ThermoHands: A Benchmark for 3D Hand Pose Estimation from Egocentric Thermal Images

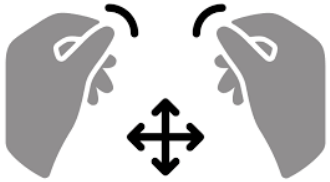
ACM SenSys 2025 Presentation (Irvine, USA)

Fangqiang Ding[1], Yunzhou Zhu[2], Xiangyu Wen[1], Gaowen Liu[3], Chris Xiaoxuan Lu[4]
[1]University of Edinburgh, [2]Georgia Institute of Technology, [3]Cisco Research,
[4]AI Centre, Department of Computer Science, UCL

Egocentric Hand Pose Estimation for Physical AI

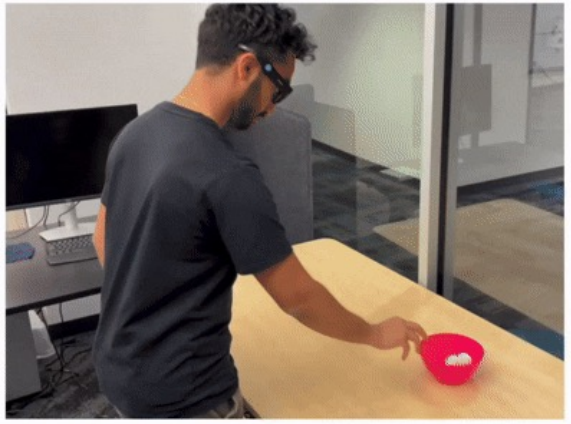
- VR/AR: controller-free spatial interaction (e.g., Meta Quest, Apple Vision Pro)

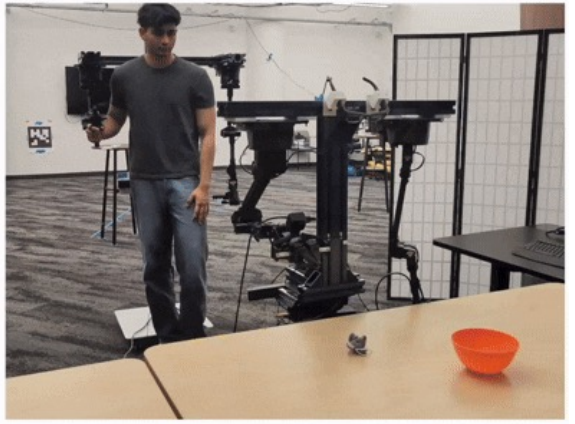
Controller

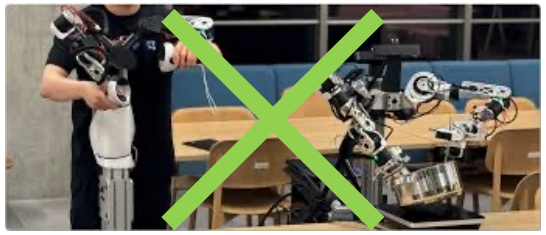


Hands-as-UI

- Robot manipulation: Imitation learning from human demonstration via egocentric video (e.g., MAPLE[1], EgoMimic[2])







Expensive teleoperation

Scalability

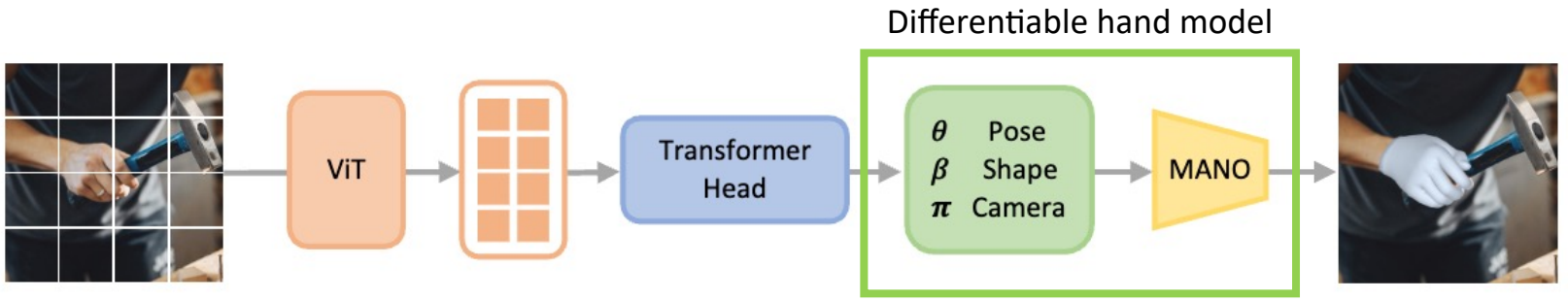
THE UNIVERSITY of EDINBURGH

[1] Gavryushin et al., MAPLE: Encoding Dexterous Robotic Manipulation Priors Learned From Egocentric Videos (arxiv, 2025)
[2] Kareer et al., EgoMimic: Scaling Imitation Learning via Egocentric Video (arxiv, 2024)

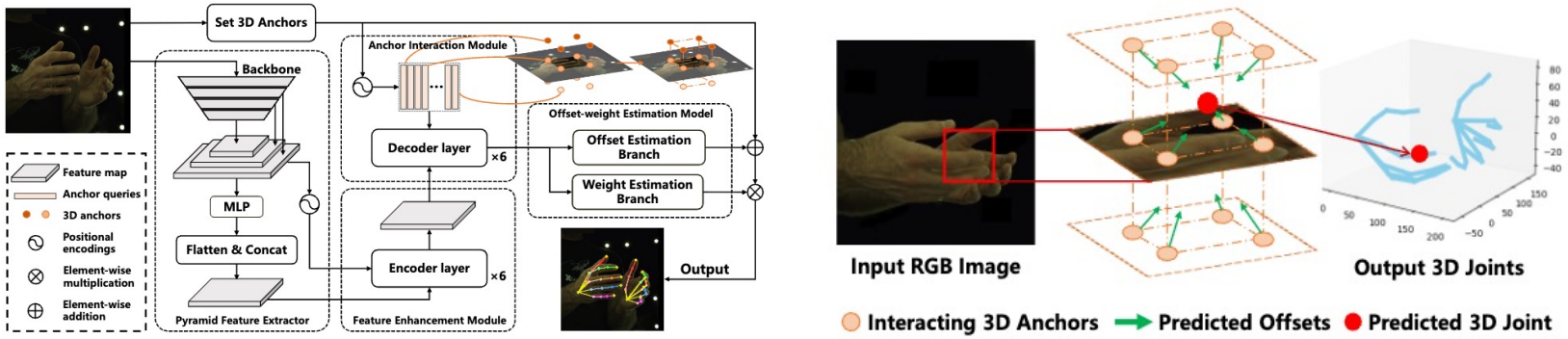
Previous works in 3D hand pose estimation

- Primarily on RGB images-based methods: model-based (output MANO parameters) v.s. model-free

Model-based method, e.g., HaMeR [1]



Model-free method, e.g., A2J-Transformer[2]



