ECCV 2026 | MALMÖ | SEPT 8–12

TransHands: Repurposing Human Pose Encoders as Hand Pose Encoders

Milo Piccioli[1] Gianluca Amprimo[1] Claudia Ferraris[2] Gabriella Olmo[1]

[1]Politecnico di Torino, Torino, Italy [2]Consiglio Nazionale delle Ricerche, Torino, Italy

Politecnico di Torino — Department of Control and Computer Engineering

Consiglio Nazionale delle Ricerche

SMILIES

Paper

Code

Contributions & Takeaway

First systematic study
To the best of our knowledge, the first systematic study on transferring body motion priors to hand pose estimation across multiple SOTA architectures.

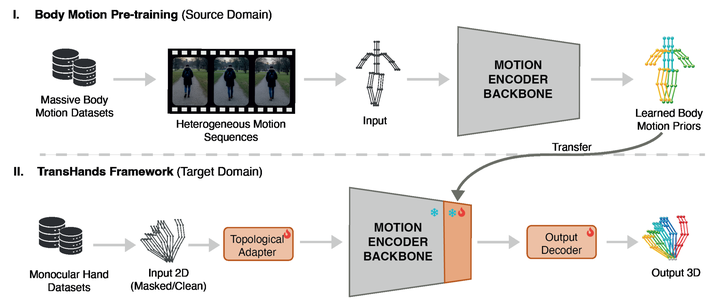
Backbone-agnostic
A modular framework that decouples topological adaptation from temporal modelling, so future motion encoders drop straight in.

Strong inductive biases
Biomechanical knowledge learned from body datasets provides strong inductive biases for modelling hand kinematics, opening new directions.

Takeaway
Human pose encoders can be effectively repurposed for hands. A lightweight adaptation bridges the gap, enabling highly data-efficient 3D hand pose estimation.

Limitations
Evaluation relies mostly on ground-truth 2D keypoints. Real-world accuracy remains bounded by the 2D detector.

1 Framework



(I) *Body Motion Pre-training* captures spatio-temporal priors from heterogeneous body datasets. **(II)** These priors are repurposed for hand pose estimation using a learnable *Topological Adapter* that aligns monocular hand sequences with the pre-trained body manifold.

2 Motivation and method

The problem. Existing hand-specific datasets are orders of magnitude smaller than their body pose counterparts, limiting the capacity of data-driven models to learn robust kinematic priors. Hands exhibit complex biomechanical constraints and high-frequency articulations that are difficult to capture and annotate at scale.

Can motion priors learned from large-scale human body datasets be transferred to model hand biomechanical knowledge?

The architecture. A frozen body encoder $\mathcal{E}_{\text{body}}$ encapsulated within learnable adaptation layers:

$$\hat{Y} = \mathcal{D}_{\text{out}}(\mathcal{P}(\mathcal{E}_{\text{body}}(\mathcal{A}_{\text{in}}(X))))$$

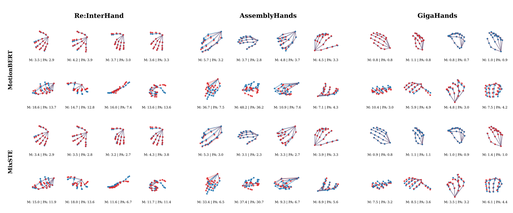
- **Topological Adapter** $\mathcal{A}_{\text{in}}$: A continuous **Neural ODE** that fluidly projects the hand topology onto the body's latent kinematic manifold, bridging the domain gap without structural redesign.
- **Motion encoder backbone** $\mathcal{E}_{\text{body}}$: Pre-trained on massive body datasets. **Partial unfreezing** ensures we retain robust body priors while selectively adapting deeper layers to hand micro-dynamics.
- **Latent space alignment** $\mathcal{P}$: **Multi-scale retention** layers standardise heterogeneous encoder outputs to a common, stable latent representation.
- **Task-specific decoder** $\mathcal{D}_{\text{out}}$: A deep MLP modelling the non-linear inverse kinematics to recover complex 3D finger articulations from the compressed features.

The training. Three stages. **S1 (Geometric Alignment):** the backbone is frozen while the adapter, projection head and decoder are optimised on Re:InterHand, with an additional self-supervised masked-motion objective. **S2-A (In-Domain Refinement):** last blocks unfrozen at a differential learning rate. **S2-B (Multi-Dataset Generalisation):** S1 specialised on AssemblyHands and GigaHands with balanced sampling.

3 Qualitative results

Prediction vs. ground truth

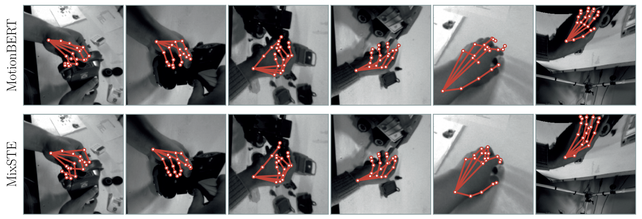
In-domain and cross-domain, from 2D ground-truth keypoints.



Prediction (red) vs. ground truth (blue). **S2-A** for Re:InterHand, **S2-B** for AssemblyHands and GigaHands. Top row per backbone: representative; bottom row: hard cases including failures on occluded frames.

