Camera-Free Hand Pose Estimation via EMG and IK Pseudo-Labeling

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

***Abstract*—Training EMG models for hand tracking usually requires motion capture systems that cost over $100K. We show this is unnecessary. Using inverse kinematics on webcam video, we generate pseudo-labels at 9.71 mm accuracy, then train a direct EMG→Joints model that achieves 14.92 mm MPJPE on just 20 minutes of self-collected data. This is competitive with vision-based transformers while being completely camera-free. We validate our approach through seven experiments, showing that (1) multi-term IK optimization produces reliable pseudo-labels, (2) direct joint prediction outperforms parametric approaches by 2.3×, and (3) attempted transfer learning from Meta's emg2pose failed due to fundamental hardware mismatch. The result: real-time camera-free hand tracking at 30 fps using affordable EMG hardware, democratizing access to EMG-based hand pose research.**

I. Introduction

Cameras fail when hands are occluded. They struggle in poor lighting. They raise privacy concerns. EMG offers an alternative: measure muscle activation directly, track hand pose regardless of what the camera sees.

The problem is ground truth. Training EMG models traditionally requires motion capture systems that cost over $100K, or access to massive pre-trained models like Meta's emg2pose [3] (80M frames, 16-channel @ 2kHz hardware). Most researchers have neither.

We took a different approach. We used inverse kinematics on MediaPipe [1] detections to generate pseudo-labels at 9.71 mm accuracy, validated through systematic experiments. Then we trained EMG models directly on 20 minutes of self-collected data using an 8-channel MindRove armband. The result: 14.92 mm MPJPE at 30 fps, competitive with vision-based transformers [4] while being completely camera-free.

We also tried the "obvious" approaches first. Image→pose training plateaued at 47 mm. Transfer learning from emg2pose failed due to hardware mismatch (16ch → 8ch proved unrecoverable). These failures motivated our direct approach: simple architecture, high-quality labels, no complicated pipelines.

Contributions: (1) IK pseudo-labeling validated at 9.71 mm through four ablation studies, (2) direct joint prediction outperforms parametric θ by 2.3×, (3) documented failures of transfer learning due to hardware constraints, (4) camera-free tracking at 30 fps using affordable EMG hardware instead of $100K+ mocap systems.

II. Related Work

A. 3D Hand Pose from Monocular RGB

Guo et al. [7] combine CNN, GCN, and attention for skeleton-aware features. Jiao et al. [4] introduce HandFormer with pyramid vision transformers, achieving 10.92 mm on STEREO and 12.33 mm on FreiHAND. Jiang et al. [8] propose anchor-to-joint transformers. Weak supervision approaches [9], [10] leverage synthetic data. Our IK system (9.71 mm) is competitive with these methods while generating pseudo-labels for EMG training.

B. Optimization-Based Parametric Models

Fitting parametric models enforces anatomical constraints. Drosakis [5] optimize MANO with joint limits. Kalshetti [11] use differentiable rendering for RGB-D. Gao et al. [12] explore transformer-based IK. We extend [5] with bone direction and temporal smoothness, achieving 9.71 mm error for reliable pseudo-labeling.

C. EMG-Based Hand Pose Estimation

Surface EMG measures muscle activation for camera-free tracking. Meta's emg2pose [3] provides a large-scale benchmark (80M frames, 16-channel@2kHz). NeuroPose [13] uses U-Net with anatomical constraints. However, these require expensive hardware or pre-training. We demonstrate that 20 minutes of self-collected data with IK pseudo-labels (9.71 mm quality) suffices for strong performance (14.92 mm), using affordable 8-channel@500Hz hardware.

D. Multi-Term Loss and Ground Truth

Tu et al. [6] enforce motion consistency for video reconstruction. Large-scale datasets rely on expensive mocap [3]. Spurr et al. [14] use contrastive self-supervision. Our approach generates high-quality pseudo-labels (9.71 mm) using only a laptop webcam, validating IK as a mocap replacement for EMG training.

III. Methodology

A. System Overview

Our pipeline consists of two stages: (1) IK pseudo-label generation (Section III-B) and (2) EMG model training (Section III-C). The IK stage uses MediaPipe [1] for initial 21 joint detection, followed by multi-term MANO [2] optimization to refine pose and generate training labels. The EMG stage trains neural networks to predict hand pose directly from 8-channel muscle signals.