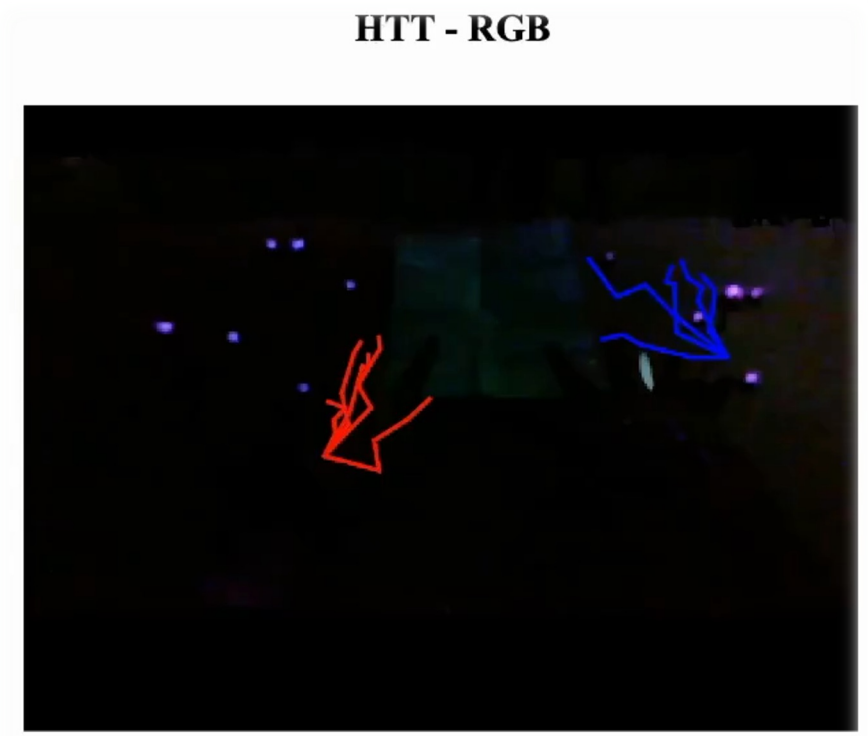
THE UNIVERSITY of EDINBURGH

[1] Pavlakos et al., HaMeR: Reconstructing Hands in 3D with Transformers (CVPR 2024)

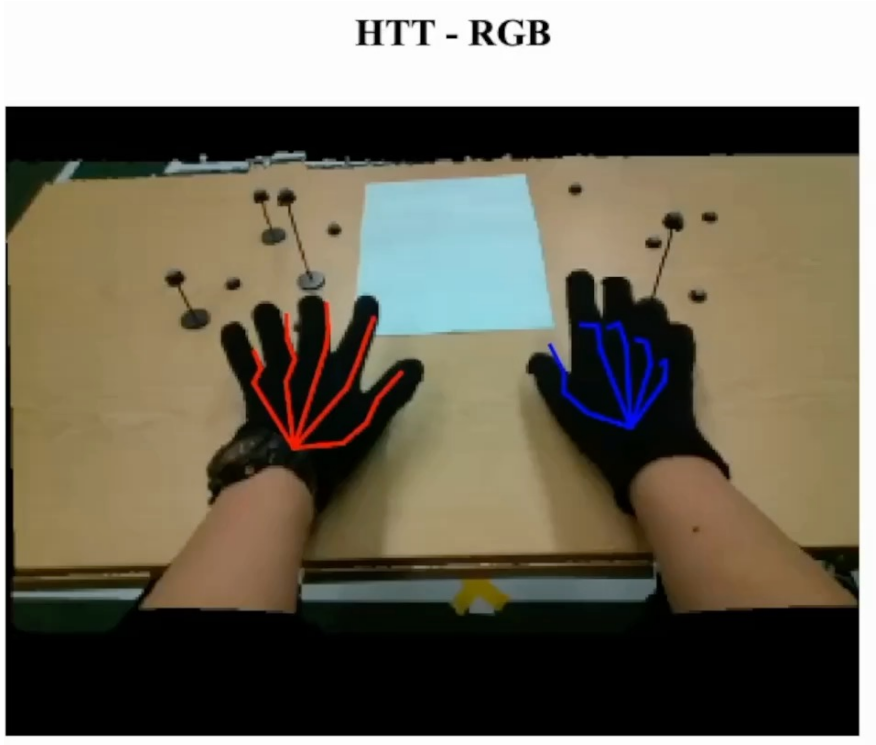
[2] Jiang et al., A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image (CVPR 2023)

Limitations of RGB images-based methods

- RGB-based models rely on **color, texture** – making them brittle in challenging real-world conditions. HTT-RGB[1] model is trained with RGB images collected in normal scenarios.



- Dark environments (e.g., night, warehouses)

