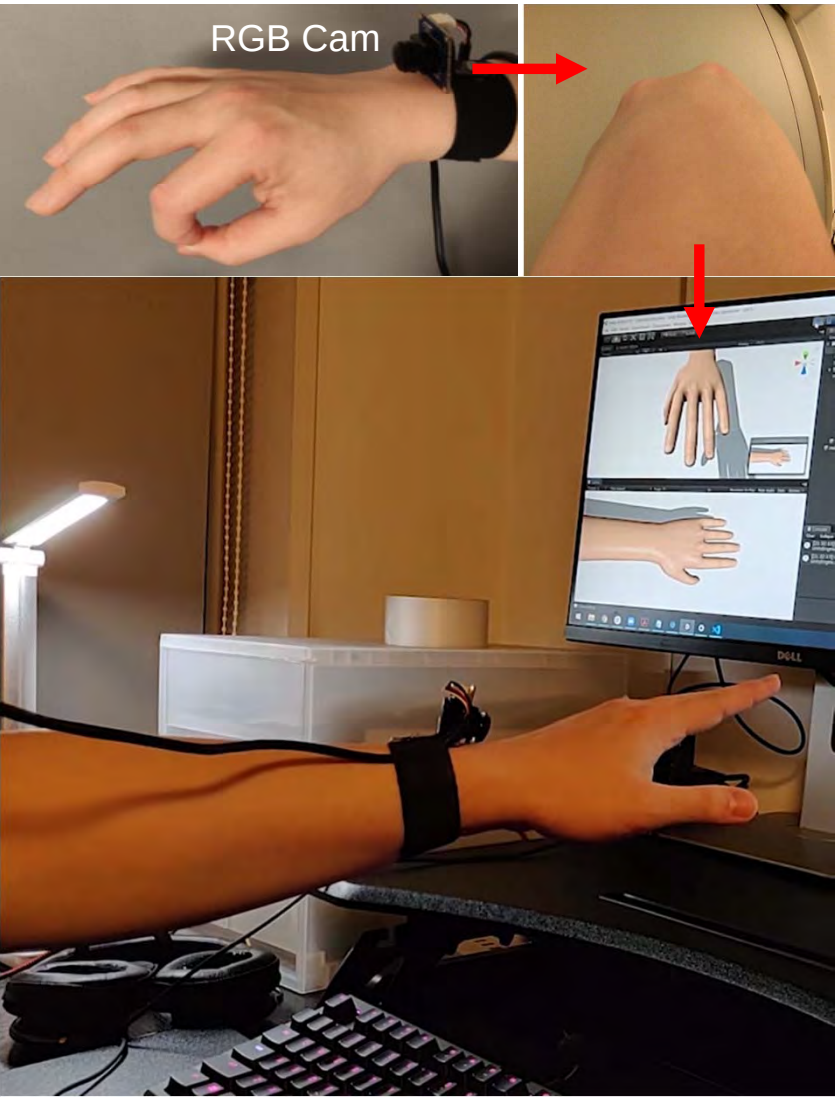


Back-Hand-Pose:

3D Hand Pose Estimation for a Wrist-worn Camera via Dorsum Deformation Network

Erwin Wu
PhD 1st Student
JSPS Research Fellow (DC1)
Tokyo Institute of Technology

Kris Kitani
Associate Professor,
Carnegie Mellon University
Research Fellow,
University of Tokyo

Hideki Koike
Professor,
School of Computing,
Tokyo Institute of Technology

Japan Society for the Promotion of Science
日本学術振興会

Carnegie Mellon University
The Robotics Institute

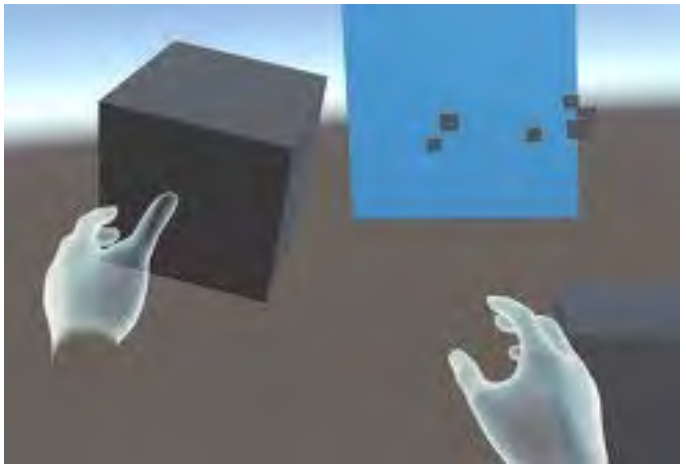
東京工業大学
Tokyo Institute of Technology

Background

Human beings get used to interact with object using their hands.

