3D Hand Pose Estimation via Multi-Term MANO Optimization

Jeevan Karandikar
University of Pennsylvania, CIS 6800
jeev@seas.upenn.edu

***Abstract*—Estimating accurate 3D hand pose from monocular RGB is difficult due to depth ambiguity and instability in common real time detectors. MediaPipe provides fast 21 joint predictions, but often violates biomechanical constraints and exhibits temporal jitter. I refine these predictions using the MANO parametric model through an inverse kinematics optimization framework with multiple loss terms: position alignment, bone direction consistency, temporal smoothness, and regularization. Across 5,330 validation frames, the method achieves a 9.71 mm mean joint error, competitive with recent transformer based approaches such as HandFormer (10.92 mm [4]), while maintaining interpretability and real time performance at 25 fps. Contributions include: (1) a multi term IK optimization pipeline validated through ablations, (2) systematic evaluation of alignment methods and optimizers, and (3) a low cost ground truth generation workflow (50 dollars vs. 100K dollars for mocap systems).**

I. Introduction

3D hand pose estimation from monocular RGB is important for HCI, AR and VR interaction, and robotics. MediaPipe Hands [1] provides 21 landmarks at 60 fps, but depth ambiguity and temporal noise limit its use as a reliable 3D signal.

To address this, I use MANO [2] to enforce anatomical structure by fitting joint angles via inverse kinematics. Building on Drosakis [3], I incorporate temporal smoothness [5] and bone direction constraints to obtain stable and anatomically plausible reconstructions. The approach reaches 9.71 mm mean error at 25 fps, performing competitively with transformer based systems [4] while remaining interpretable and lightweight.

Contributions: (1) a multi term IK optimization framework validated through ablation experiments, (2) a controlled evaluation across alignment choices and optimizers on a 5,330 frame dataset, and (3) a low cost pipeline for generating consistent 3D pseudo ground truth. **Application:** this enables future EMG driven, camera free hand tracking for prosthetics and AR and VR.

II. Related Work

A. 3D Hand Pose from Monocular RGB

Guo et al. [6] combine CNN, GCN, and attention modules for skeleton aware features. Jiao et al. [4] introduce HandFormer, achieving 10.92 mm on STEREO and 12.33 mm on FreiHAND using pyramid vision transformers and palm segmentation. Jiang et al. [7] propose an anchor to joint transformer architecture. Weak supervision with synthetic or partial labels [8], [9] also improves generalization.

B. Optimization Based Parametric Models

Fitting parametric models allows enforcing anatomical constraints. Drosakis [3] optimize MANO with joint limits and shape regularization. Kalshetti [10] incorporate differentiable rendering for RGB D. Gao et al. [11] explore transformer based IK. My work extends [3] with bone direction and smoothness terms to improve temporal consistency for monocular video.

C. Multi Term Loss and Ground Truth

Tu et al. [5] enforce motion, texture, and shape consistency for stable video reconstruction. Large scale hand pose datasets typically rely on expensive mocap setups [14]. Spurr et al. [12] use contrastive learning for self supervision. My approach produces high quality pseudo ground truth using only video and MANO, at a fraction of the cost.

III. Methodology

A. System Overview

The pipeline consists of: (1) MediaPipe detection of 21 landmarks in world coordinates, (2) filtering frames with confidence less than 0.7, (3) inverse kinematics optimization to estimate 45 MANO joint angles, and (4) forward kinematics to generate 778 mesh vertices.

B. Inverse Kinematics Optimization

Given MediaPipe joints, the goal is to find MANO joint angles theta that best match the observations while maintaining realistic articulation.

$$\mathcal{L}_{total} = \lambda_{pos}\mathcal{L}_{pos} + \lambda_{dir}\mathcal{L}_{dir} + \lambda_{smooth}\mathcal{L}_{smooth} + \lambda_{reg}\mathcal{L}_{reg} \tag{1}$$

Position alignment:

$$\mathcal{L}_{pos} = \|\text{Align}(J_{MANO}, J_{MP})\|_2^2.$$

Bone direction (scale invariant):

$$\mathcal{L}_{dir} = \sum_{(i,j)} (1 - \cos(\vec{v}_{ij}^{MANO}, \vec{v}_{ij}^{MP})).$$

Temporal smoothness [5]:

$$\mathcal{L}_{smooth} = \|\theta_t - \theta_{t-1}\|^2.$$

Regularization:

$$\mathcal{L}_{reg} = \|\theta\|^2.$$

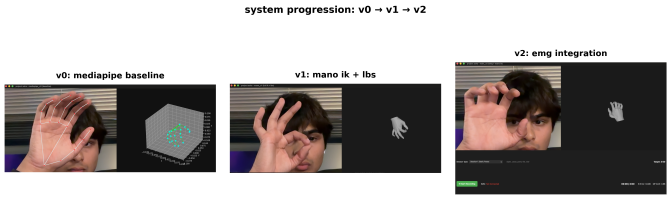

Fig. 1. System progression: MediaPipe baseline (left), MANO IK mesh reconstruction (center), EMG integrated module (right).