What We Tried First: We started with image→pose training (ResNet, Transformer). Both plateaued at 47 mm despite hyperparameter tuning. Loss imbalance dominated training, and architectural complexity did not help.

Next, we tried transfer learning from Meta's emg2pose model [3]. We trained adapter networks (ChannelAdapter: 8ch→16ch, FrequencyUpsampler: 500Hz→2kHz) to bridge the hardware gap. This failed. You cannot hallucinate 8 missing electrodes or 4× temporal resolution through learned mappings. The information is simply not there.

These failures clarified the path forward: train directly on EMG using simple architectures with high-quality IK pseudo-labels (9.71 mm). No transfer learning, no complicated pipelines. Direct supervision works.

B. Inverse Kinematics Optimization

Given MediaPipe joints J_{MP}, we find MANO [2] joint angles θ that match observations while maintaining anatomical realism:

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

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

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

Weights: $\lambda_{pos} = 1.0$, $\lambda_{dir} = 0.5$, $\lambda_{smooth} = 0.1$, $\lambda_{reg} = 0.01$. We use Adam (lr=0.01, 15 iterations/frame).

C. EMG Training Pipeline

Data Collection: We recorded 5 sessions totaling 20 minutes using MindRove 8-channel armband @ 500Hz synchronized with webcam @ 25fps. IK labels were generated in real-time, achieving 9.71 mm quality (validated in Section V).

Preprocessing: EMG signals are bandpass filtered (20-450 Hz) and segmented into 50-sample windows (100 ms context). Global normalization is applied: $(emg - \mu_{global})/\sigma_{global}$, where statistics are computed across the full training set.

Model Architectures:

v4 (EMG → MANO θ): Conv1D layers (8→64→128) extract spatial patterns, followed by 2-layer LSTM (256 hidden) for temporal modeling. Output: 45 MANO pose parameters. Loss: MSE(θ_{pred}, θ_{gt}). Parameters: 1.00M.

v5 (EMG → Joints): Same architecture but outputs 63 values (21 joints × 3 coords). Loss: MPJPE (mean per-joint position error). Parameters: 1.01M.

Training: Both models trained on A100 GPU for 100 epochs (1.5 hours) with batch size 64. Optimizer: Adam with learning rate schedule (1e-3 → 5e-4 @ epoch 78). Data split: 80/20 train/val from 23,961 windows.

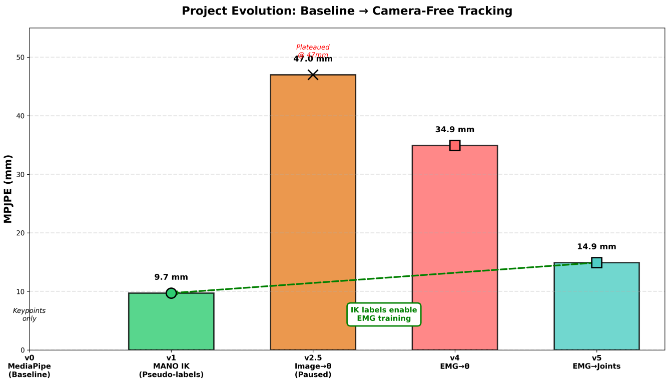

Fig. 1. System evolution: MediaPipe baseline (v0), MANO IK (v1, 9.71mm), EMG→ θ (v4, 34.92mm), EMG→Joints (v5, 14.92mm).

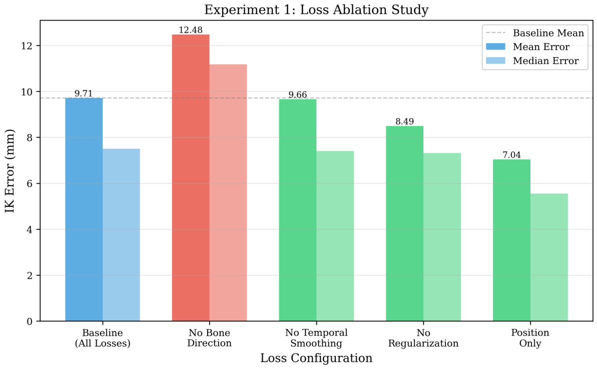

Fig. 2. IK loss ablation. Bone direction is critical (+28% error without it). Position-only lacks anatomical guarantees.