Touch Screen

Gesture Control

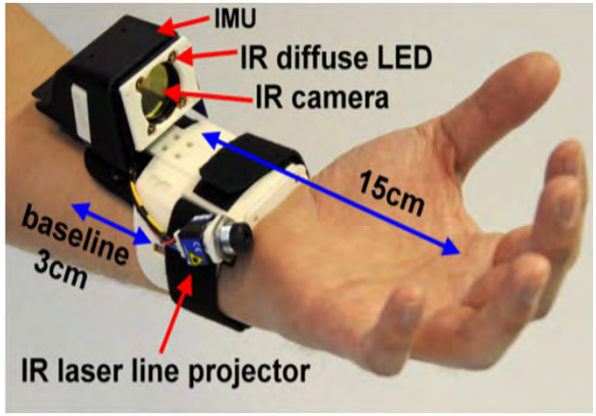


VR/AR

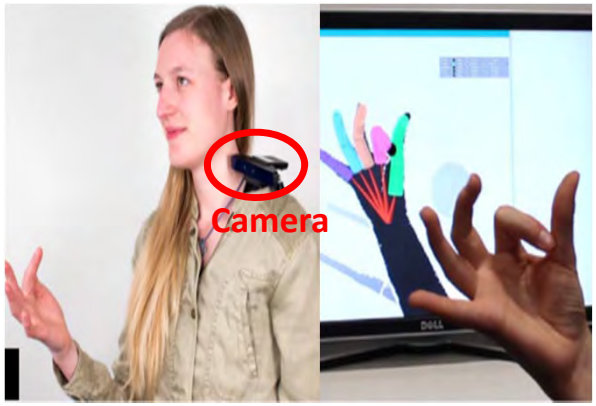
Limitations of Previous Work

Oculus Quest Hand

Glove Technologies



Digits (UIST2012)



FingerInput (ISS2018)

Leap Motion

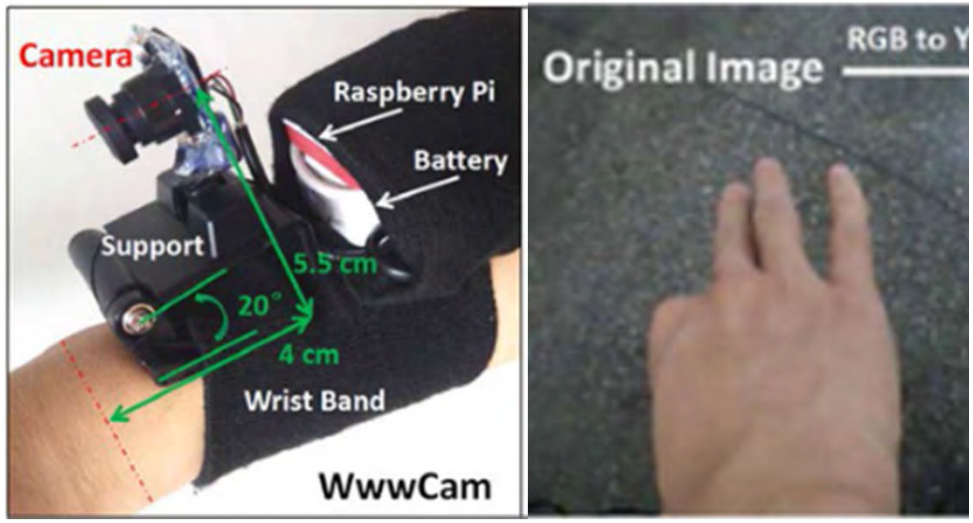
Zhou et al. (CVPR2020)

Related Work

CyclopsRing
(Chan et al. UIST 2015)

Using ring-like fisheye camera between the finger

Finger Angle-Based Hand Gesture Recognition
(Chen et al. Appl. Sci. 2018)



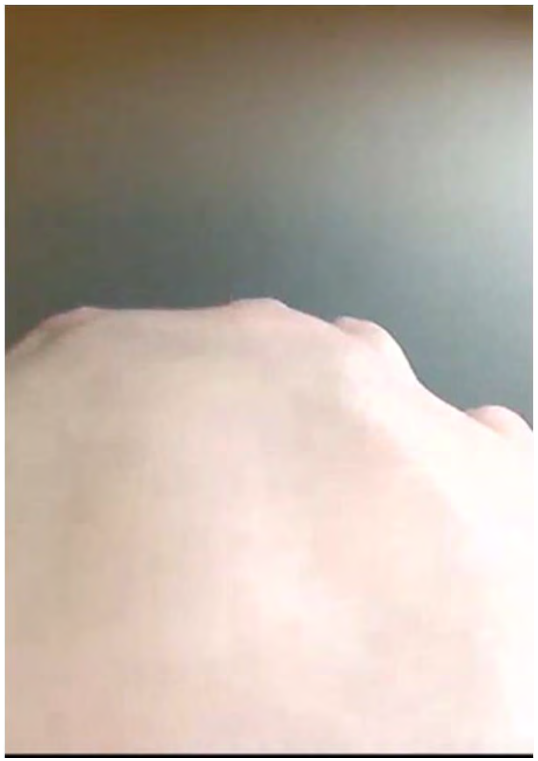
Using an elevated wrist-worn camera to recognize finger angle

Question

Do we really need to see the fingers?