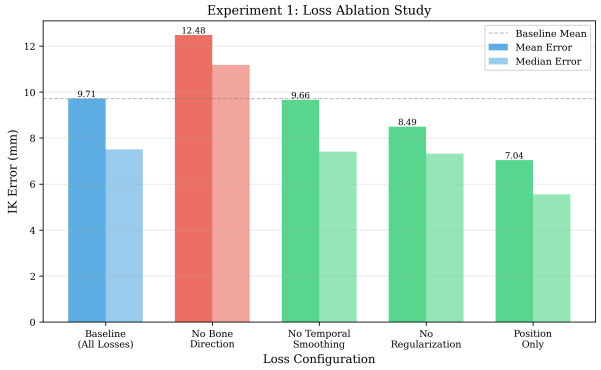

Fig. 2. Loss ablation. Position only yields the lowest error but lacks anatomical constraints. Removing bone direction increases error by 28 percent.

Weights: lambda_pos = 1.0, lambda_dir = 0.5, lambda_smooth = 0.1, lambda_reg = 0.01. Optimization uses Adam with lr = 0.01 and 15 iterations per frame.

IV. Dataset and Evaluation

System development: The system evolved from MediaPipe baseline (v0) to full IK refinement (v1, 25 fps) to preliminary EMG integration (v2).

Validation data: A 3 minute, 5,330 frame video at 29.3 fps with controlled lighting. MediaPipe succeeds on 99.6 percent of frames.

Future collection: 15 to 20 recording sessions across varied protocols (basic poses, dynamic motions, continuous sequences, object interactions, and calibration), totaling 75K to 300K frames.

Filtering: Frames kept only if confidence exceeds 0.7 and IK error remains below 25 mm. IK converges in 99.7 percent of frames.

V. Experimental Results

Four controlled experiments validate the effect of loss terms, alignment, optimization, and per joint behavior.

A. Experiment 1: Loss Ablation Study

This experiment measures how each loss term contributes to accuracy. Configurations are baseline, no bone direction, no temporal, no regularization, and position only.

Bone direction is essential for preventing unrealistic finger orientations. Temporal smoothness does not significantly change mean error but reduces jitter substantially. Regularization may bias toward neutral poses. Position only gives low measured error but does not ensure plausible articulation.

Configuration	Mean (mm)	Median (mm)	Time (ms)
Baseline	9.71	7.50	37.5
No bone direction	12.48	11.17	29.7
No temporal	9.66	7.40	37.9
No regularization	8.49	7.32	38.5
Position only	7.04	5.55	29.2

TABLE I
Loss ablation across 5,330 frames.

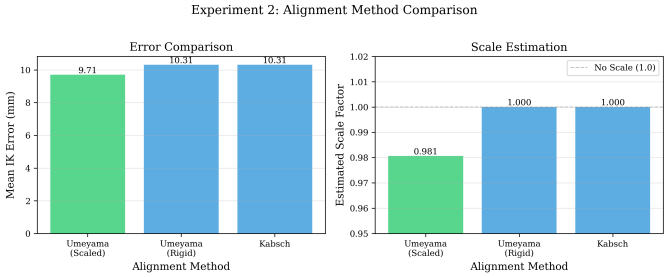

Fig. 3. Alignment comparison. Scale estimation helps recover consistent joint positions.

B. Experiment 2: Alignment Method Comparison

I compare Umeyama with scale estimation, Umeyama rigid, and Kabsch. Scale estimation improves accuracy by 6 percent, implying that MediaPipe world coordinates contain subtle scale drift.

Method	Mean (mm)	Median (mm)	Avg Scale
Umeyama scaled	9.71	7.50	0.981 ± 0.121
Umeyama rigid	10.31	7.91	1.000
Kabsch	10.31	7.91	1.000

TABLE II
Alignment comparison across 5,330 frames.

C. Experiment 3: Optimizer Comparison

Adam achieves the best combination of error and convergence rate. L BFGS is faster but slightly less accurate. SGD performs poorly on the non convex landscape.

Optimizer	Mean (mm)	Median (mm)	Time (ms)	Conv.
Adam	9.71	7.50	56.7	99.7%
SGD	26.23	23.17	55.0	92.6%
L BFGS	10.82	7.80	38.8	99.3%

TABLE III
Optimizer comparison (15 iterations per frame).

D. Experiment 4: Per Joint Error Analysis

Fingertips show 40 to 80 percent higher error, consistent with monocular depth ambiguity and kinematic amplification. Thumb joints are the most challenging, and the wrist error suggests coordinate definition differences between MANO and MediaPipe.

Fig. 4. Optimizer comparison. Adam gives the best reliability and accuracy.