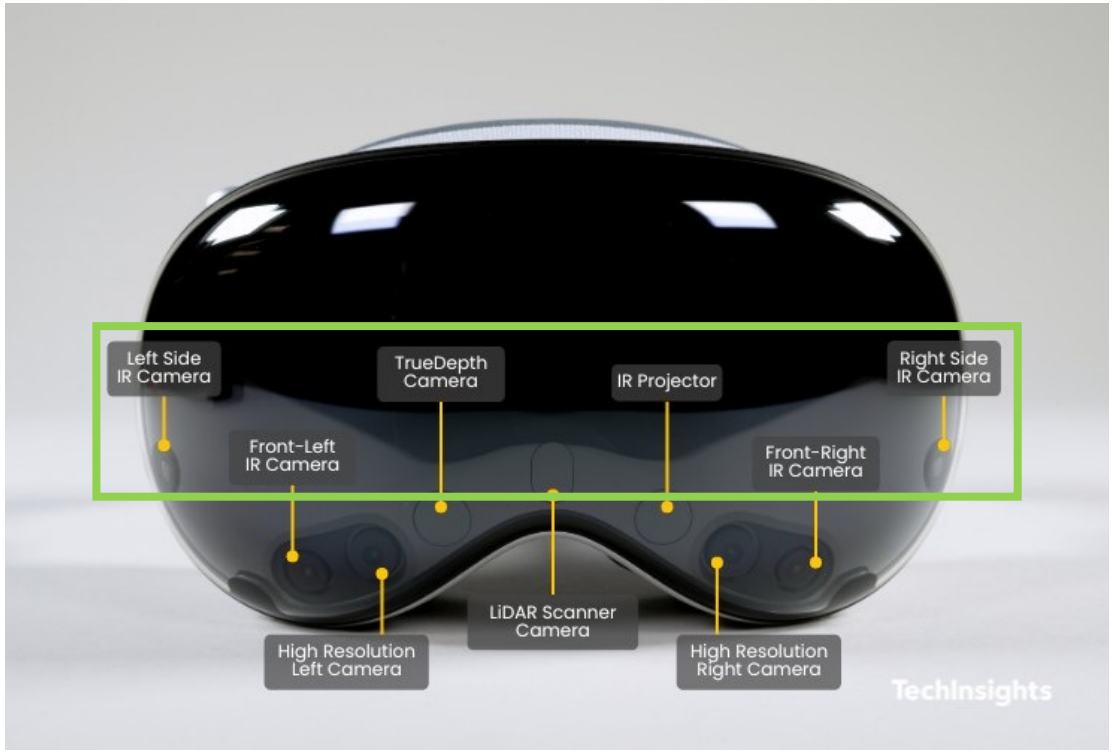


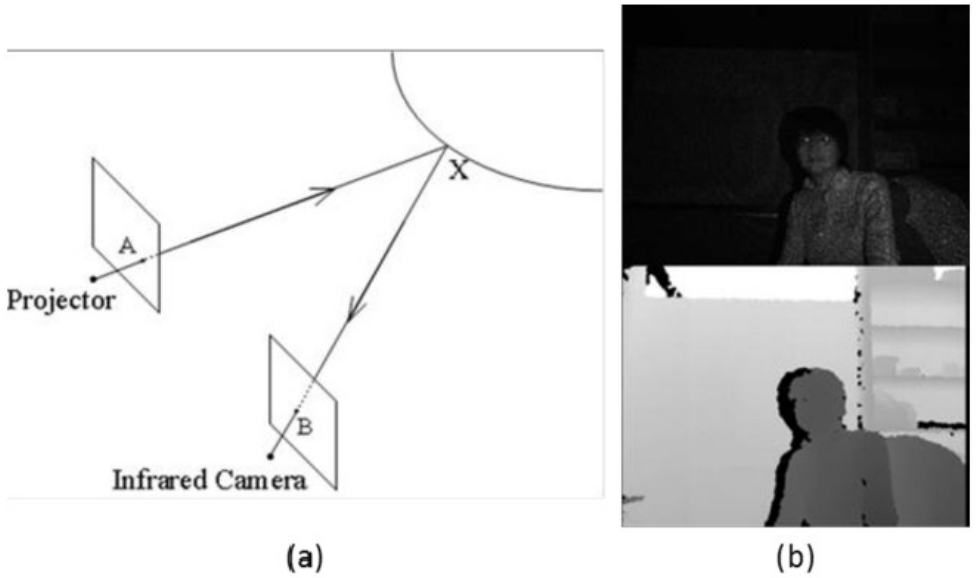
- Handwear like gloves (e.g., cold weather, healthcare)

[1] Wen et al., Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos (CVPR 2023)

Active near infrared (NIR) camera as a supplement

- Current VR/AR devices (e.g., Apple Vision Pro) are equipped with active NIR systems, which provide **depth** measurement, and also facilitate hand tracking under **low-lighting** conditions.





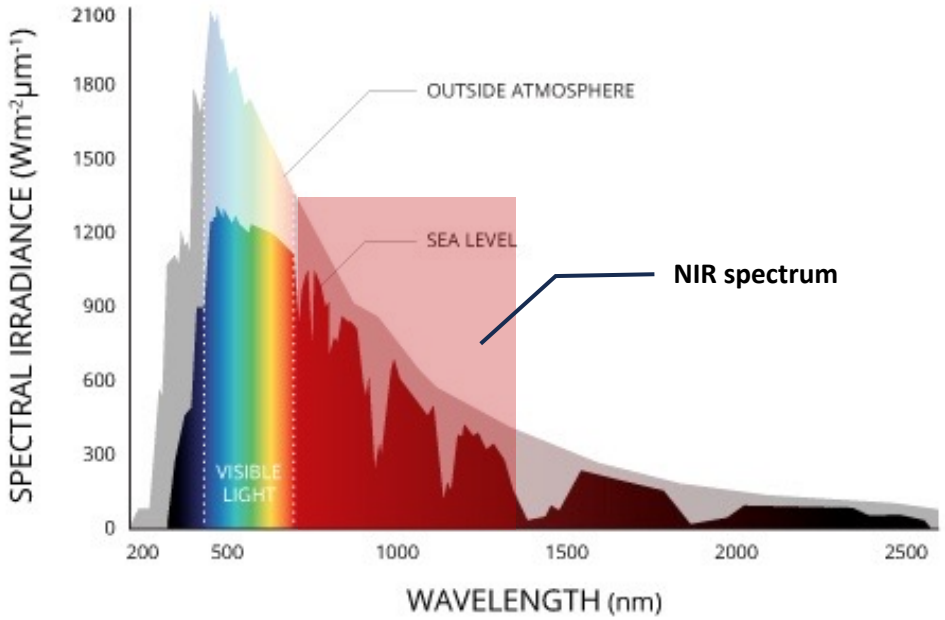
Credit to Wang et al. [1]

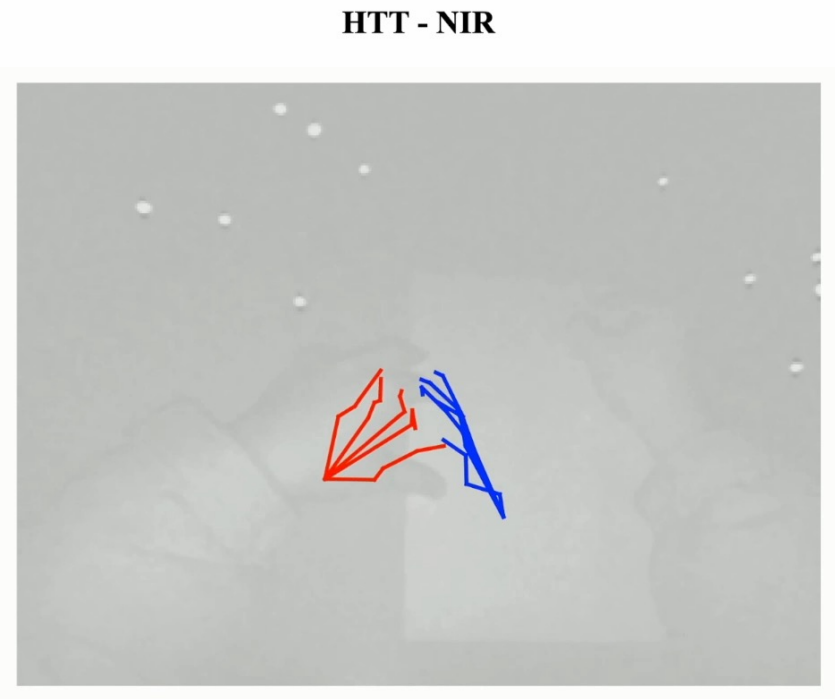
[1] Wang et al., 3D scene reconstruction by multiple structured-light based commodity depth cameras, (ICASSP 2012)

NIR camera - sensitive to other NIR resources

- Sunlight contains strong NIR component (750-1400nm), which saturate active NIR sensors.

Full spectrum of sunlight as seen on Earth[1]

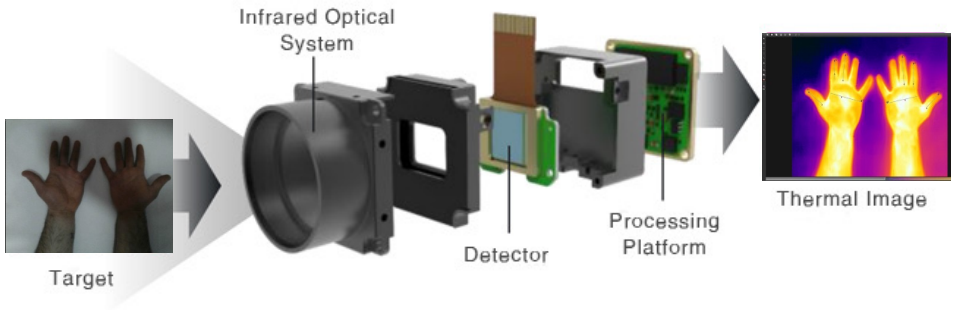




[1] Fordriest Environmental, Inc, Solar Radiation and Photosynethically Active Radiation." Fundamentals of Environmental Measurements, 2014
[2] Wen et al., Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos (CVPR 2023)