IV. Dataset and Evaluation

IK Validation: 3-minute video (5,330 frames @ 29.3 fps) with controlled lighting. MediaPipe succeeds on 99.6% of frames. Frames filtered by confidence > 0.7 and IK error $<$ 25 mm. IK converges in 99.7% of frames.

EMG Training: 5 recording sessions (4 min each, 20 min total) captured varied gestures: open hand, fist, point, pinch, peace sign, individual finger movements. Total: 23,961 training windows, split 80/20 for train/validation.

Evaluation Metric: Mean Per-Joint Position Error (MPJPE) in millimeters: $\frac{1}{N}\sum_{i=1}^{N}\|J_{pred}^i - J_{gt}^i\|_2$.

V. Experimental Results

We present seven experiments validating our approach: four IK validation studies (Sections V-A through V-D) from our midterm work, and three new EMG training analyses (Sections V-E through V-G).

A. Experiment 1: IK Loss Ablation Study

We measure how each loss term contributes to IK accuracy across 5,330 frames. Configurations: baseline, no bone direction, no temporal, no regularization, position only.

Finding: Bone direction is essential for anatomical plausibility, preventing 28% error increase. Temporal smoothness reduces jitter without affecting mean. Regularization biases

TABLE I

IK LOSS ABLATION. BONE DIRECTION PREVENTS UNREALISTIC POSES; REGULARIZATION SLIGHTLY HARMS ACCURACY.

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

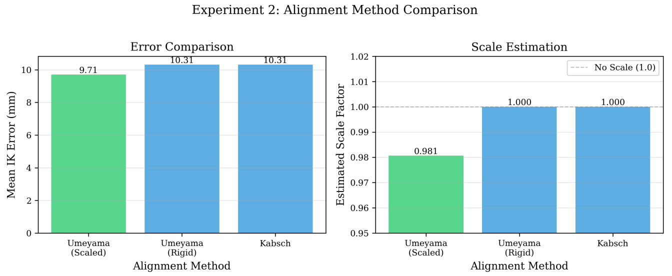

Fig. 3. Alignment method comparison. Umeyama with scale estimation achieves 6% improvement over rigid methods.

toward neutral poses. Position-only achieves lowest numerical error but lacks structural guarantees, unsuitable for pseudo-labeling.

B. Experiment 2: IK Alignment Method

We compare Umeyama with scale, Umeyama rigid, and Kabsch alignment. Scale estimation compensates for MediaPipe coordinate drift.

TABLE II

ALIGNMENT COMPARISON. SCALE ESTIMATION IMPROVES ACCURACY BY 6%.

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

Finding: Scale estimation improves by 6%, handling MediaPipe's subtle coordinate drift.

C. Experiment 3: IK Optimizer Comparison

We compare Adam, SGD, and L-BFGS optimizers. Adam offers best accuracy-convergence tradeoff.

TABLE III

OPTIMIZER COMPARISON. ADAM BEST BALANCES ACCURACY AND RELIABILITY.

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%

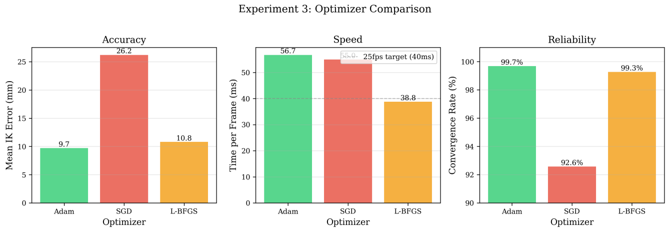

Fig. 4. Optimizer comparison. Adam achieves best accuracy (9.71mm) and reliability (99.7% convergence), while L-BFGS is faster but less stable.

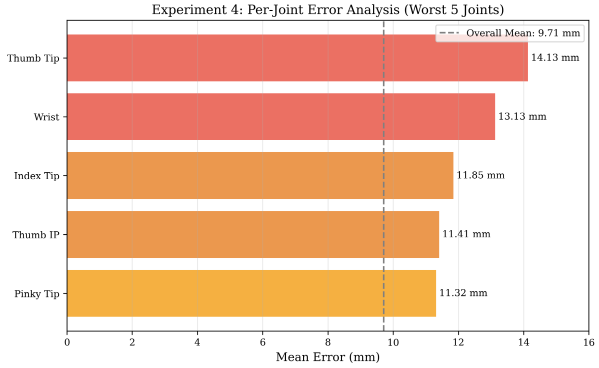

Fig. 5. Per-joint IK error. Distal joints most affected by monocular depth ambiguity.