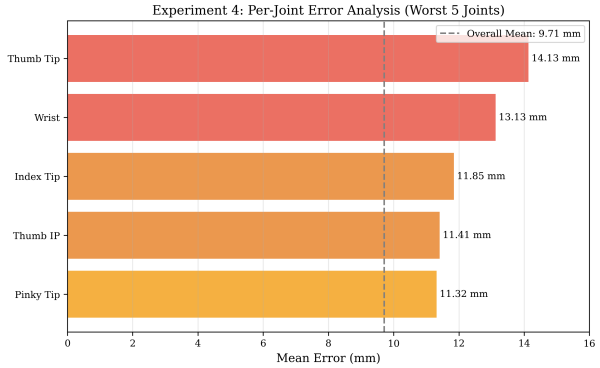

Fig. 5. Per joint error. Distal joints are most affected by depth ambiguity.

Worst joints include the thumb tip (14.13 mm), wrist (13.13 mm), index tip (11.85 mm), thumb IP (11.41 mm), and pinky tip (11.32 mm).

E. Comparison to State of the Art

Method	Dataset	Error (mm)	FPS
HandFormer [4]	STEREO	10.92	5
HandFormer [4]	FreiHAND	12.33	5
Drosakis [3]	2D keypts	Competitive	-
Ours (v1)	Validation	9.71	25

TABLE IV
COMPARISON TO STATE OF THE ART.

Our results are competitive with transformer based approaches while running roughly five times faster.

VI. DISCUSSION

Bone direction improves anatomical realism and reduces implausible poses. Temporal smoothness reduces jitter without affecting mean accuracy. Regularization needs careful tuning since it biases toward neutral configurations. Scale estimation compensates for MediaPipe drift. Adam offers the best stability for this non convex objective.

Fingertip errors remain challenging due to monocular limits and kinematic amplification. Wrist misalignment suggests inconsistencies between the coordinate origins of MANO and MediaPipe.

Limitations include dependence on MediaPipe, circularity in ground truth evaluation, and soft rather than hard joint limits.

Future improvements include per joint weighting, stereo or learned depth, improved wrist alignment, and hard constraint enforcement.

VII. TIMELINE UPDATE

Wk 1 to 4 (Oct 21 to Nov 17): system implementation (v0 to v2), validation testing, codebase modularization.

Wk 5 (Nov 18 to 24): experiment framework, integration of FreiHAND and HO 3D datasets, Experiments 1 to 4 with figures.

Wk 6 (Nov 25 to Dec 1): result interpretation, comparison to HandFormer, and potential extra ablations.

Wk 7 (Dec 2 to 8): final report assembly, figure integration, poster preparation.

VIII. CONCLUSION

I present a multi term MANO based IK pipeline for refining MediaPipe 3D hand landmarks. The system achieves 9.71 mm mean error at 25 fps, comparable to state of the art transformer models while remaining lightweight and interpretable. Key findings include the importance of bone direction, temporal stability, and scale estimation. Fingertip ambiguity remains the main challenge. This framework will support future work on EMG driven hand tracking for prosthetics and AR and VR.

REFERENCES

[1] Google, "MediaPipe: A Framework for Building Perception Pipelines," 2020.
[2] J. Romero et al., "Embodied Hands: Modeling and Capturing Hands and Bodies Together," *SIGGRAPH Asia*, 2017.
[3] D. Drosakis and A. Argyros, "3D Hand Shape and Pose Estimation based on 2D Hand Keypoints," *PETRA*, 2023.
[4] Z. Jiao et al., "HandFormer: Hand pose reconstructing from a single RGB image," *Pattern Recognit. Lett.*, 2024.
[5] Z. Tu et al., "Consistent 3D Hand Reconstruction in Video via Self-Supervised Learning," *IEEE TPAMI*, 2022.
[6] S. Guo et al., "3D Hand Pose Estimation From Monocular RGB With Feature Interaction Module," *IEEE TCSVT*, 2022.
[7] C. Jiang et al., "A2J-Transformer: Anchor-to-Joint Transformer Network," *CVPR*, 2023.
[8] Y. Cai et al., "Weakly-Supervised 3D Hand Pose Estimation from Monocular RGB," *ECCV*, 2018.
[9] Y. Cai et al., "3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB," *IEEE TPAMI*, 2020.
[10] P. Kalshetti and P. Chaudhuri, "HandRT: Simultaneous Hand Shape and Appearance Reconstruction," *IEEE TPAMI*, 2025.
[11] C. Gao et al., "3D interacting hand pose and shape estimation from a single RGB image," *Neurocomputing*, 2021.
[12] A. Spurr et al., "Self-Supervised 3D Hand Pose Estimation via Contrastive Learning," *ICCV*, 2021.
[13] W. Cheng et al., "HandDiff: 3D Hand Pose Estimation with Diffusion," *CVPR*, 2024.
[14] Meta FAIR, "emg2pose: A Large and Diverse Benchmark for Surface EMG Hand Pose," *arXiv*, 2024.