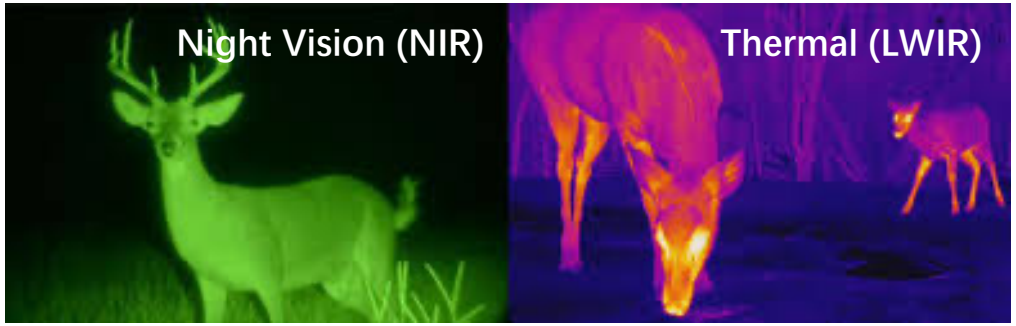
Thermal camera – offer a robust sensing solution for hand pose estimation

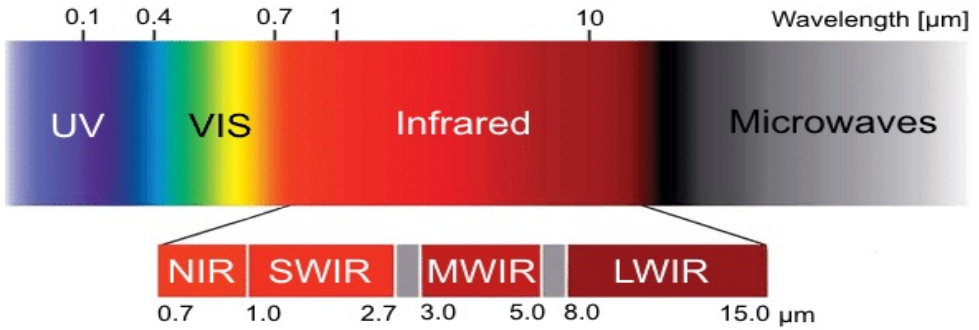
- Detect **long-wave infrared (LWIR)** thermal radiation emitted by the human hands for imaging, without relying on any light source



TELEDYNE FLIR Thermal Camera Products

- Different from night vision, thermal imaging is
  - completely passive sensing
  - capture heat (LWIR) rather than NIR reflection
  - hardly affected by **sunlight**





Research Question

Can egocentric thermal imagery be effectively used for 3D hand pose estimation under various conditions (lighting, handwear), and how does it compare to other spectral imagery (RGB, NIR, depth)?

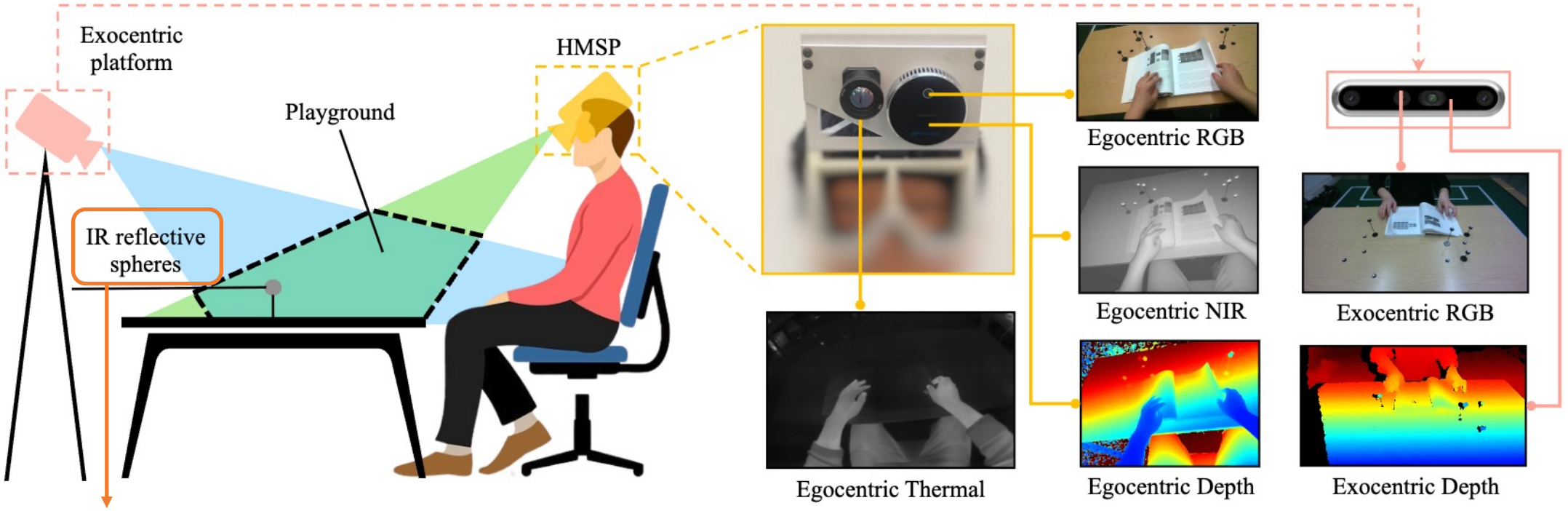
Contribution

- The first-of-its-kind benchmark, **ThermoHand**, of thermal imaging for egocentric 3D hand pose estimation

  - ✓ **Diverse dataset** with 96k synchronized multi-spectral, multi-view images from 28 subject, 19 actions.
  - ✓ **Automatic annotation** pipeline designed to generate 3D hand pose ground truth.
  - ✓ **Bespoken baseline** method, TherFormer, for thermal image-based 3D hand pose estimation.
  - ✓ **Extensive evaluation** on state-of-the-art methods and various spectral images under different conditions.
  - ✓ **Released repository,** including dataset, method implementation and automatic annotation tools.

Multiple-Spectral hand pose dataset

- **Data capture setup** with the customized head-mounted sensor platform (HMSP) and exocentric platform

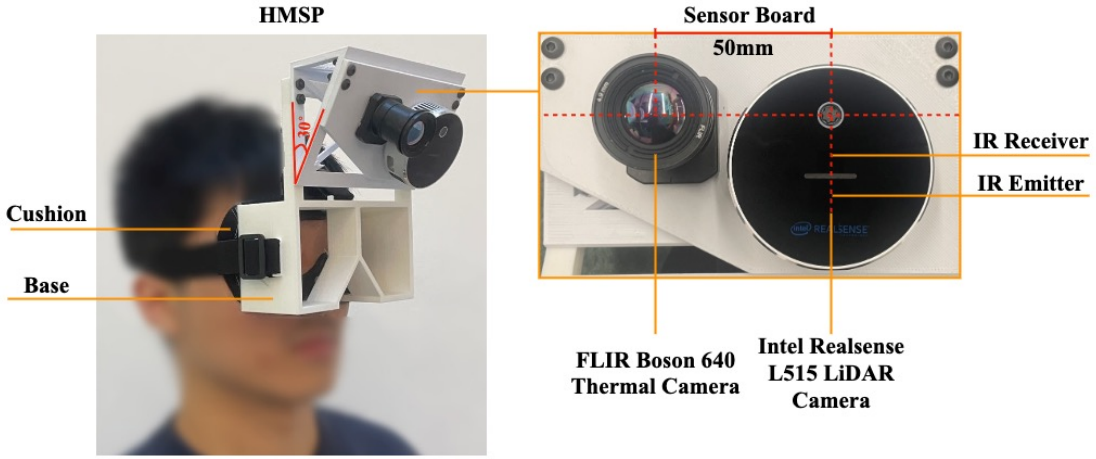


- Our participants are asked to perform predefined **hand-object and hand-virtual** interaction actions within the playground above the table.