Background

Some Commercial Products



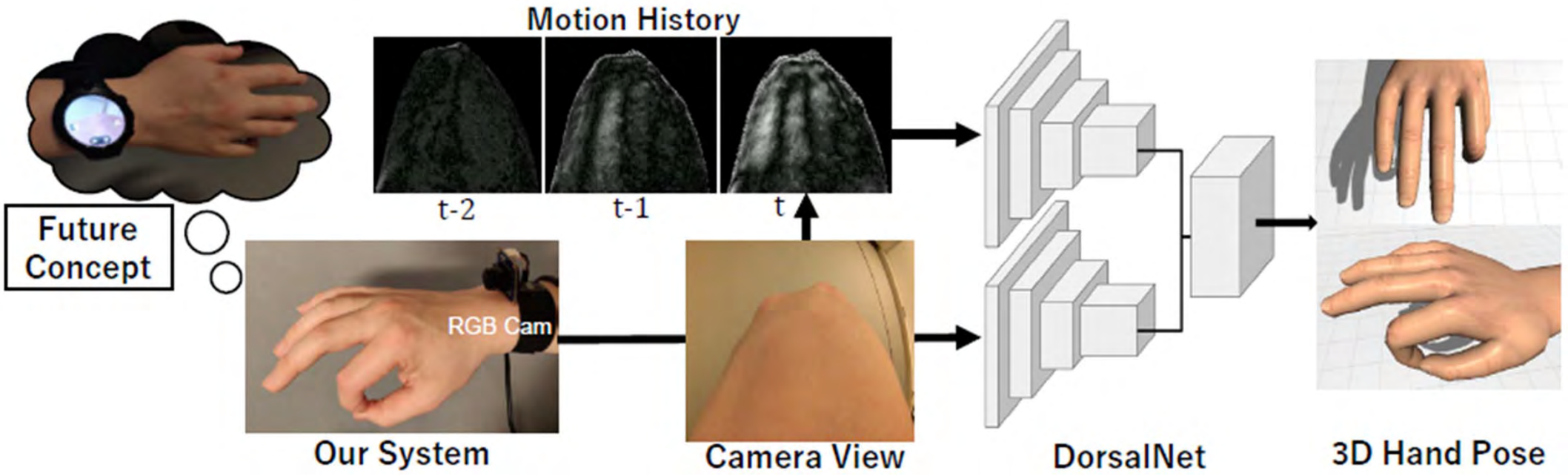
Camera View

Our Previous Work

Gesture Recognition from back hand image using depth camera.

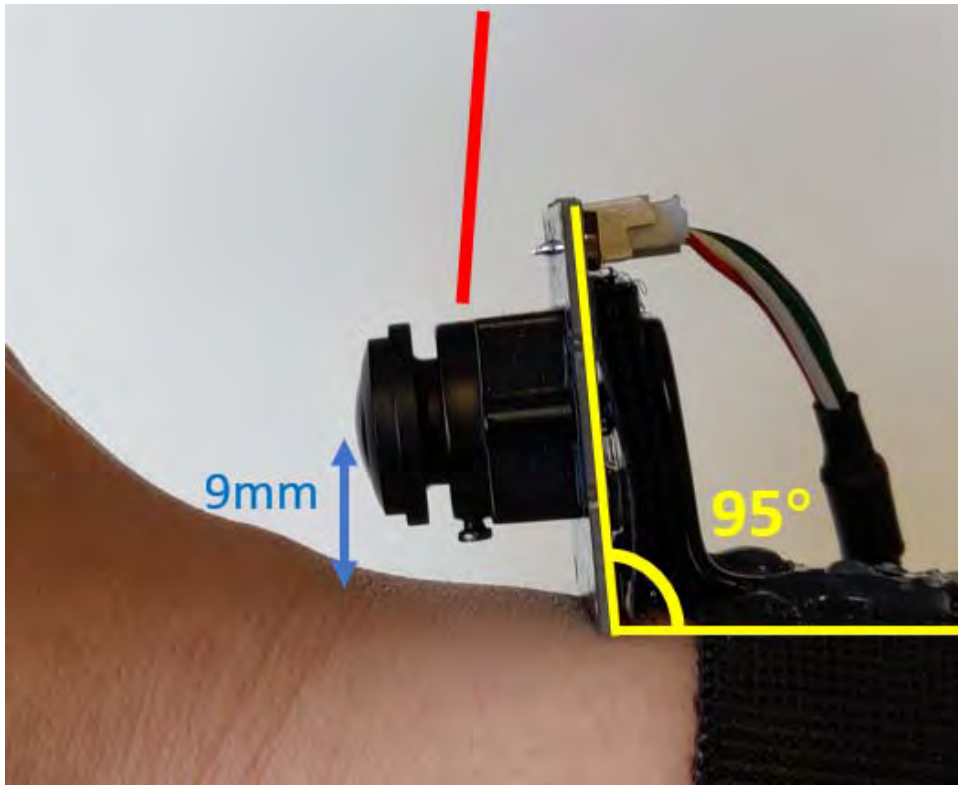
Opisthenar
Wu et al. (UIST 2019)

