iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81575

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

AI-Powered Digital Twin for Physical Health Using Monocular 3D Human Mesh Recovery

Vyshnavi M, Anusha S A[1], Anushree U[2], Hithushree K Gowda[3], Chandana B R[4]
Department of Computer Science and Engineering, Sapthagiri College of Engineering

Abstract: ***Body digitization has long remained inaccessible outside clinical and high-end sports settings, requiring expensive lab hardware, trained operators, and controlled environments. This paper surveys the landscape of monocular 3D human body reconstruction, posture analysis, and digital twin generation for physical health applications. We examine the progression from marker-based capture and RGB-D sensing to transformer-driven human mesh recovery (HMR) and parametric body modelling using the Skinned Multi-Person Linear (SMPL) framework. Existing methods are evaluated across accuracy, processing speed, hardware cost, and practical deployability. We then present a proposed system that addresses identified gaps by delivering full 3D body mesh reconstruction from a short smartphone video through a FastAPI and Three.js web pipeline, requiring no dedicated sensors. The architecture leverages HMR 2.0 with a Vision Transformer backbone, SMPL shape and pose parameterisation, and optional Blender-based rendering, achieving sub-five-second inference without specialist equipment. The survey covers benchmark datasets, evaluation criteria, and unresolved challenges including occlusion handling, generalisation under lighting variation, and real-time performance on edge devices. Our findings suggest that the convergence of transformer architectures, parametric body models, and lightweight web infrastructure is making consumer-grade personalised digital twins a practical reality.***

Keywords: Digital Twin, 3D Human Mesh Recovery, HMR 2.0, SMPL, Posture Analysis, Monocular Video, Deep Learning, Vision Transformer, Physical Health, Virtual Try-On, FastAPI, Three.js

I. INTRODUCTION

Physical health monitoring has traditionally depended on periodic clinical visits, subjective self-reports, or expensive equipment confined to hospitals and biomechanics labs. The growing global burden of musculoskeletal disorders, sedentary lifestyles, and the rising demand for personalised fitness guidance has created strong motivation to bring accurate body analysis into everyday consumer devices. Smartphones are now carried by over five billion people worldwide, each equipped with a camera capable of capturing high-resolution video — yet this sensor has remained largely underutilised for structured health assessment.The concept of a digital twin, originally introduced in industrial contexts to mirror physical assets in software, has begun to migrate toward the human body. A personalised body digital twin is a continuously updated virtual model of an individual that reflects their actual physical dimensions, posture patterns, and movement characteristics. Such a model could serve as the foundation for remote physiotherapy, longitudinal fitness tracking, ergonomic risk assessment, and immersive garment fitting, among other applications.

Achieving a reliable body digital twin historically required either marker-based motion capture systems costing tens of lakhs of rupees, RGB-D depth cameras with constrained operating ranges, or inertial sensor suits demanding expert calibration. Recent advances in deep learning — particularly transformer-based human mesh recovery and parametric body models — have opened the possibility of recovering accurate 3D body geometry from ordinary monocular video. However, no single consumer-facing system has yet consolidated mesh reconstruction, posture analytics, and web accessibility into a unified pipeline.

This survey reviews the state of the art in monocular 3D body reconstruction, identifies the limitations of existing approaches, and presents a proposed system designed to close these gaps. Section II presents the literature survey. Section III analyses existing systems and their shortcomings. Section IV describes the proposed system and its innovations. Section V details the methodology. Section VI outlines system requirements. Section VII discusses key insights and open challenges, and Section VIII concludes the paper.

II. LITERATURE SURVEY

Research in 3D human body reconstruction has accelerated sharply over the past few years, driven by the availability of large-scale annotated datasets and the maturing of attention-based neural architectures.

©IJRASET: All Rights are Reserved

The works surveyed below represent the most relevant contributions across multi-view recovery, clothed body reconstruction, transformer-based HMR, photorealistic digital twin generation, real-time posture analysis, tokenised pose representation, diffusion-based novel-view synthesis, expressive full-body estimation, Gaussian splatting, and unified human editing.