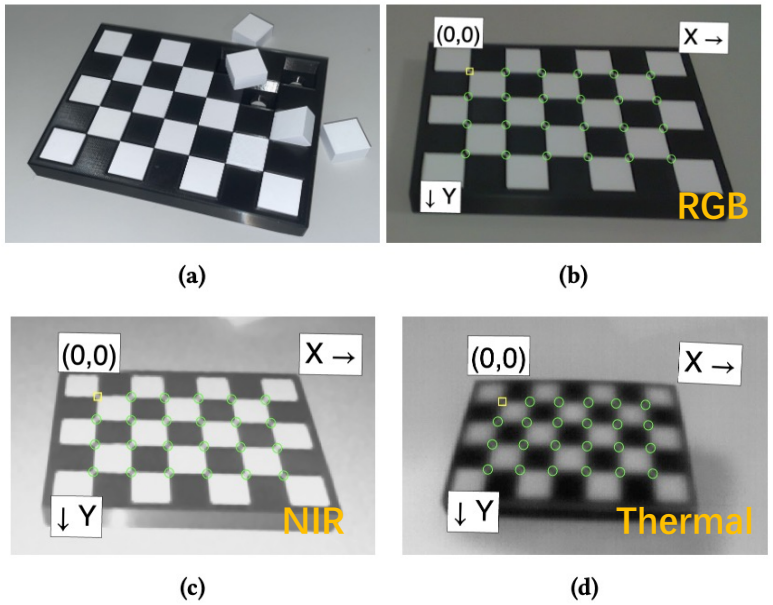
THE UNIVERSITY of EDINBURGH

Multiple-Spectral hand pose dataset

- Design of Head-mounted sensor platform (HMSP)



- Thermal calibration chessboard (self-designed)

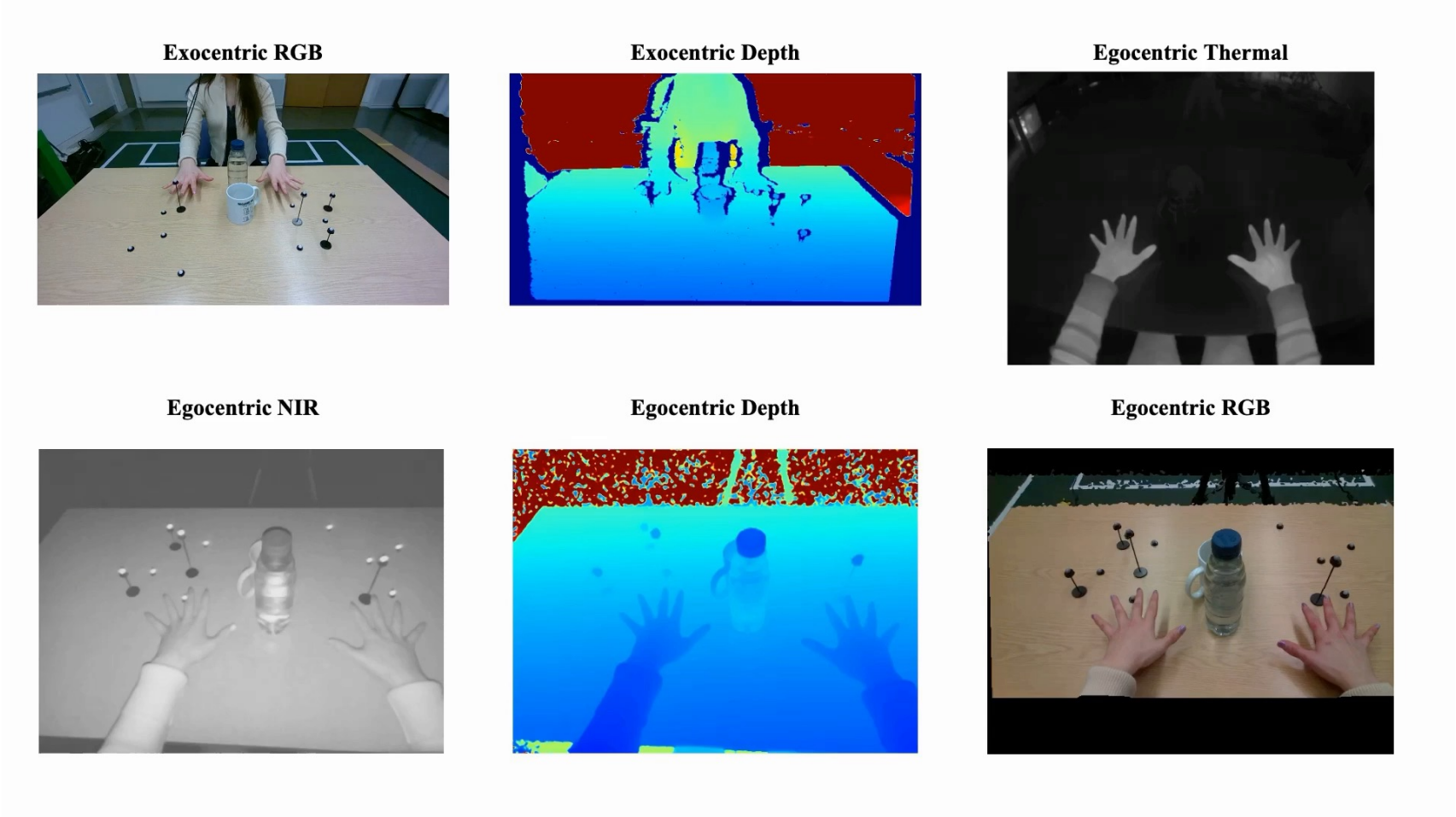


- Dataset statistics – 3 hours with 96K sync frames totally

	Normal office (Main)				Other settings				Total
Setting	train	val	test	sum	darkness	sun glare	gloves	kitchen	
#frames	47,436	12,914	24,002	84,352	3,188	2,508	3,068	2,808	95,924
#seqs	172	43	86	301	12	12	12	14	352
#subjects	16	4	8	28	1	1	1	2	-

