Transformer-Based Full-Body Pose Estimation for Rehabilitation via RGB Camera and IMU Fusion

Yuanshuo Tan
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
tanyuanshuo@sjtu.edu.cn

Xinyuan He
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
hexinyuan@sjtu.edu.cn

Guoxing Liu
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
lgxing1900@sjtu.edu.cn

Licheng Zhong
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
zlicheng@alumni.sjtu.edu.cn

Huiming Pan
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
panhuiming@sjtu.edu.cn

Kezhe Zhu
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
kezhe_zhu@sjtu.edu.cn

Peter B. Shull*
School of Mechanical Engineering
Shanghai Jiaotong University
Shanghai, China
pshull@sjtu.edu.cn

Abstract*—Rehabilitation training plays a vital role in the recovery of lower back and cervical spine function. Human pose estimation can support this process by guiding and evaluating rehabilitation movements. However, specialized rehabilitation exercises often involve severe self-occlusions, posing significant challenges for vision-based pose estimation methods. We thus propose a full-body pose estimation framework tailored for rehabilitation exercises, which fuses monocular images and inertial measurement unit (IMU) signals using a temporal transformer. Multimodal data was collected from six subjects performing 22 specialized rehabilitation movements (e.g., single-leg open book, cross-leg body rotation, standing iliotibial band stretch, standing lumbar extension). The collected data comprises synchronized images, 2D and 3D human keypoint coordinates, and IMU signals. Our approach first employs a convolutional neural network (CNN) to extract 2D keypoints from image sequences. These keypoints, combined with IMU signals, are then processed by a temporal transformer to estimate 3D joint coordinates. On the collected data, a vision-only baseline yields a 2D joint position error of 7.33 $\pm$ 2.08 pixels and a 3D joint error of 10.05 $\pm$ 2.67 cm. In comparison, the proposed method achieves lower errors, with 5.50 $\pm$ 0.75 pixels for 2D joints and 8.27 $\pm$ 1.03 cm for 3D joints. By leveraging inertial data, our method enhances the robustness of pose estimation under challenging conditions such as self-occlusion, demonstrating its potential for both clinical and home-based rehabilitation applications.

***Index Terms*—human pose estimation, rehabilitation exercises, multimodal sensing**

I. Introduction

Pain conditions such as lower back pain [1] and spine disorders [2] affect a substantial portion of the global population, creating a growing demand for effective and accessible rehabilitation solutions. Rehabilitation training is critical for restoring physical function and improving patients' quality of life. However, whether rehabilitation occurs in clinical settings or at home, current methods often lack visual guidance and quantitative tools for tracking and assessing patient movements. To address these limitations, human pose estimation approaches can be integrated into rehabilitation scenarios. These systems estimate keypoint positions or joint angles using either visual data [3], [4] or inertial sensor inputs [5]–[7]. Such integration enables visual feedback, objective tracking, and movement evaluation in both clinical and home environments, thereby improving rehabilitation outcomes.

However, both vision-based and IMU-based pose estimation methods face challenges in rehabilitation contexts. Vision-based approaches are limited by the camera's field of view and struggle with severe self-occlusion. Sparse IMUs make inverse kinematics estimation ill-posed, while using too many IMUs can be intrusive and impractical.

To address these challenges, this paper proposes a novel pose estimation approach that leverages an RGB camera and six IMUs. The approach first uses a CNN backbone to extract 2D keypoint coordinates from monocular image sequences. A temporal transformer then fuses these 2D keypoints with IMU signals, modeling temporal dependencies to estimate 3D joint coordinates. We also collect a large-scale set of multimodal rehabilitation motion data, which includes images, 2D and 3D keypoint annotations, and IMU signals. Our approach achieves a 2D joint error of 5.50 $\pm$ 0.75 pixels and a 3D joint error of 8.27 $\pm$ 1.03 cm on the collected data, offering a reliable

*Corresponding author. This work is supported by the National Natural Science Foundation of China under grant W2441018.

solution for estimating complex rehabilitation movements.

II. Methods

A. Data Collection

We collected over 230,000 frames from 6 healthy subjects (4 male and 2 female; height: $1.69 \pm 0.05\text{m}$; mass: $63.2 \pm 6.9\text{kg}$; BMI: $22.3 \pm 3.0\text{kg/m}^2$) performing 22 professional rehabilitation exercises (Table I). All the subjects provided written informed consent before being tested, and the experimental procedure was reviewed and approved by the ethics committee of Shanghai Jiao Tong University (No. E2021013P).