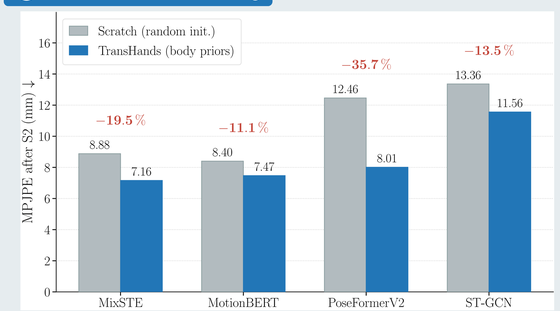
Zero-shot from an off-the-shelf detector

AssemblyHands, no target-domain supervision.

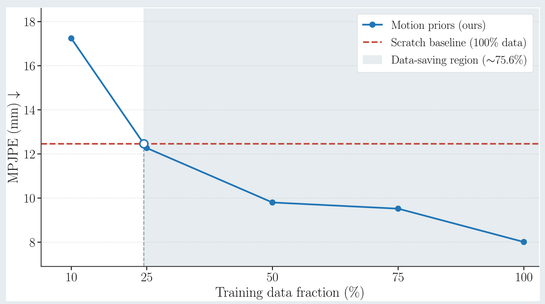


Predicted pose (red) over the frame, placed by the rotation and scale best fitting the detected 2D. No ground truth, no metric depth.

4 The value of transfer learning



Identical protocol and data, only the initialization differs. Red figures show the error reduction at S2 versus training from scratch. Our transfer method achieves a superior final result on **all four backbones**.



With PoseFormerV2, TransHands matches the scratch baseline trained on the *full* dataset (12.46 mm) using **less than 25% of the annotations**.

5 Accuracy across domains

In-domain results on Re:InterHand

Total is the backbone size. **Tr-S1** and **Tr-S2** are the trainable parameters per stage. **Gain** is the S1 → S2-A MPJPE improvement.

Backbone	Params (M)			MPJPE ↓		PA-MPJPE ↓		Gain (%)
	Total	Tr-S1	Tr-S2	S1	S2-A	S1	S2-A	
MotionBERT	65.70	2.24	15.00	9.51	7.47	6.81	5.62	+21.45
MixSTE	36.02	2.24	10.79	**9.09**	**7.16**	**6.69**	**5.41**	+21.34
PoseFormerV2	16.62	2.24	5.89	12.62	8.01	7.88	5.88	**+36.53**
ST-GCN	6.61	4.10	5.00	12.95	11.56	8.32	7.66	+10.73

Transformer-based architectures significantly outperform the graph convolutional network in the final refinement stage. **MixSTE** achieves the lowest error (**7.16 mm**), while **PoseFormerV2** shows the highest relative adaptation gain (**+36.53%**). **ST-GCN** appears to hit a capacity ceiling.

Multi-dataset generalisation

Zero-shot frozen backbone (**S1**) vs. multi-dataset refinement (**S2-B**), in mm.

	AssemblyHands				GigaHands			
	MPJPE ↓		PA-MPJPE ↓		MPJPE ↓		PA-MPJPE ↓	
Backbone	S1	S2-B	S1	S2-B	S1	S2-B	S1	S2-B
MotionBERT	42.87	11.73	22.06	6.96	139.00	**2.21**	29.80	**1.89**
MixSTE	**38.58**	**10.92**	**21.71**	**6.47**	121.20	2.36	28.27	2.03
PoseFormerV2	39.70	13.05	21.93	7.34	**115.90**	2.82	27.68	2.37
ST-GCN	54.80	20.66	24.15	11.34	137.02	6.12	**26.99**	4.38

On GigaHands, a high zero-shot MPJPE ($>$ 115 mm) alongside a stable PA-MPJPE (27–30 mm) indicates **global rigid misalignment** rather than structural distortion. Multi-dataset refinement resolves this gap, yielding very low errors aided by the dataset's smoother dynamics.

6 Robustness and generalisation in real-world contexts

RGB → 3D, zero-shot on AssemblyHands

Lift estimates 2D keypoints with an off-the-shelf detector then uplifts them. *RGB* estimates 3D directly.

Method	Type	MPJPE ↓	PA-MPJPE ↓
ArcticNet-SF	RGB	110.76[†]	–
V-HPOT	RGB	**92.09**[‡]	–
SMPLer-X	RGB	98.20	**19.40**
MediaPipe 3D	Lift	130.77	29.38
Ours (MotionBERT)	Lift	92.97	23.82
Ours (MixSTE)	Lift	93.44	24.28

[†]As reported in Prakash et al. [‡]As reported in Mucha et al.

On identical 2D inputs, we recover ∼38 mm over MediaPipe's GHUM uplifting. TransHands remains competitive with V-HPOT (root-aligned), showing robust performance across different regimes.

Downstream gesture recognition

Frozen S2-B representations evaluated on Jester and EgoGesture.

Backbone	Jester		EgoGesture	
	Top-1 (%) ↑	Top-5 (%) ↑	Top-1 (%) ↑	Top-5 (%) ↑
MotionBERT	88.44	**95.80**	**88.29**	97.36
MixSTE	**88.72**	95.76	88.08	**97.53**

Reaching ∼88% on both datasets with a *frozen* lifting encoder confirms that the learned motion priors encode high-level semantic dynamics.