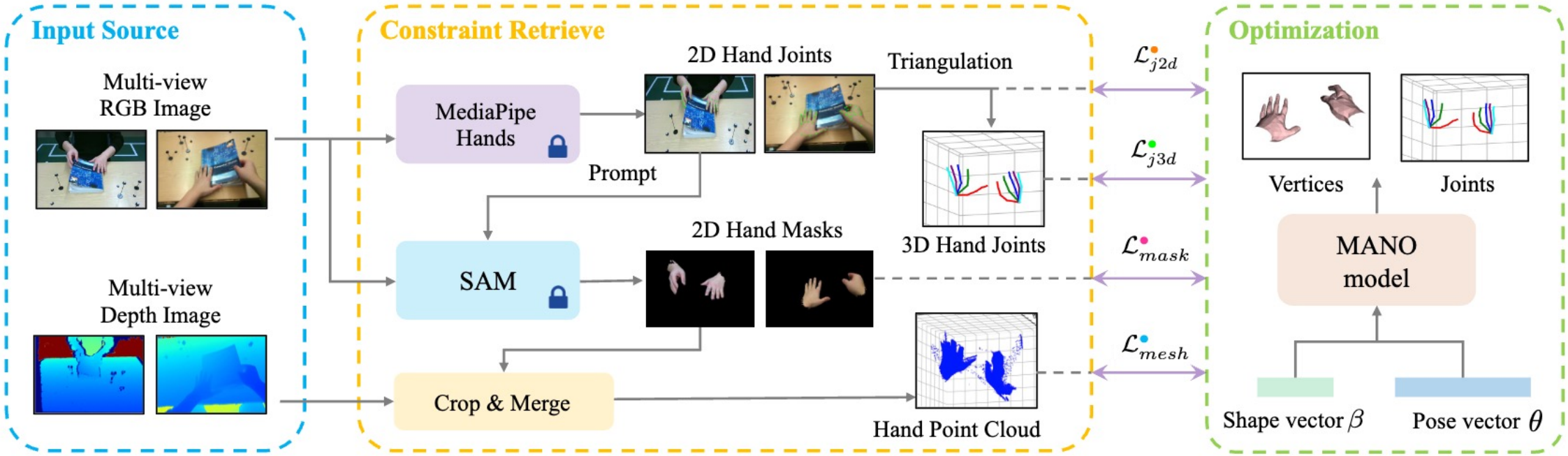
Multiple-Spectral hand pose dataset

- Example of multi-view, multi-spectral data collection (synchronized w.r.t. thermal images, 8.5fps)



Automatic annotation pipeline

- Annotation pipeline: extracting constraints from multi-view RGB and depth images and fit the MANO model to these information by minimizing the combined optimization objective.

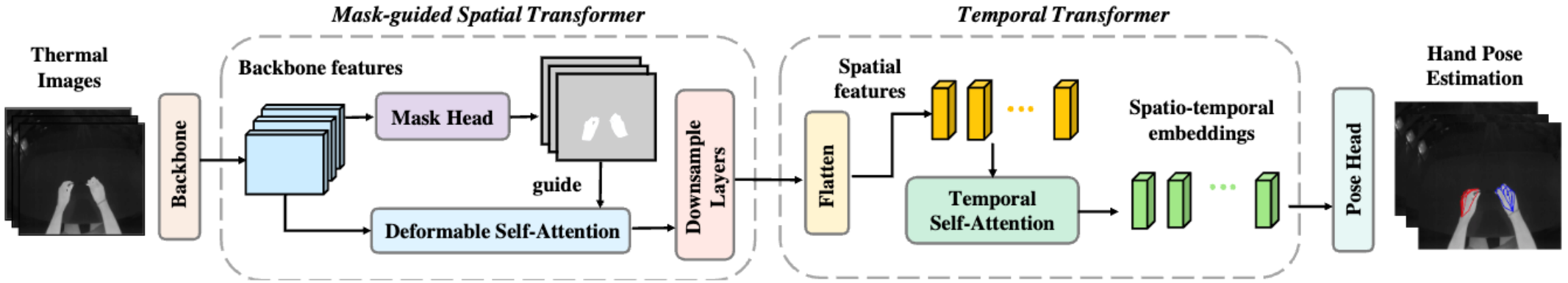


- Optimization objective

$$\theta^* = \arg\min_{\theta} \lambda_{j2d}\mathcal{L}_{j2d}^{\bullet} + \lambda_{mask}\mathcal{L}_{mask}^{\bullet} + \lambda_{j3d}\mathcal{L}_{j3d}^{\bullet} + \lambda_{mesh}\mathcal{L}_{mesh}^{\bullet} + \lambda_{reg}\mathcal{L}_{reg}^{\bullet}$$

TherFormer: a baseline for thermal image-based hand pose estimation

- Transformer-based network design
  - mask-guided spatial transformer: increase the robustness to background clutter
  - Temporal transformer: solve ambiguity and occlusion



- Multi-task loss function

$$L_{Hand} = \|P^{2D} - P_{gt}^{2D}\|_1 + \lambda_1 \|P^{depth} - P_{gt}^{depth}\|_1 + \lambda_2 \|P^{3D} - P_{gt}^{3D}\|_1$$

$$L_{Mask} = -\frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} [\, w_{pos} \cdot M_{wh} \cdot \log(\hat{M}_{wh}) + w_{neg} \cdot (1 - M_{wh}) \cdot \log(1 - \hat{M}_{wh})\,]$$

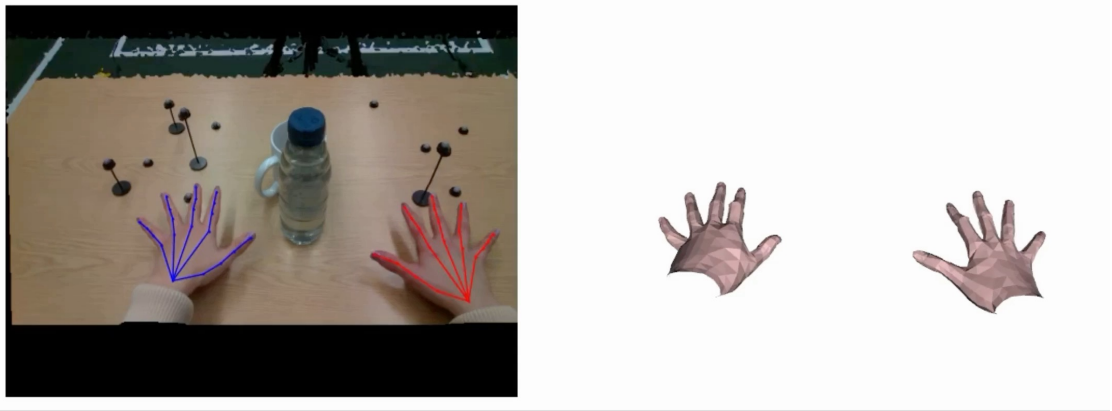
Evaluation of the annotation method

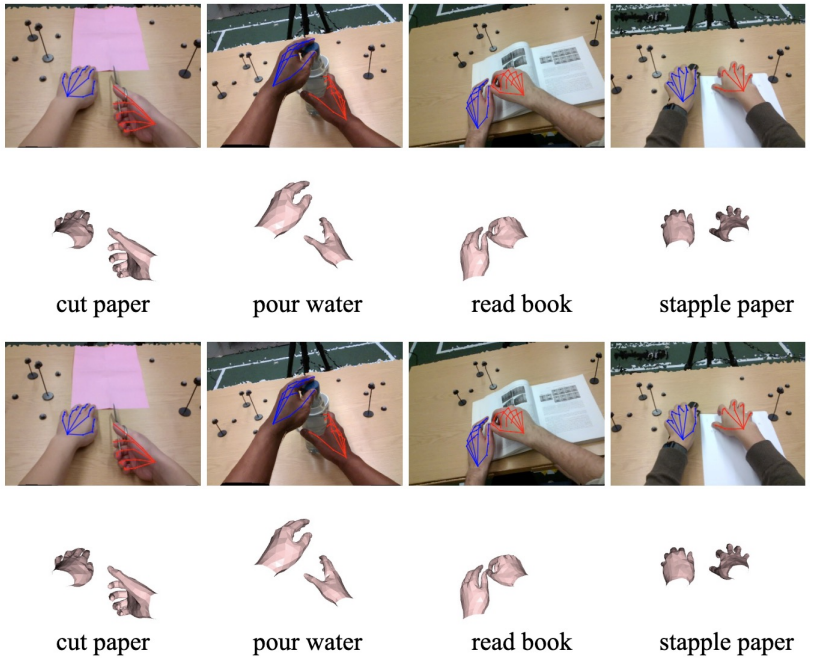
- Annotation accuracy evaluation with different optimization errors combined