Each subject wore a motion capture suit with 39 reflective markers attached according to the Plug-in Gait scheme [8], along with six IMUs (MTw, Xsens Technologies, Netherlands) positioned on the head, abdomen, forearms, and shanks (Fig. 1). The Vicon system and IMUs were synchronized at 60 Hz, and the global inertial coordinate frame was calibrated to align with the Vicon coordinate system. An RGB camera (Nikon Z5, 1920×1080@60 Hz) was positioned beside the motion capture area to record the movements. Camera-to-Vicon calibration was performed using the method proposed by Zhang et al. [9], and synchronization was achieved via a clapping motion.

Participants performed the exercises in three postures–supine, standing, kneeling, with each movement repeated 3–5 times per side.

TABLE I
Rehabilitation Exercises Classification

Posture	Rehabilitation Exercise	Targeted Area
Supine	McGill curl-up	Pelvis
	External rotator stretch	Pelvis
	Single-leg open book	Lumbar region
	Knee-flexion transverse abdominis activation	Thoracic spine
	Dead bug	Lumbar region
	Double-arm lift glute bridge	Lumbar region
	Cross-leg body rotation	Pelvis
	Sciatic nerve mobilization	Hip-knee complex
	Dynamic intra-abdominal pressure training	Lumbar region
Standing	Shortstop squat	Hip-knee complex
	Basic squat	Hip-knee complex
	Standing iliotibial band stretch	Hip-knee complex
	Sumo squat	Hip-knee complex
	Standing lumbar extension	Lumbar region
	Single-leg contralateral ankle touch	Lumbar region
Kneeling	Catcow full spine	Thoracic spine
	Unilateral bird dog	Lumbar region
	Quadruped posterior reach	Lumbar region
	Quadruped donkey kick	Lumbar region
	Oscillatory bird dog	Lumbar region
	Kneeling dynamic frog stretch	Pelvis
	Kneeling thoracic rotation mobilization	Thoracic spine

B. Data Processing

1) 2D and 3D Keypoint Coordinates: A human skeleton model is first constructed in Visual3D (C-Motion, MD, USA) based on 3D marker positions. The ground truth 3D coordinates of 16 human keypoints (Fig. 1) are then derived from this model. The corresponding 2D keypoint coordinates are computed using calibrated camera intrinsics and extrinsics.

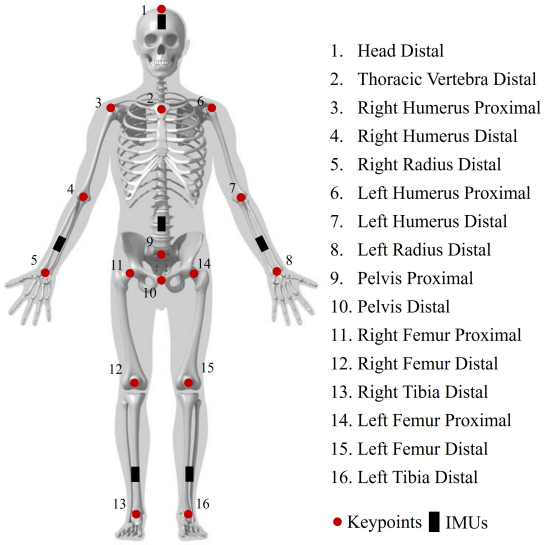

Fig. 1. Illustration of the 16 annotated body keypoints and the 6 IMU sensor placements. IMUs are positioned on the head, abdomen, left/right forearms, and left/right shanks.

2) IMU Signals: During data processing, the rotation matrices and acceleration values from the IMUs are transformed into the RGB camera coordinate system to ensure spatial alignment between inertial and visual data.

3) Images: Images are cropped, resized to 256x256 pixels, and centered on the subject to improve computational efficiency.

4) Data Augmentation: The image data augmentation methods include random rotation, scaling, translation, color adjustments, and Gaussian occlusion. All ground truth annotations are transformed accordingly to maintain consistency with image-level augmentations.

C. Human Pose Estimation Approach for Rehabilitation Exercises

We propose a two-stage pose estimation framework (Fig. 2). The model takes as input a sequence of L consecutive RGB images along with rotation and acceleration signals from six IMUs. The output is the 3D relative coordinates of human keypoints in the Lth frame, represented relative to the root joint (pelvis). The model consists of two stages: one for extracting 2D human keypoints from images, and the other for estimating 3D keypoints from the 2D keypoints and IMU signals.

