal network processes these features, regressing SMPL parameters such as pose, shape, and translation. Finally, Kalman filtering is applied to smooth the motion trajectories, producing stable and coherent 3D human motion throughout the sequence.

2.3 Post-Optimization with Kalman Filtering

Even with sophisticated temporal modeling, monocular video-based human motion predictions often exhibit subtle inconsistencies, particularly in complex or fast-moving sequences. These inconsistencies, commonly manifesting as jittery or unstable trajectories, can significantly detract from the realism of the reconstructed motion. To mitigate this, we introduce a Kalman filtering post-optimization step, designed to smooth the motion trajectories and ensure they exhibit natural, fluid movement across frames.

Kalman filtering is a well-established recursive estimation technique, particularly effective in dynamic systems where real-time, noisy measurements need to be smoothed. In our approach, we apply Kalman filtering to the predicted SMPL translation and pose parameters across frames, effectively reducing noise and stabilizing the predicted trajectory. The Kalman filter uses a combination of the current state and observed data, adjusting the predicted trajectory incrementally based on the likelihood of noise or abrupt changes in motion.

By integrating Kalman filtering into our pipeline, we ensure that the final motion trajectories are coherent and natural, addressing the common issue of jitter in monocular video reconstructions. This step is essential for applications requiring smooth, realistic motion, such as AR/VR or human motion analysis, where continuity and temporal coherence are critical.

2.4 Loss Function.

To ensure that the network learns to predict accurate and stable SMPL parameters, we carefully design our loss function to balance multiple objectives, each targeting a key aspect of the reconstruction process. Our loss function is a weighted combination [2] of 2D projection loss, 3D loss, and SMPL parameter loss, with a specific emphasis on improving the translation parameters to ensure realistic motion trajectories.

2D Projection Loss (L_{kp2D}): This loss penalizes the discrepancy between the projected 3D keypoints and the ground truth 2D keypoints. By enforcing consistency between the predicted and observed 2D positions, we guide the network to maintain alignment between the reconstructed human mesh and the 2D input. We use L_1 loss to compute this error:

$$L_{\text{kp2D}} = ||\pi(X) - x^*||_1 \tag{1}$$

3D Keypoint Loss (L_{kp3D}): To ensure the network accurately predicts the 3D structure, we minimize the distance between the predicted 3D keypoints and the ground truth 3D keypoints. This encourages the model to produce realistic 3D meshes that match the actual body structure. Similar to the 2D loss, we employ L_1 loss:

$$L_{\text{kp3D}} = ||X - X^*||_1 \tag{2}$$

SMPL Parameter Loss (L_{smpl}): This term focuses on constraining the SMPL parameters (pose θ, shape β, and translation τ) to ensure they remain close to the ground truth. We assign higher weights to the pose and translation parameters to emphasize accurate motion estimation, using L_2 loss for this component:

$$L_{\text{smpl}} = ||\theta - \theta^*||_2^2 + ||\beta - \beta^*||_2^2 + ||\tau - \tau^*||_2^2 \tag{3}$$

By combining these loss components, our training process effectively guides the network to balance both spatial accuracy and temporal coherence. The higher emphasis on translation and pose parameters ensures that the predicted motion remains stable and realistic, while the 2D and 3D losses keep the reconstructed human mesh closely aligned with the input video.

3 EXPERIMENTS

To evaluate the effectiveness of our proposed method, we conducted experiments on the Minions dataset, measuring metrics such as Mean Per Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and trajectory jitter. Our experimental design prioritized iterative improvements to key components, focusing on refining specific metrics rather than performing extensive comparisons with baseline methods.

3.1 Experimental Details

We conducted the experiments using two NVIDIA RTX 3090 GPUs, taking advantage of their substantial computational power to handle the intensive training required for the temporal network and ViT encoder. The overall training pipeline was split into two distinct stages, followed by a post-optimization step.