Our Approach



Hardware Design

2.1mm 120° lens 30-FPS Camera



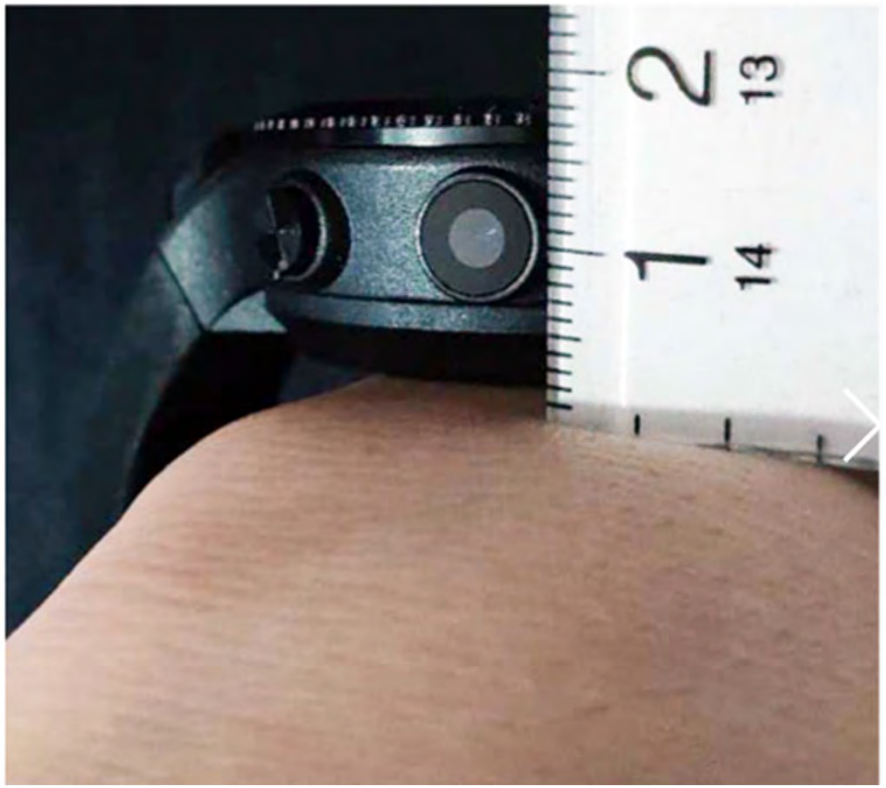