1) Extracting 2D Human Keypoints from Images: A sequence of L consecutive images is fed into a pre-trained ResNet34 [10] backbone, which produces L heatmaps of size $N \times w \times h$, denoted as $\{H_j^{(t)} | j = 1, 2 \dots N, t = 1, 2 \dots L\}$, where N is the number of keypoints. The coordinates of each keypoint in heatmap space are computed and then mapped to the original image space, resulting in $\{p_j^{(t)} \in \mathbb{R}^2 | j = 1, 2 \dots N, t = 1, 2 \dots L\}$.

2) Estimating 3D Keypoints from 2D Keypoints and IMU Signals: Rotation matrices and acceleration from the

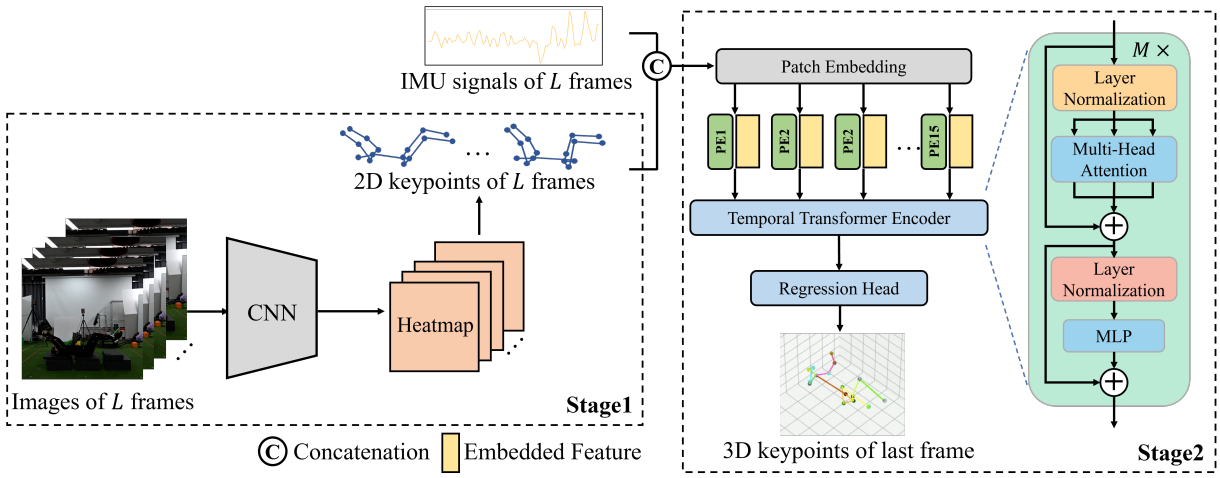

Fig. 2. The framework of the proposed pose estimation approach. Stage 1: A CNN is used to extract heatmaps from the image sequence, which are then processed to obtain the sequential 2D human keypoints. Stage 2: The 2D human keypoint sequence is concatenated with the IMU signals and fed into a Temporal Transformer to extract the 3D human keypoint coordinates of the final frame.

six IMUs are transformed into the RGB camera coordinate frame, resulting in $\{R_j^{(t)}|j=1,2\dots6,t=1,2\dots L\}$ and $\{a_j^{(t)}|j=1,2\dots6,t=1,2\dots L\}$. Each IMU provides 12 values per frame (9 from the rotation matrix and 3 from the acceleration), yielding a total of 72 values across all six IMUs. Combined with 2D keypoints, the input sequence is represented as $X \in \mathbb{R}^{L\times(2N+72)}$.

Each frame input $x_i \in \mathbb{R}^{2N+72}$ is first projected into a latent space using a shared linear embedding matrix $E \in \mathbb{R}^{(2N+72)\times C}$, where C is the embedding dimension. This yields a sequence of embedded features $Z \in \mathbb{R}^{L\times C}$. To encode temporal order, a learnable positional embedding matrix $E_{\text{pos}} \in \mathbb{R}^{L\times C}$ is added element-wise to Z, forming the final input to the Transformer encoder:

$$Z_0 = Z + E_{\text{pos}} = [x_1E, x_2E, \dots, x_LE] + E_{\text{pos}} \quad (1)$$

The temporal Transformer encoder follows the architecture proposed by Zheng et al. [11]. It produces an output $Y \in \mathbb{R}^{L\times C}$, which is passed through a convolutional layer and a regression head to predict the 3D joint coordinates of the last frame, denoted as $\hat{p} \in \mathbb{R}^{3N}$.