	Ego-view optimization		Multi-view optimization			
Errors	$\mathcal{L}_{mask} + \mathcal{L}_{j2d}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh}$	$\mathcal{L}_{mask}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh}$	$\mathcal{L}_{mask} + \mathcal{L}_{j2d} + \mathcal{L}_{mesh} + \mathcal{L}_{j3d}$
mean (std)	37.29 (± 18.02)	7.03 (± 2.57)	8.13 (± 0.57)	1.29 (± 0.43)	1.28 (± 0.43)	1.01 (± 0.34)

nearly 1cm joint error

- Example of automatic annotation results



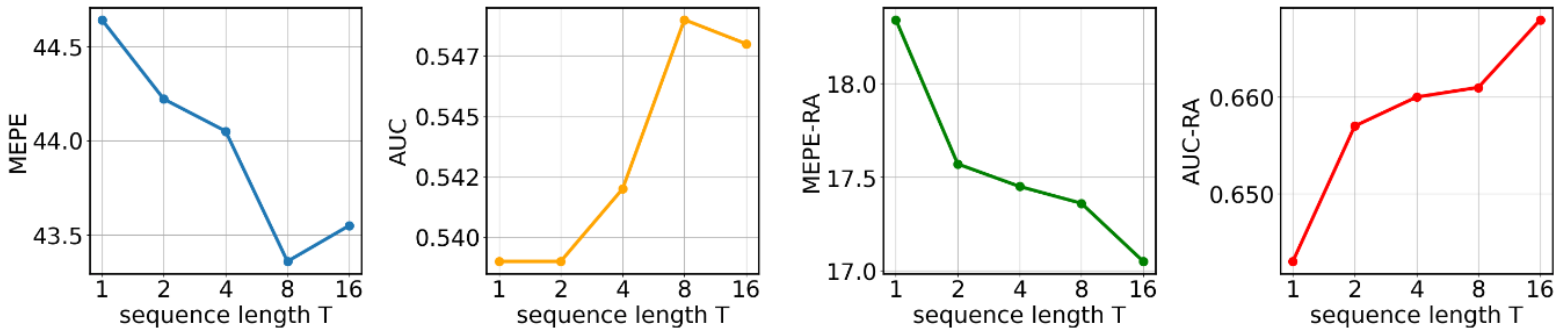


Evaluation - thermal Image-based 3D hand pose estimation

- Comparison between TherFormer to SOTA methods. S – single image input; V – video input.

	Method	Input	MEPE (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC-RA ↑	fps ↑
(a)	HaMeR* [105]	Single	-	-	20.88	0.598	118
(b)	A2J-Transformer [46]	Single	51.68	0.474	20.76	0.603	34
(c)	HTT [15]	Single	49.09	0.489	20.69	0.599	**211**
(d)	TherFormer-S	Single	**44.64**	**0.539**	**18.34**	**0.643**	136
(e)	(c) w/o spatial transformer	Single	48.79	0.491	20.15	0.609	174
(f)	(c) w/o mask guidance	Single	48.83	0.494	18.89	0.625	141
(g)	HTT [15]	Sequence	47.07	0.512	17.49	0.659	**129**
(h)	TherFormer-V	Sequence	**43.36**	**0.549**	**17.36**	**0.661**	52

- Support flexible input length: impact of temporal sequence length to TherFormer



THE UNIVERSITY of EDINBURGH

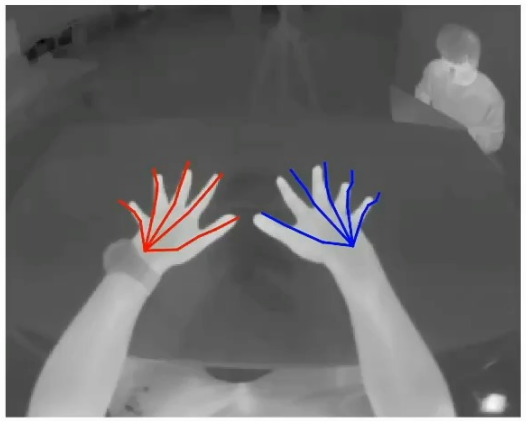
Evaluation – comparison between spectrum

- Visualization of 3D hand pose estimation under challenging conditions.
  - Lighting conditions: darkness and sun glare