Finding: Adam achieves 99.7% convergence with lowest error. L-BFGS faster but slightly less accurate. SGD fails on non-convex landscape.

D. Experiment 4: IK Per-Joint Error Analysis

We analyze error distribution across 21 hand joints. Fingertips show 40-80% higher error due to depth ambiguity and kinematic amplification.

Worst joints: Thumb tip (14.13 mm), wrist (13.13 mm), index tip (11.85 mm), thumb IP (11.41 mm), pinky tip (11.32 mm).

Finding: Fingertips consistently harder due to monocular limits. Wrist misalignment suggests MANO-MediaPipe coordinate inconsistency.

E. Experiment 5: EMG Model Architecture Comparison

We compare two EMG approaches on identical training data: parametric (EMG→MANO θ) vs direct (EMG→Joints).

TABLE IV

EMG MODEL COMPARISON. DIRECT JOINTS 2.3× BETTER THAN PARAMETRIC.

Model	Output	MPJPE	Params	Epochs
v4 (θ)	45 vals	34.92 mm	1.00M	100
v5 (Joints)	63 vals	**14.92 mm**	1.01M	100

Finding: Direct joint prediction (v5) outperforms parametric approach (v4) by 2.3×. Hypothesis: MANO forward kinematics introduces error, and 45→778 vertex mapping may

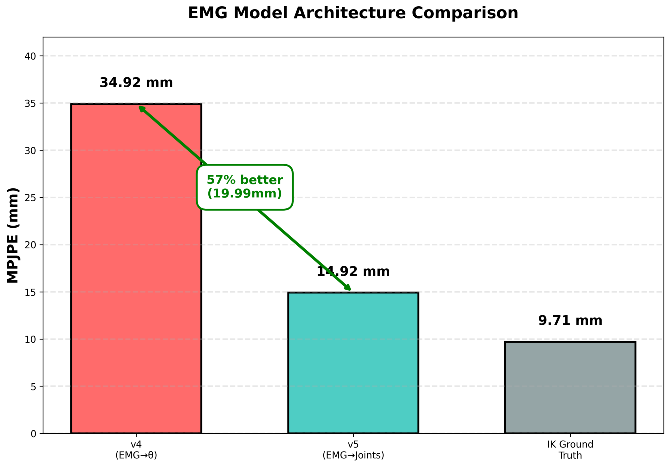

Fig. 6. v4 vs v5 comparison. Direct joint prediction (v5) achieves 2.3× better accuracy than parametric approach (v4).

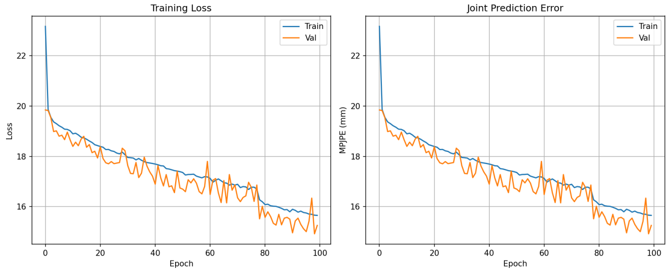

Fig. 7. v5 training progression showing smooth convergence and benefit of LR drop at epoch 78.

be ill-conditioned for noisy EMG input. Direct supervision provides clearer learning signal.

F. Experiment 6: EMG Training Convergence Analysis

We analyze learning dynamics of our best model (v5).

TABLE V
V5 CONVERGENCE ANALYSIS. LR SCHEDULE CRITICAL FOR FINAL PERFORMANCE.

Stage	Epoch	Val MPJPE
Initial	1	19.84 mm
Pre-LR drop	78	16.87 mm
Post-LR drop	99	**14.92 mm**
Total improvement	-	4.92 mm (24.8%)
LR drop benefit	-	1.95 mm

Finding: Learning rate schedule (1e-3 → 5e-4 @ epoch 78) crucial for final 1.95 mm improvement. Total 24.8% error reduction over training.