Image Preprocessing

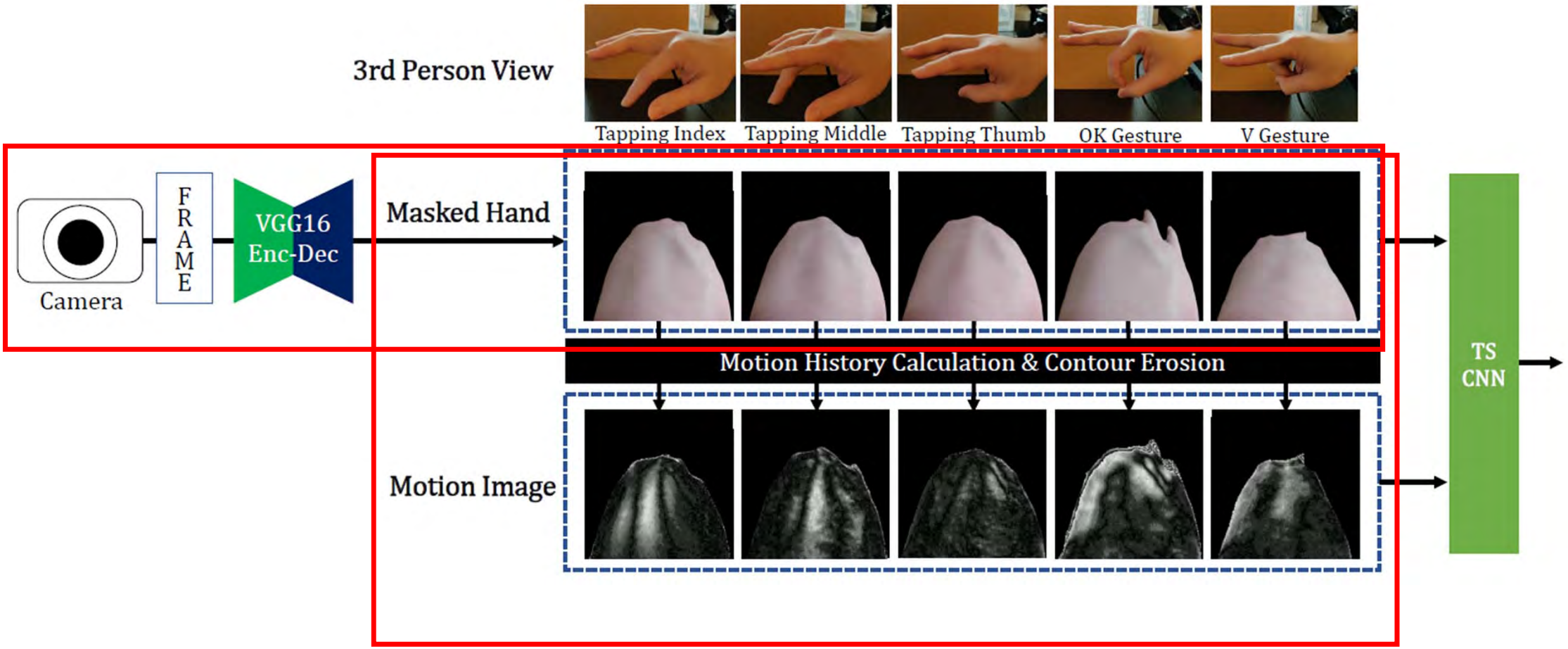
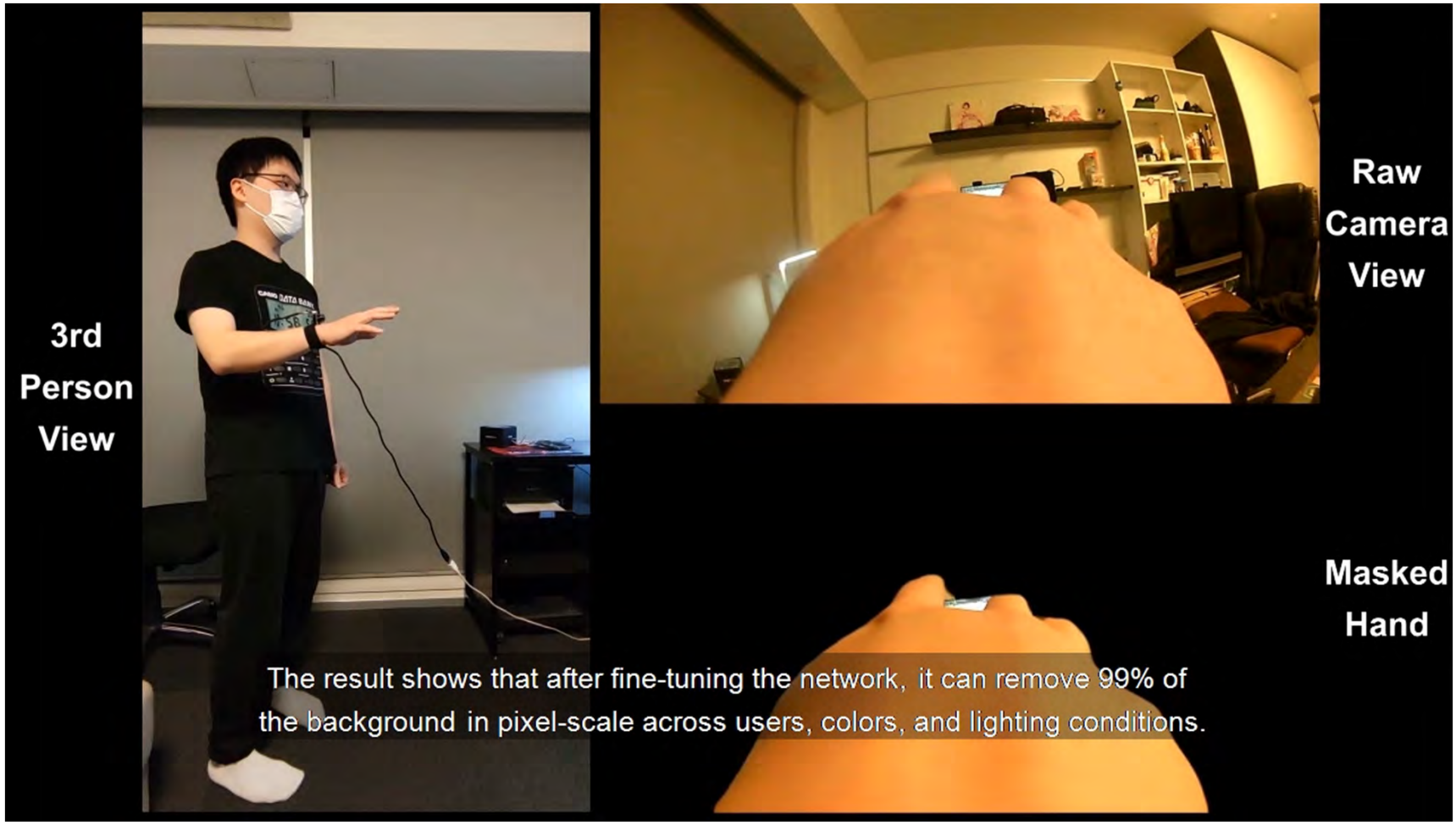