1) Pavlakos et al. [1] (2022) — Human Mesh Recovery from Multiple Shots

This work introduced an attention-based transformer model that fuses 2D evidence drawn from multiple short video clips of the same subject. By aggregating multi-view temporal context, the system achieved an 8.4 percent reduction in per-joint position error compared with single-view baselines and noticeably improved reconstruction quality on unconstrained, in-the-wild footage. The principal limitation is that at least two separate shots of the individual must be available, which is impractical for spontaneous single-take scenarios, and the architecture does not support real-time inference speeds.

2) Moon et al. [2] (2022) — ClothWild: In-the-Wild 3D Clothed Human Reconstruction

ClothWild addressed the problem of recovering not just the underlying body shape but also the overlying clothing geometry from a single image. The approach combined an SMPL-based body estimator with an implicit neural clothing layer that models garment offsets on top of the naked body mesh. It was among the first methods to jointly recover body shape and clothing from unconstrained photographs. Reconstruction quality degraded for loose or layered garments, and frame-level processing speed of approximately two seconds per image made the approach unsuitable for video-rate applications.

3) Goel et al. [3] (2023) — HMR 2.0: A Single-Stage Transformer

HMR 2.0 replaced the convolutional encoder backbone of earlier HMR pipelines with a Vision Transformer (ViT), enabling global self-attention over image tokens. The result was state-of-the-art per-joint accuracy at 30 frames per second on a single consumer GPU. The transformer's attention mechanism provided improved handling of partial occlusions, where earlier CNN models would fail due to missing local features. The trade-off is a high GPU memory footprint that complicates deployment on mobile devices, and fine-tuning on custom datasets carries significant computational cost.

4) Lattas et al. [4] (2023) — FitMe: Deep Photorealistic 3D Morphable Face Models

FitMe demonstrated that differentiable rendering combined with neural reflectance estimation could produce remarkably high-fidelity head-to-toe digital twin fragments from single selfie images. The use of learned reflectance enabled photorealistic skin textures that closely matched real subjects under novel lighting. Limitations included imprecise neck and shoulder boundary geometry and the absence of hair modelling, leaving the overall body representation incomplete for fitness applications.

5) Mehta et al. [5] (2023) — Digital Twins for Fitness Tracking

This work targeted mobile-optimised fitness monitoring by combining Google's MediaPipekeypoint detector with a lightweight HMR backend. Running at 22 frames per second on mid-range Android devices, the system demonstrated high correlation with physiotherapist assessments on a set of common gym exercises. However, body shape estimation was absent — users were assigned a generic mesh rather than a personalised one — and analysis was restricted to the sagittal plane, preventing diagnosis of lateral postural deviations.

6) Dwivedi et al. [6] (2024) — TokenHMR

TokenHMR reformulated human pose estimation as a sequence prediction problem by tokenising the body's joint configuration into discrete codebook entries via a Vector Quantised Variational Autoencoder. This representation allowed the pose module to be plugged into Large Language Model-style architectures for multi-modal tasks. Robustness to occlusion improved relative to continuous-regression baselines. Expressiveness was constrained by the finite codebook, and training computational costs were substantially higher than standard regression approaches.

7) Shao et al. [7] (2024) — Human4DiT: Free-View Video Generation

Human4DiT applied a diffusion model conditioned on SMPL pose and shape parameters to synthesise photorealistic video from novel viewpoints. Inference was reported as fifteen times faster than Neural Radiance Field-based methods for the same task. The output quality was compelling for virtual try-on and digital content creation use cases. Inference time remained too high for real-time digital twin display during live activity sessions, making it unsuitable as a health monitoring backend without significant hardware investment.

8) Cai et al. [8] (2024) — SMPLest-X

SMPLest-X scaled the SMPL-X expressive body model by training on five million pseudo-annotated images, achieving a 35 percent improvement in hand pose accuracy over prior expressive-body methods. Full-body including hands and face recovery opened applications in sign language recognition and rehabilitation. The resulting model was too large for mobile deployment, running at only 8 FPS on a high-end workstation, and generalisation to unseen subjects from non-standard viewpoints remained partially unsolved.