3) Loss Function: To jointly supervise the heatmaps and the 3D keypoint coordinates of the last frame, the loss function is designed as:

$$L = \lambda_{\text{H}} L_{\text{Heatmap}} + \lambda_{\text{J}} L_{\text{Joints}} \quad (2)$$

where:

$$L_{\text{Heatmap}} = \frac{1}{Nwh}\sum_{i=1}^{L}\sum_{j=1}^{N}\sum_{x=1}^{w}\sum_{y=1}^{h}[H_j^{(i)}(x,y) - \hat{H_j^{(i)}}(x,y)]^2 \quad (3)$$

$$L_{\text{Joints}} = \frac{1}{N}\sum_{j=1}^{N}||p-\hat{p}||_2^2 \quad (4)$$

where $\|\cdot\|_2$ denotes the L2 norm. The loss weights are empirically set to $\lambda_{\text{H}} = 10.0$ and $\lambda_{\text{J}} = 2.0$.

D. Performance Evaluation

We adopt leave-one-subject-out cross-validation on the collected data to train and evaluate our approach, and compare it with a vision-only baseline, Integral Pose [12]. The accuracy of estimated keypoints is assessed using 2D-MPJPE (Mean Per Joint Position Error in 2D), 3D-MPJPE (Mean Per Joint Position Error in 3D), and PA-MPJPE (Procrustes Aligned Mean Per Joint Position Error). The temporal sequence length L for both image and IMU inputs is set to 10.

An ablation study is conducted to evaluate the contribution of IMU signals to pose estimation. Separate models are trained using vision + 6 IMUs, vision + 4 IMUs, vision + 2 IMUs, and vision only, and evaluated using the same metrics.

III. Results

A quantitative comparison between the proposed method and the vision-only baseline, Integral Pose, was conducted using three metrics on data collected from professional rehabilitation exercises (Table II). The proposed method consistently outperforms Integral Pose across all evaluation metrics. For 2D pose estimation, the proposed approach achieves a lower error of 5.50 $\pm$ 0.75 pixels, compared to 7.33 $\pm$ 2.08 pixels for Integral Pose. For 3D keypoints, the proposed approach achieves an error of 8.27 $\pm$ 1.03 cm, while Integral Pose yields 10.05 $\pm$ 2.67 cm. Structural accuracy, measured by PA-MPJPE, is also improved: 6.31 $\pm$ 0.92 cm for our method versus 7.47 $\pm$ 1.60 cm for Integral Pose. Notably, our method shows both better accuracy and lower standard deviation across all subjects, indicating enhanced robustness and generalizability in rehabilitation scenarios.

In the ablation study, the vision + IMU model outperforms the vision-only model across all three metrics (Fig. 3), with the best performance achieved using four IMUs. The 2D-MPJPE is 5.16 pixels, the 3D-MPJPE is 7.44 cm, and the PA-MPJPE is 5.81 cm.

TABLE II
COMPARISON OF POSE ESTIMATION PERFORMANCE (MPJPE IN PIXELS/CM)

Metric	Method	S1	S2	S3	S4	S5	S6	Avg (± std)
2D-MPJPE	Ours	**6.28**	**5.05**	**6.77**	**5.12**	**5.05**	**4.74**	**5.50 ± 0.75**
	Integral	8.23	9.29	10.40	5.16	5.49	5.41	7.33 ± 2.08
3D-MPJPE	Ours	**9.78**	**8.41**	**9.29**	**7.09**	**8.07**	6.99	**8.27 ± 1.03**
	Integral	11.00	13.57	12.15	7.10	8.83	**6.14**	10.05 ± 2.67
PA-MPJPE	Ours	8.01	**6.24**	**6.67**	**5.41**	**6.29**	**5.21**	**6.31 ± 0.92**
	Integral	**7.72**	10.57	8.08	5.92	6.38	6.14	7.47 ± 1.60