3.1.1 Stage 1: Temporal Network Training. We began by training the temporal network for 200 epochs using the pre-trained HMR2.0 image encoder (weights from 4DHumans). The goal was to establish an initial baseline for motion parameter prediction, particularly focusing on pose and shape reconstruction. The temporal network, using a Transformer, was trained with a learning rate of $1e^{-5}$. At this stage, the primary objective was to capture the temporal dependencies and provide a coherent estimation of human motion over time.

3.1.2 Stage 2: Kalman Filtering for Trajectory Smoothing. Once the initial SMPL parameters were predicted, we applied Kalman filtering as a post-processing step to reduce noise and jitter in the motion trajectories. This step significantly enhanced the continuity and naturalness of the reconstructed motion.

3.1.3 Stage 3: Enhanced Training with ROPE, Stronger Data Augmentation, and Multi-Task Learning. In the third stage, we introduced several key enhancements to the training process to achieve better performance metrics. We incorporated Rotary Position Embedding (ROPE) into the Transformer-based temporal network, which allowed the model to better capture positional information across sequences. Additionally, we applied stronger data augmentation techniques to enrich the training dataset, improving the model's generalization capabilities. We also adopted a multi-task learning framework, enabling the model to learn pose estimation alongside related tasks simultaneously. This approach leveraged shared representations and provided regularization benefits, leading to significant improvements in the accuracy and robustness of the predicted SMPL parameters.

3.2 Experimental Results

Table 1 presents the results of different stages of our method, with MPJPE, PA-MPJPE, and trajectory jitter as evaluation metrics.

HMR2.0+T: This baseline combines the pre-trained HMR2.0 encoder with the temporal network, resulting in an MPJPE of 391.68mm and a jitter score of 3.78.

HMR2.0+T+KF: Adding Kalman filtering for post-optimization reduced jitter to 0.60 and improved MPJPE to 369.39mm, highlighting the effectiveness of trajectory smoothing.

HMR2.0+T+KF (Enhanced): We enhanced the temporal network by incorporating Rotary Position Embedding (ROPE), applying stronger data augmentation, and adopting a multi-task learning framework. These improvements reduced the MPJPE to 271.48 mm and jitter to 0.42, attributed to better temporal feature extraction and more effective noise reduction strategies.

The results show a clear progression in performance through the stages of our method. Initially, the baseline (HMR2.0+T) provided reasonable results, though with notable jitter in motion predictions. The introduction of Kalman filtering (HMR2.0+T+KF) significantly reduced jitter, demonstrating the importance of post-optimization for stabilizing motion trajectories. Further enhancements in the temporal modeling and optimization techniques (HMR2.0+T+KF (Enhanced)) brought additional improvements, particularly in MPJPE, as the model was able to capture temporal dependencies more effectively, resulting in more accurate and stable pose and shape estimations.

Models	MPJPE	PA-MPJPE	Jitter
HMR2.0+T	391.68	11.41	3.78
HMR2.0+T+KF	369.39	**11.30**	0.60
HMR2.0+T+KF(Enhanced)	**271.48**	11.60	**0.42**

Table 1: Quantitative Results at Different Stages of the Pipeline. The table presents the performance metrics at various stages of the pipeline, including MPJPE, PA-MPJPE, and trajectory jitter.

4 CONCLUSION

In this paper, we presented a method for 3D human mesh reconstruction and motion trajectory estimation from monocular video, which won first place in the ACM MM'24 Multimodal Human Motion Capture Challenge (Track 1). Our method leverages a ViT-based image encoder, a Transformer temporal network, and Kalman filtering for trajectory smoothing, achieving state-of-the-art results on the Minions dataset. Future work will focus on refining our approach to further enhance its robustness and applicability in real-world scenarios.

REFERENCES

[1] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14783–14794.

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7122–7131.

[3] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*. Springer, 590–606.

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[5] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. *arXiv preprint arXiv:2409.06662* (2024).

[6] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. 2024. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2070–2080.

[7] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

[8] Rejin Varghese and M Sambath. 2024. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE, 1–6.

[9] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. 2024. TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. *arXiv preprint arXiv:2403.17346* (2024).

[10] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems* 35 (2022), 38571–38584.