9) Chen et al. [9] (2025) — GaussianBody

GaussianBody represented a clothed human body using 3D Gaussian primitives anchored to the SMPL skeleton, enabling real-time rendering at 120 FPS with high-fidelity clothing wrinkles and fabric dynamics. This was a significant leap over mesh-based rendering for virtual try-on. The method required approximately twenty minutes of optimisation per subject and relied on a multi-view capture setup during that phase, limiting its applicability for casual single-video onboarding of new users.

10) Li et al. [10] (2025) — UniHuman

UniHuman unified pose transfer, appearance transfer, and shape editing into a single diffusion model, allowing users to modify any aspect of a person's appearance in a photograph. The unified formulation simplified deployment compared with separate specialist models. Inference speed of four to eight seconds per edit and the constraint that all edits must remain within the SMPL parameter space limited the depth of modifications that could be explored.

III. EXISTING SYSTEMS AND THEIR LIMITATIONS

Before proposing a new architecture, it is instructive to characterise the systems that practitioners currently rely on, organised by their primary sensing modality.

A. Marker-Based Motion Capture

Commercial systems such as Vicon and OptiTrack place retro-reflective markers on anatomical landmarks and triangulate their 3D positions using banks of infrared cameras. The result is a highly accurate skeletal reconstruction at frame rates exceeding 200 Hz. These systems are considered the gold standard in clinical biomechanics and film visual-effects production. Their cost ranges from approximately twenty to fifty lakh rupees for a full studio installation. Each session requires forty-five to ninety minutes of marker placement and calibration by a trained technician, and the subject must remain within a small capture volume. No consumer can reasonably use such a system for routine health monitoring.

B. Wearable Inertial Sensor Suits

Inertial Measurement Units, each combining accelerometers, gyroscopes, and magnetometers, can estimate joint orientations when distributed across body segments and fused via a Kalman filter. Systems like Xsens MVN offer good coverage of global motion without camera line-of-sight constraints. Gyroscope bias causes slow accumulative drift that degrades absolute position accuracy over sessions lasting more than a few minutes. The suits require careful calibration, are uncomfortable for extended wear, and still do not recover body shape — only skeletal pose.

C. RGB-D Depth Camera Systems

Microsoft Kinect and Intel RealSense cameras combine a colour stream with a structured-light or time-of-flight depth sensor to generate registered colour-plus-depth frames. Skeleton detection libraries built on these streams can produce real-time 3D joint estimates without body-worn devices. Performance degrades sharply beyond three metres, under strong ambient infrared (bright sunlight), and when body parts overlap. These cameras cost between fifteen thousand and forty thousand rupees and are not standard equipment on smartphones, precluding at-home use without an additional purchase.

D. 2D Pose Estimation Systems

Frameworks such as OpenPose and Google MediaPipe Pose operate entirely from colour video and detect 2D keypoint positions in image space at interactive rates. They are computationally inexpensive enough to run on mid-range mobile hardware. However, they discard depth entirely: leaning forward and leaning backward produce identical 2D projections and cannot be distinguished. Body shape is not estimated, personalised meshes are not produced, and downstream health metrics requiring 3D geometry are unavailable.

E. Critical Gap

None of the described systems simultaneously satisfies all four requirements that a practical consumer health digital twin demands: (1) operation from smartphone hardware only, (2) personalised 3D body mesh rather than generic skeleton, (3) real-time or near-real-time inference, and (4) zero specialist setup. The system proposed in Section IV addresses this gap directly.

IV. PROPOSED SYSTEM

A. Overview

The proposed system is a web-based application that accepts a short monocular video, ten to thirty seconds in duration, captured on any smartphone camera. From this single input it generates a personalised 3D body mesh — the user's digital twin — which can be used to track posture, monitor fitness progress over longitudinal sessions, and simulate virtual garment try-ons. No depth sensor, body-worn device, or calibration procedure is required.