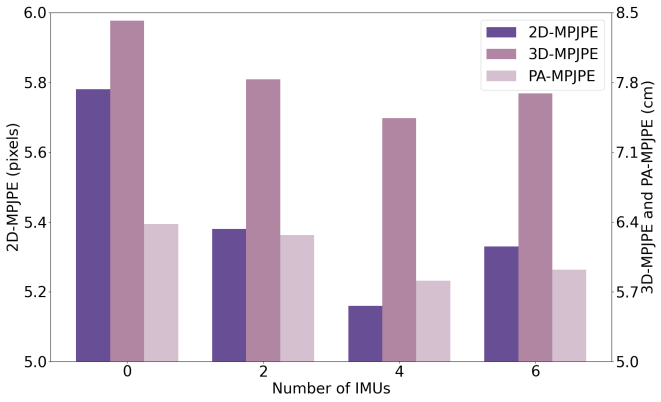

Fig. 3. Ablation study on the impact of input modality on pose estimation performance. The model is trained and tested on the collected rehabilitation motion data using different input modalities: vision only, vision + 2 IMUs, vision + 4 IMUs and vision + 6 IMUs.

IV. DISCUSSION & CONCLUSION

This paper introduces a vision-IMU fusion method for pose estimation in rehabilitation, aiming to enhance accuracy through inertial sensing. The proposed approach outperforms the vision-only baseline, Integral Pose, across all metrics, achieving both higher accuracy and greater robustness. In rehabilitation, where subject variability is high, lower deviations reflect greater robustness and better user experience.

Our ablation study confirms the strong complementary effect of IMU data on visual input, with the fusion of vision and IMU achieving higher accuracy than vision alone. Notably, the combination of vision with four IMUs outperforms that with six. The two additional IMUs, placed on the head and abdomen, contribute less because these regions exhibit limited motion during many rehabilitation exercises, particularly in supine and kneeling postures. As a result, the signals from these sensors may introduce redundancy or noise during model training.

Previous works have also explored pose estimation in rehabilitation contexts. Cotton [13] proposed a method that integrates vision and IMU data, but it processes each frame independently, which may result in abrupt pose transitions. Li et al. [14] developed a home-based pose estimation system, but their work focused on lower-body rehabilitation rather than full-body recovery. In contrast, we focus on full-body rehabilitation movements and have collected a unique multimodal dataset from professionally guided exercises. By employing temporal fusion of vision and IMU signals, our approach enables continuous full-body pose estimation in complex rehabilitation movements, thereby helping fulfill the growing need for intelligent rehabilitation solutions.

Overall, this paper proposes a pose estimation method that fuses data from a monocular RGB camera and six IMUs, validated using motion data collected from specialized rehabilitation movements. The results confirm its accuracy, while ablation study demonstrates the complementary role of IMUs. These findings highlight the potential of vision-IMU fusion for rehabilitation. Future work will focus on improving real-time performance and enabling interactive movement guidance to further enhance user experience.

REFERENCES

[1] S. Chen, M. Chen, X. Wu, S. Lin, C. Tao, H. Cao, Z. Shao, and G. Xiao, "Global, regional and national burden of low back pain 1990–2019: a systematic analysis of the global burden of disease study 2019," *Journal of orthopaedic translation*, vol. 32, pp. 49–58, 2022.

[2] G. Y. Shim, J. Choi, H. J. Kim, R. Kwon, M. S. Kim, M. C. Yoo, M. Rahmati, W. Cho, and D. K. Yon, "Global, regional, and national burden of spine pain, 1990-2019: a systematic analysis of the global burden of disease study 2019," *Archives of physical medicine and rehabilitation*, vol. 105, no. 3, pp. 461–469, 2024.

[3] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[4] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1944–1953.

[5] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[6] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu, "Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 167–13 178.

[7] H. Pan, H. Wang, D. Li, K. Zhu, Y. Gao, R. Yin, and P. B. Shull, "Automated, imu-based spine angle estimation and imu location identification for telerehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 21, no. 1, p. 96, 2024.

[8] IDMIL. (2020) Plug-in-gait marker placement. IDMIL. [Online]. Available: https://www.idmil.org/wp-content/uploads/2020/08/Plug-in-GaitMarkerPlacement.pdf

[9] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2002.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 656–11 665.

[12] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545.

[13] R. J. Cotton, "Kinematic tracking of rehabilitation patients with markerless pose estimation fused with wearable inertial sensors," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 508–514.

[14] Y. Li, C. Wang, Y. Cao, B. Liu, J. Tan, and Y. Luo, "Human pose estimation based in-home lower body rehabilitation system," in *2020 International Joint Conference on Neural Networks (IJCNN)*. Ieee, 2020, pp. 1–8.