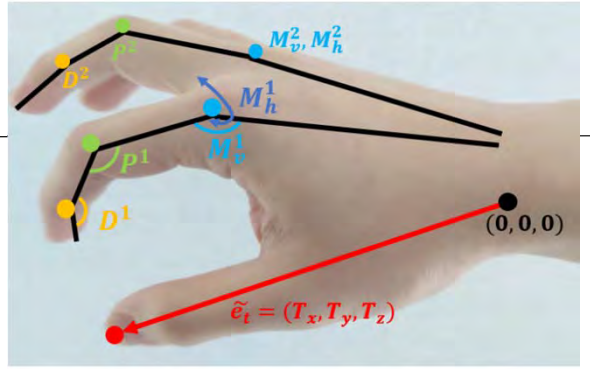
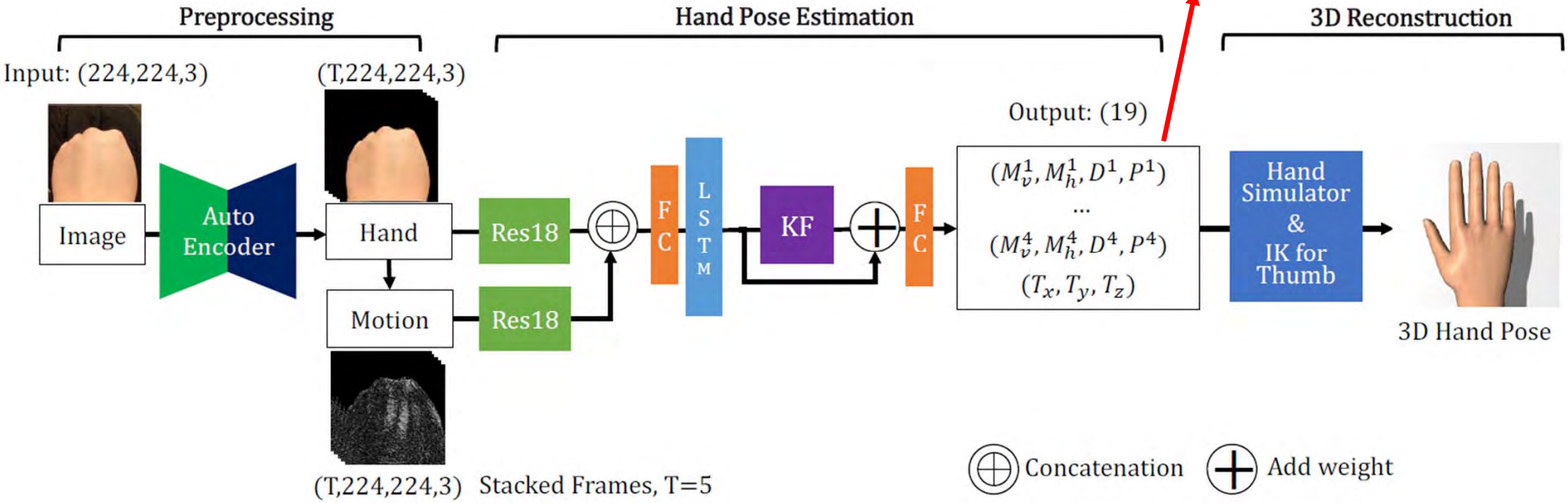