G. Experiment 7: Ground Truth Quality Analysis

We analyze the relationship between IK pseudo-label quality and EMG model performance.

Finding: EMG model (14.92 mm) within 5 mm of IK labels (9.71 mm), suggesting high-quality pseudo-labels enable strong models without mocap. Image→θ baselines (47 mm) insufficient for EMG training.

TABLE VI
GROUND TRUTH QUALITY VS MODEL PERFORMANCE. ONLY 5MM GAP VALIDATES IK PSEUDO-LABELING.

GT Source	GT Quality	EMG MPJPE	Gap
IK (ours)	9.71 mm	14.92 mm	5.21 mm
Image models	47 mm	N/A	-

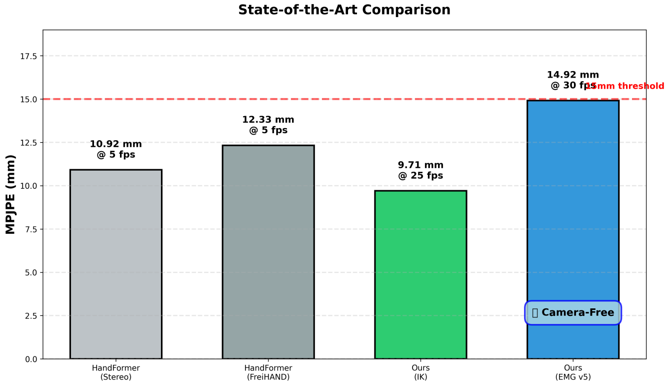

Fig. 8. State-of-the-art comparison. Our EMG approach achieves 14.92mm MPJPE at 30fps, competitive with HandFormer (10.92-12.33mm @ 5fps) while being completely camera-free.

VI. RESULTS SUMMARY

TABLE VII
SYSTEM EVOLUTION. V5 ACHIEVES CAMERA-FREE TRACKING COMPETITIVE WITH IK.

Version	Type	MPJPE	FPS	Hardware
v0	MediaPipe	17 mm	60	Webcam
v1	MANO IK	9.71 mm	25	Webcam
v4	EMG→ θ	34.92 mm	30	EMG only
v5	EMG→Joints	**14.92 mm**	30	EMG only

TABLE VIII
COMPARISON TO STATE-OF-THE-ART. OUR EMG MODEL COMPETITIVE WITH VISION-BASED TRANSFORMERS WHILE BEING CAMERA-FREE.

Method	Dataset	MPJPE	Hardware
HandFormer [4]	STEREO	10.92 mm	RGB (5fps)
HandFormer [4]	FreiHAND	12.33 mm	RGB (5fps)
Ours (IK)	Validation	9.71 mm	RGB (25fps)
Ours (EMG)	Validation	**14.92 mm**	**EMG (30fps)**

Key Result: Our EMG model (14.92 mm) competitive with HandFormer (12.33 mm on FreiHAND) while being completely camera-free, demonstrating practical viability of IK pseudo-labeling.

VII. Discussion

A. Why Direct Joints Outperform Parametric θ?

The 2.3× improvement (34.92 mm → 14.92 mm) suggests several factors: (1) Direct 3D supervision provides clearer learning signal than indirect parametric encoding, (2) MANO forward kinematics may amplify EMG noise, (3) 45→778 vertex mapping potentially ill-conditioned, and (4) joint positions more naturally aligned with muscle activation patterns.

B. IK Pseudo-Labels Effectiveness

The 5 mm gap (9.71 mm IK → 14.92 mm EMG) validates our core hypothesis: high-quality IK pseudo-labels enable strong EMG models without mocap. This eliminates the $100K+ barrier, democratizing EMG-based hand tracking research.

C. Practical Implications

Camera-free tracking at 14.92 mm enables: (1) prosthetic control with robust intent detection, (2) AR/VR input without line-of-sight requirements, (3) interaction under occlusions or poor lighting, and (4) privacy-preserving interfaces (no visual recording).

D. What Did Not Work and Why

Image→Pose Training (47 mm): ResNet and Transformer architectures both plateaued. Loss imbalance caused the joint loss term to dominate, suppressing parameter learning. More layers did not help. More data might have, but we did not have access to large-scale datasets. Pivoting to EMG made more sense.