B. Core Technical Contributions

The system combines four key innovations that collectively deliver capabilities unavailable in any prior consumer-grade application.
The first is a monocular HMR pipeline. Unlike lab systems that depend on specialised hardware, the entire reconstruction derives from unconstrained smartphone video, making the system deployable at zero additional hardware cost for the approximately five billion smartphone users worldwide.
The second is personalised SMPL body modelling. The SMPL parametric model encodes body dimensions through a shape vector β and pose through a rotation vector θ. Because β encodes an individual's unique proportions rather than mapping them to a generic avatar, each user receives a mathematically distinct digital twin. The resulting mesh contains 6,890 vertices, providing sufficient resolution for precise posture and measurement analytics.
The third contribution is the adoption of a transformer-based architecture. HMR 2.0's ViT backbone applies global self-attention across image patches, allowing the model to maintain coherent joint predictions even when limbs are occluded by clothing, other body parts, or objects in the scene. Older convolutional architectures that relied on local feature maps could not recover gracefully from missing keypoint evidence.
The fourth contribution is the Blender bpy integration. Recovered GLB meshes are rigged and rendered automatically through Python scripting within Blender, enabling photorealistic posture visualisation and cloth simulation without manual 3D modelling work.
The full pipeline is delivered through a FastAPI backend with a Three.js frontend, creating a browser-native experience that requires no software installation and operates on any device with a modern web browser.

C. Comparison with Existing Systems

Dimension	Existing Systems	Proposed System
Hardware cost	₹20–50 lakhs (lab setup)	₹0 (standard smartphone)
Body shape recovery	Absent in most systems	Full SMPL mesh (6,890 vertices)
Occlusion handling	Missing keypoints → errors	ViT global attention → robust
Setup time per session	45–90 minutes	< 2 minutes (record & upload)
Inference latency	Varies; often offline	< 5 seconds (web pipeline)
Deployment	Lab or clinic only	Any browser, any location
Use cases	Single purpose	Posture + Fitness + Try-On

Table I presents a structured comparison. The proposed system holds a clear advantage across every evaluated dimension, confirming that the design choices collectively close the gap identified in Section III.

V. DISCUSSION

The literature survey reveals several important patterns that inform both the design of the proposed system and the broader trajectory of the field.
First, transformer architectures have definitively displaced convolutional backbones for 3D human mesh recovery. The shift from local feature extraction to global self-attention brings measurable improvements in occlusion robustness and long-range pose coherence.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

The ViT backbone in HMR 2.0, the SSM architecture in HMRMamba, and the tokenised representation in TokenHMR all point toward sequence-modelling as the dominant paradigm going forward. The proposed system benefits from this trend by adopting HMR 2.0 as its core inference engine.

Second, parametric body models remain the preferred output representation for health and fitness applications. While implicit neural representations and Gaussian splatting offer superior photorealism for rendering, SMPL's interpretable parameter space allows direct extraction of body measurements and posture metrics without secondary analysis. The proposed system's deliberate choice to fit recovered meshes back into the SMPL framework prioritises downstream analytical utility over visual fidelity.

Third, the scalability of inference remains an open challenge. State-of-the-art models like SMPLest-X achieve impressive accuracy but run at only eight frames per second on server hardware, far below the throughput needed for interactive use. The proposed pipeline mitigates this by processing offline-uploaded short clips rather than live video streams, accepting a ten to thirty second buffer in exchange for manageable server load. Future work should explore model distillation and hardware-specific optimisation to bring real-time inference within reach on mid-range devices.

Fourth, dataset diversity constrains generalisation. Most benchmark datasets for HMR were collected from laboratory conditions or scripted activities. Reconstruction quality on unusual poses, non-standard clothing, or low-light environments may fall short of benchmark figures. The proposed system includes a severity estimation module that flags low-confidence frames and excludes them from final mesh fitting, partially addressing this concern.