Image Preprocessing



Network Architecture





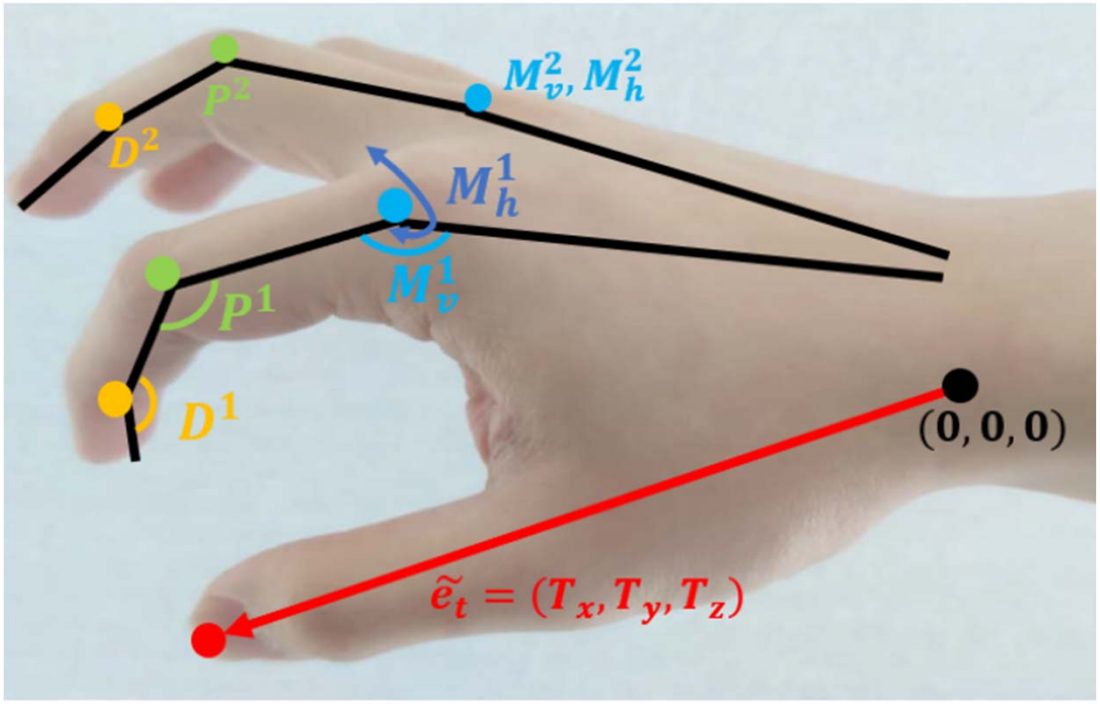
Hand Representation

$$(M_v^1, M_h^1, D^1, P^1)$$

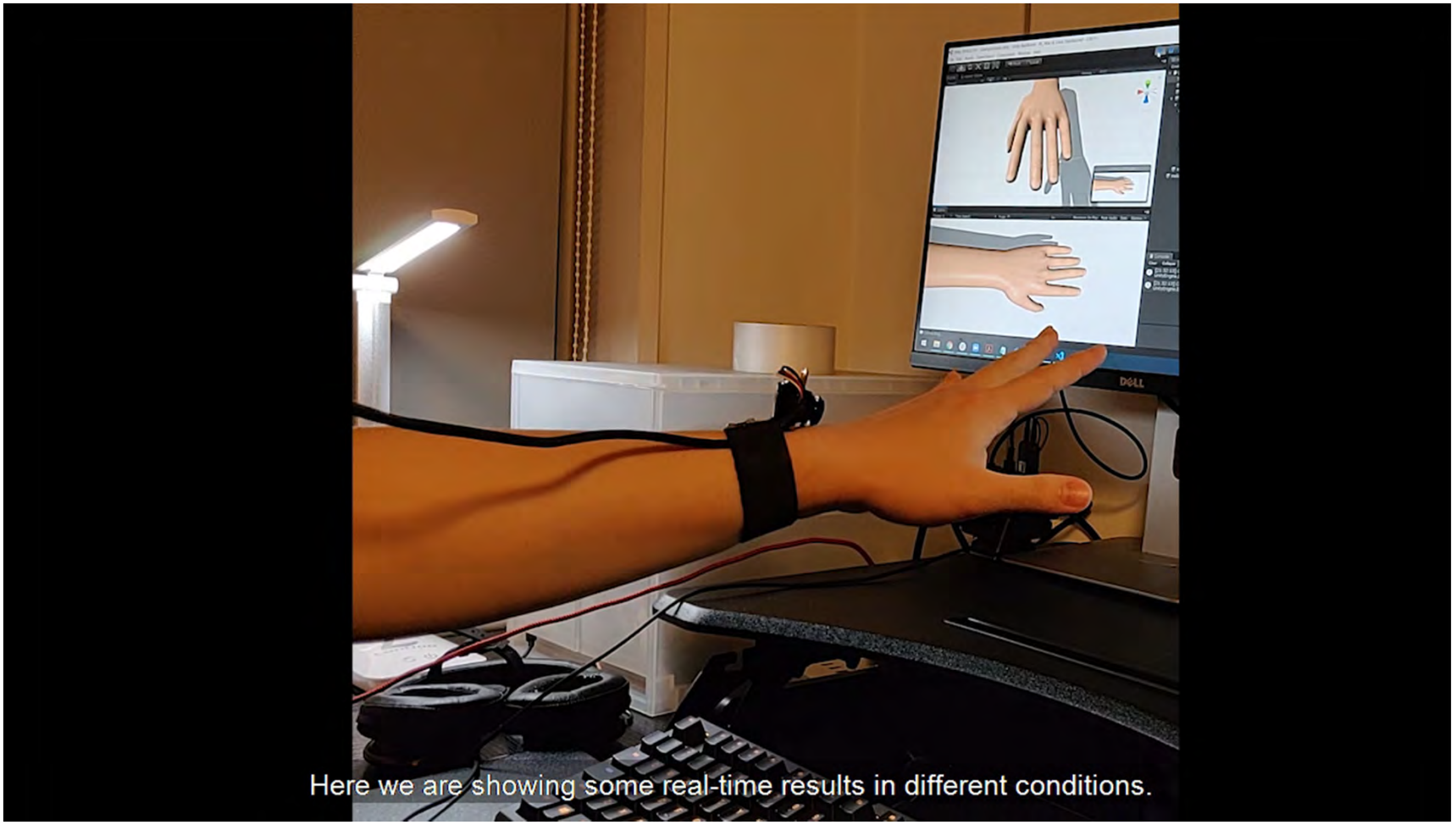
$$\dots$$

$$(M_v^4, M_h^4, D^4, P^4)$$

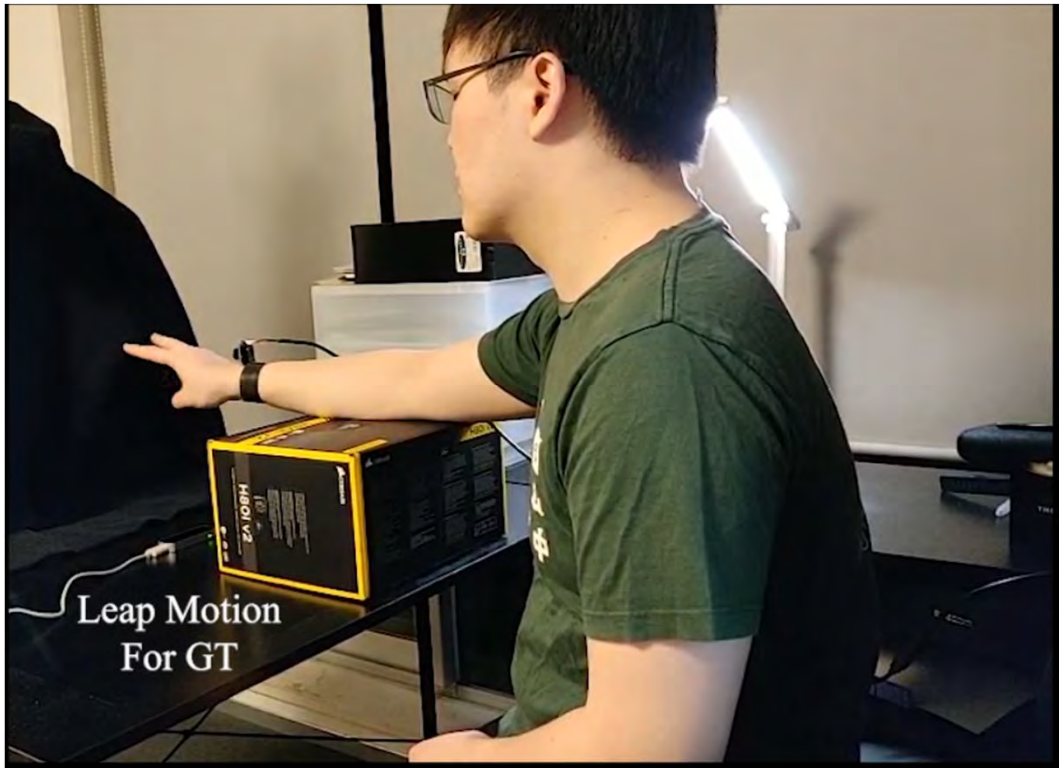
$$(T_x, T_y, T_z)$$