Thermal vs. RGB in a Dark Environment

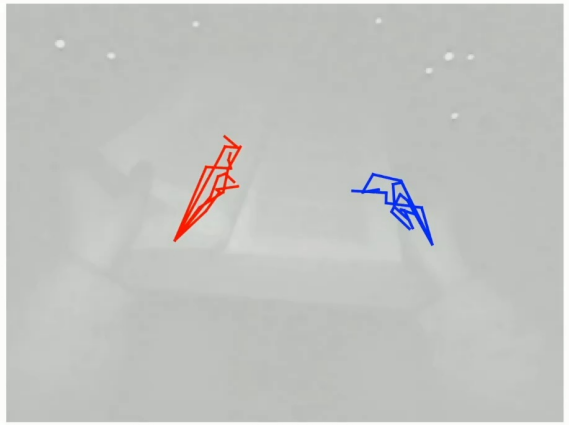
HTT - RGB

HTT - Thermal

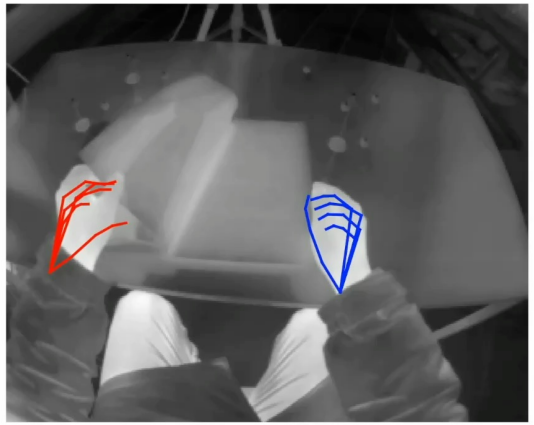


Thermal vs. NIR with Sun Glare

HTT - NIR

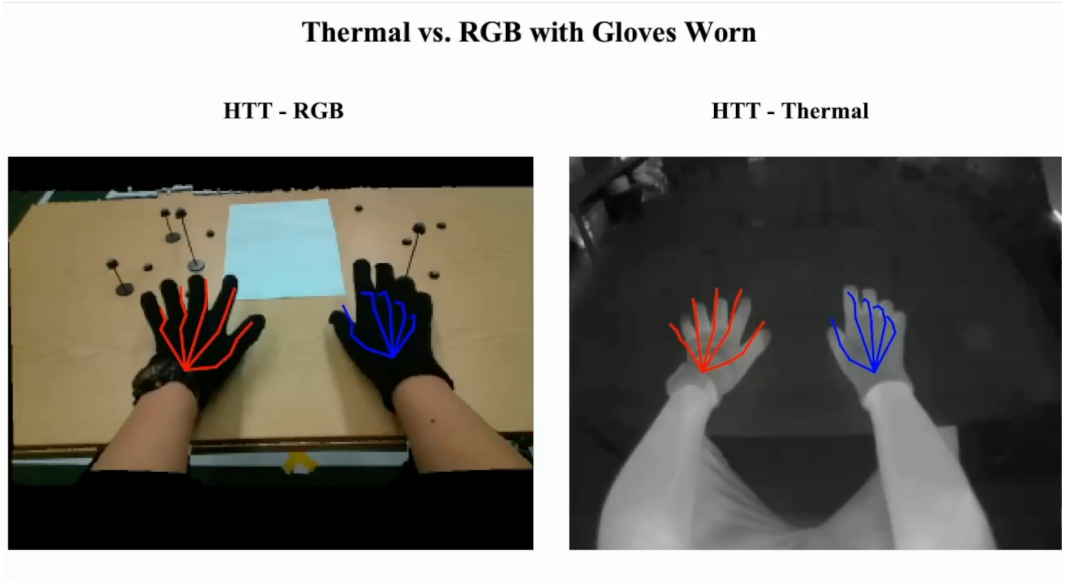


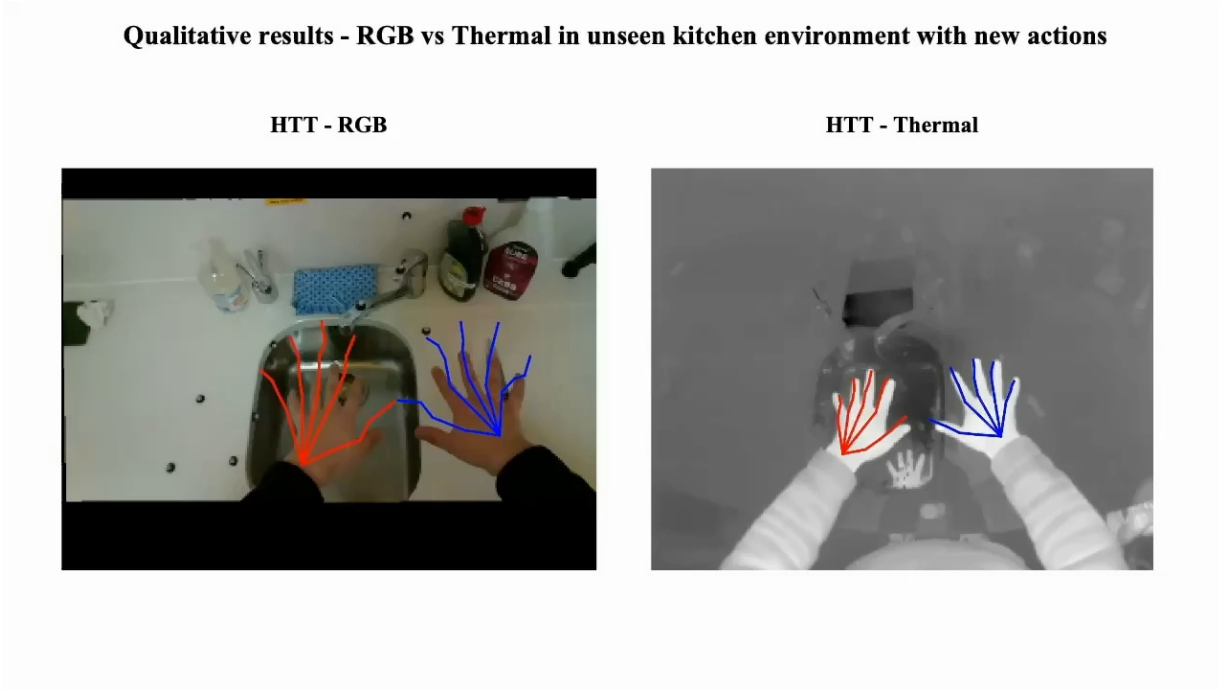
HTT - Thermal



Evaluation – comparison between spectrum

- Handwear like glove
- Generalization to unseen action and environment





Evaluation – comparison between spectrum

- Evaluation results under challenging conditions

	TherFormer-V (glove)		TherFormer-V (sun glare)	
Spectrum	MEPE-RA (mm) ↓	AUC ↑	MEPE-RA (mm) ↓	AUC ↑
RGB	51.94	0.141	38.24	0.252
Depth	45.96	0.206	42.27	0.254
NIR	39.83	0.282	90.84	0.093
Thermal	**39.23**	**0.302**	**32.56**	**0.363**

Thermal images can not only serve as supplements in challenging cases but also can be a viable alternative to other spectra in normal conditions.

- Evaluation results under normal conditions

	HTT (Sequence)		TherFormer-V		TherFormer-S		Best	
Spectrum	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑	MEPE (mm) ↓	AUC ↑
RGB	43.30	0.542	43.50	0.542	44.61	0.529	43.30	0.542
Depth	41.62	0.559	**39.70**	**0.581**	**39.84**	**0.579**	**39.70**	**0.581**
NIR	**41.57**	**0.562**	40.79	0.575	40.98	0.575	40.79	0.575
Thermal	47.07	0.512	43.36	0.549	44.64	0.539	43.36	0.549

Key takeaways

🖐️ Egocentric Hand Pose Estimation Powers Physical AI
- 🕶️ Enables intuitive interaction in VR/AR
- 🤖 Fuels robot manipulation learning from human demos

⚠️ RGB & NIR Have Real-World Limitations
- ❌ RGB fails in low-light, and handwear condition
- ❌ NIR breaks outdoors — sunlight overwhelms IR signals

🔥 Thermal Imaging (LWIR): A Robust Alternative
- ✅ Detects emitted heat, not reflected light
- ✅ Works in darkness, with gloves, and under sunlight

📦 ThermoHands: The first-of-it-kind benchmark work
- 📸 96K multi-view, multi-spectral images
- ⚙️ Automatic 3D annotation pipeline
- 🧠 Custom thermal baseline: TherFormer
- 📊 Extensive cross-spectrum evaluation
- 🔓 Public dataset, code, and tools released

ThermoHands

A Benchmark for 3D Hand Pose Estimation from Egocentric Thermal Images

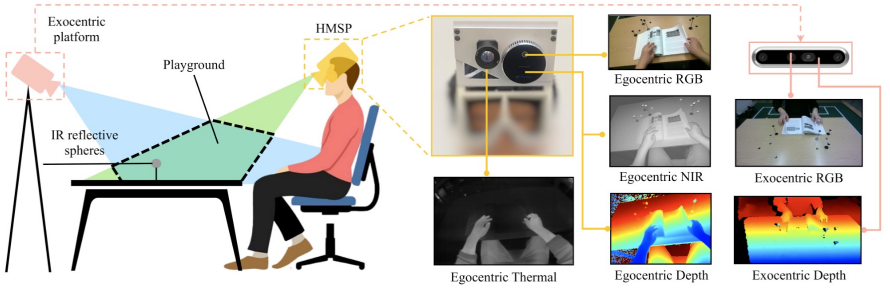
ACM Sensys 2025

Fangqiang Ding[*1] Yunzhou Zhu[*2] Xiangyu Wen[1] Gaowen Liu[3] Chris Xiaoxuan Lu[4]

[*] Denotes equal contribution

1 THE UNIVERSITY of EDINBURGH 2 Georgia Institute of Technology 3 cisco 4 UCL

arXiv Code Demo Data-A Data-B



Code

Paper

Thank you!

Page

Video