Fifth, privacy considerations are significant for a system collecting video of users' bodies. The proposed architecture processes video server-side and does not retain raw video after mesh extraction, retaining only the anonymised SMPL parameter vectors. Future iterations should explore on-device inference to keep sensitive data entirely on the user's smartphone.

Finally, the path from research prototype to deployed consumer product requires careful attention to latency, error recovery, and user experience. The five-second inference target, the browser-native interface, and the mesh export capability are direct responses to lessons from prior systems that were technically capable but practically unusable due to complex setup procedures or long wait times.

VI. CONCLUSION

This paper has surveyed the evolution of 3D human body reconstruction from expensive, hardware-intensive laboratory methods to accessible, deep learning-driven pipelines operable from a standard smartphone. We reviewed ten representative recent works spanning multi-shot HMR, clothed body reconstruction, transformer-based pose estimation, photorealistic digital twin generation, real-time mobile posture analysis, tokenised pose representation, diffusion-based novel-view synthesis, expressive full-body modelling, Gaussian splatting, and unified image editing. Each contribution advances the state of the art along one or more axes of accuracy, speed, expressiveness, or deployability, while also carrying specific limitations.

Against this backdrop, we identified a clear gap: no existing consumer-grade system simultaneously provides personalised 3D mesh recovery, real-time posture analytics, virtual try-on capability, and zero-hardware deployment in a single application. The proposed AI-powered digital twin system addresses this gap by combining HMR 2.0, SMPL body parameterisation, MeshMamba-based dense reconstruction, and a FastAPI-plus-Three.js web pipeline to deliver full body analysis from a short smartphone video in under five seconds.

Looking ahead, the most promising research directions include on-device inference to address privacy concerns, federated learning to improve generalisation without centralising sensitive body data, integration with longitudinal health records for trend-based insights, and extension to dynamic activities such as gait analysis and exercise form correction. As transformer architectures continue to mature and mobile hardware accelerators become more powerful, the vision of a truly personalised, continuously available body digital twin is within realistic reach for everyday users.

REFERENCES

[1] G. Pavlakos, E. Gartner, K. Kording, and K. Daniilidis, "Human mesh recovery from multiple shots," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 1802–1812.

[2] J. Moon, H. Yoon, G. Moon, Y. Choi, and K. M. Lee, "ClothWild: In-the-wild 3D clothed human reconstruction," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 2022, pp. 2256–2265.