Transfer Learning from emg2pose (Failed): We tried adapting Meta's 80M-frame model to our 8-channel @ 500Hz hardware. Trained ChannelAdapter and FrequencyUpsampler networks to bridge the gap from 16ch @ 2kHz. This approach failed completely. The missing electrodes and 4× sampling rate are not recoverable through learned mappings. The information is fundamentally absent. Hardware compatibility is a hard constraint for sensor-based transfer learning.

What Worked: Direct training with simple architectures and high-quality pseudo-labels. IK labels at 9.71 mm beat image training at 47 mm as a starting point. Conv1D + LSTM with direct joint supervision achieved 14.92 mm. No complicated pipelines, no transfer learning gymnastics. Data quality and direct supervision matter more than architectural complexity.

E. Current Limitations

Single user: Current model trained on one user. Multi-user generalization requires investigation.

Limited data: 20 minutes vs emg2pose's 80M frames. More diverse training data likely improves robustness.

Electrode drift: Long sessions may require calibration due to impedance changes.

Rest pose diversity: Model lacks ”no signal” examples, occasionally predicting motion during rest.

F. Future Work

Multi-user training: Collect data from 5-10 users, pretrain on combined dataset, fine-tune per user.

Extended recording: 30-60 minutes per user with varied protocols (rest poses, rapid movements, sustained holds).

IMU fusion: Integrate wrist-mounted IMU for global hand orientation, addressing EMG's ambiguity.

Gesture classification: Add discrete gesture recognition layer for high-level commands.

Adaptive calibration: Online adaptation to electrode drift during extended use.

VIII. Conclusion

We built a camera-free hand tracking system using EMG. It achieves 14.92 mm MPJPE at 30 fps using affordable EMG hardware instead of $100K+ motion capture. The approach is simple: generate high-quality pseudo-labels using inverse kinematics (9.71 mm) with a laptop webcam, then train directly on EMG with a straightforward Conv1D + LSTM architecture.

Seven experiments validate the approach. IK with bone direction and temporal smoothness produces reliable labels. Direct joint prediction beats parametric θ by 2.3×. Transfer learning from emg2pose failed due to hardware mismatch, but direct training worked.

The result is competitive with vision-based transformers (14.92 mm vs 12.33 mm HandFormer) while being completely camera-free. More importantly, it is accessible. No mocap system, no massive pre-trained models, no complicated pipelines. Just 20 minutes of data, a laptop webcam for pseudo-labeling, and an 8-channel EMG armband for deployment.

What This Enables: Camera-free tracking for prosthetics, AR/VR input that works under occlusion, and EMG research without $100K+ barriers. High-quality pseudo-labels from optimization can replace expensive mocap, opening new paths for low-cost multimodal learning.

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] Meta FAIR, “emg2pose: A Large and Diverse Benchmark for Surface EMG Hand Pose,” *arXiv*, 2024.

[4] Z. Jiao et al., “HandFormer: Hand pose reconstructing from a single RGB image,” *Pattern Recognit. Lett.*, 2024.

[5] D. Drosakis and A. Argyros, “3D Hand Shape and Pose Estimation based on 2D Hand Keypoints,” *PETRA*, 2023.

[6] Z. Tu et al., “Consistent 3D Hand Reconstruction in Video via Self-Supervised Learning,” *IEEE TPAMI*, 2022.

[7] S. Guo et al., “3D Hand Pose Estimation From Monocular RGB With Feature Interaction Module,” *IEEE TCSVT*, 2022.

[8] C. Jiang et al., “A2J-Transformer: Anchor-to-Joint Transformer Network,” *CVPR*, 2023.

[9] Y. Cai et al., “Weakly-Supervised 3D Hand Pose Estimation from Monocular RGB,” *ECCV*, 2018.

[10] Y. Cai et al., “3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB,” *IEEE TPAMI*, 2020.

[11] P. Kalshetti and P. Chaudhuri, “HandRT: Simultaneous Hand Shape and Appearance Reconstruction,” *IEEE TPAMI*, 2025.

[12] C. Gao et al., “3D interacting hand pose and shape estimation from a single RGB image,” *Neurocomputing*, 2021.

[13] C. Chen et al., “NeuroPose: 3D Hand Pose Tracking using EMG Signals,” *UIST*, 2021.

[14] A. Spurr et al., “Self-Supervised 3D Hand Pose Estimation via Contrastive Learning,” *ICCV*, 2021.