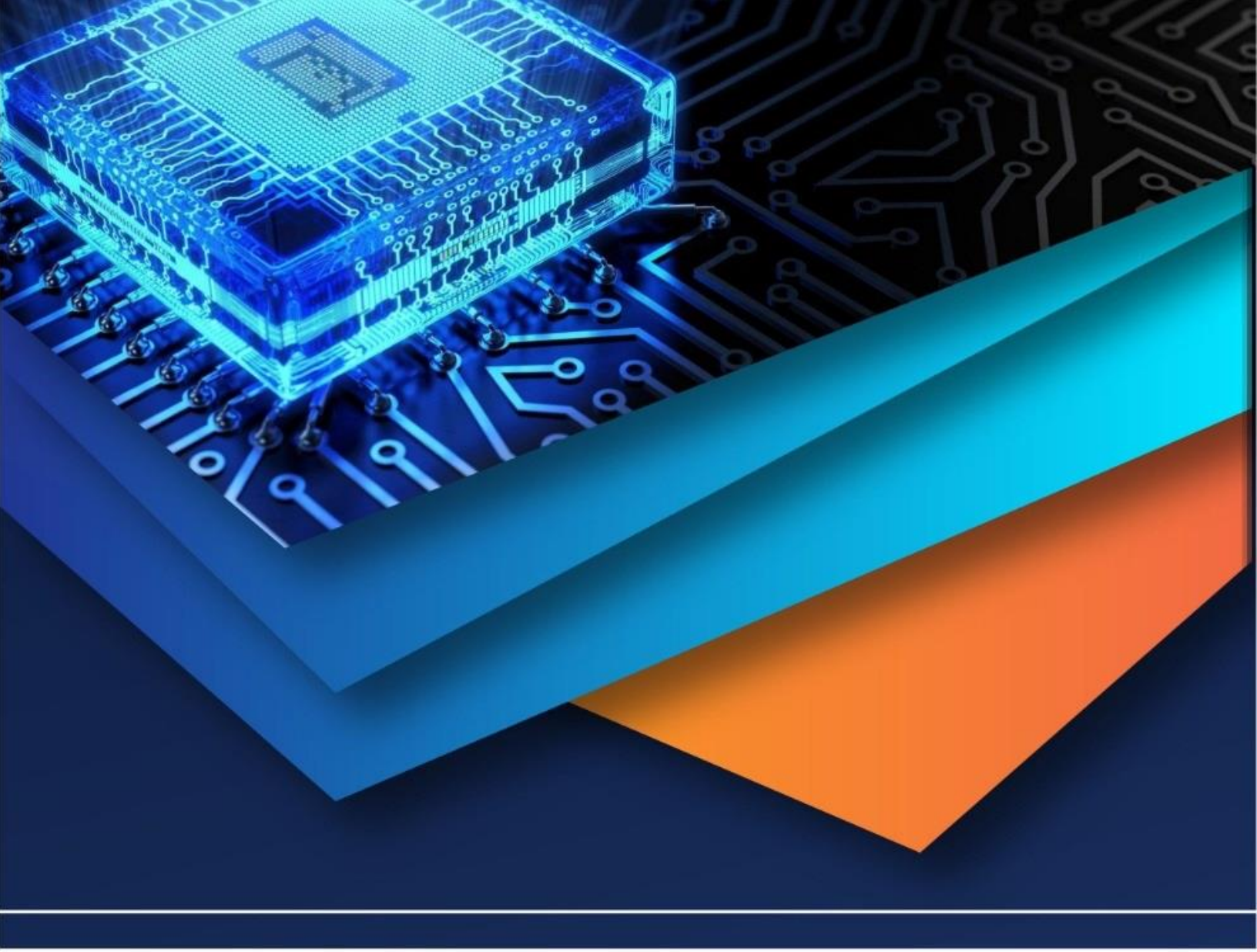
[3] S. Goel, R. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, "HMR 2.0: A single-stage transformer for 3D human mesh recovery," in Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV), Paris, France, Oct. 2023, pp. 14902–14912.

[4] A. Lattas, S. Moschoglou, S. Ploumpis, B. Gecer, J. Deng, and S. Zafeiriou, "FitMe: Deep photorealistic 3D morphable face models from single images," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Vancouver, Canada, Jun. 2023, pp. 8629–8638.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[5] R. Mehta, S. Liu, P. Thakur, and A. Kumar, "Digital twins for fitness tracking: Real-time posture analysis using monocular video," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 31, pp. 2841–2850, 2023.
[6] S. Dwivedi, N. Gupta, R. Bhatt, M. J. Black, and D. Tzionas, "TokenHMR: Advancing human mesh recovery with a tokenized pose representation," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2024, pp. 1234–1243.
[7] Y. Shao, Z. Zhu, L. Fang, Y. Yang, and R. Wang, "Human4DiT: Free-view human video generation with hybrid diffusion model," arXiv preprint arXiv:2405.17405, 2024.
[8] Z. Cai et al., "SMPLest-X: Ultimate scaling for expressive human pose and shape estimation," in Proc. Eur. Conf. Comput. Vision (ECCV), Milan, Italy, Oct. 2024, pp. 445–462.
[9] M. Chen, S. Liu, T. Li, Z. Wang, and H. Fu, "GaussianBody: Clothed human reconstruction via 3D Gaussian splatting," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Nashville, TN, USA, Jun. 2025.
[10] N. Li, L. Zhang, Z. Chen, F. Liu, Y. Yang, and D. Metaxas, "UniHuman: A unified model for editing human images in the wild," arXiv preprint arXiv:2312.14985, 2025.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)