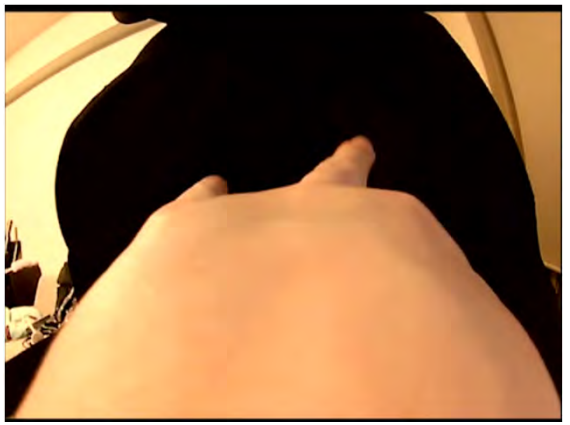
Demo


Here we are showing some real-time results in different conditions.

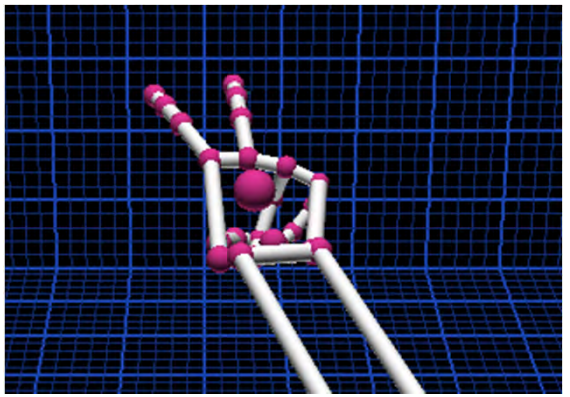
Study: Data Collection



Data Collection Environment



Raw Data



Ground Truth

Study: Data Collection

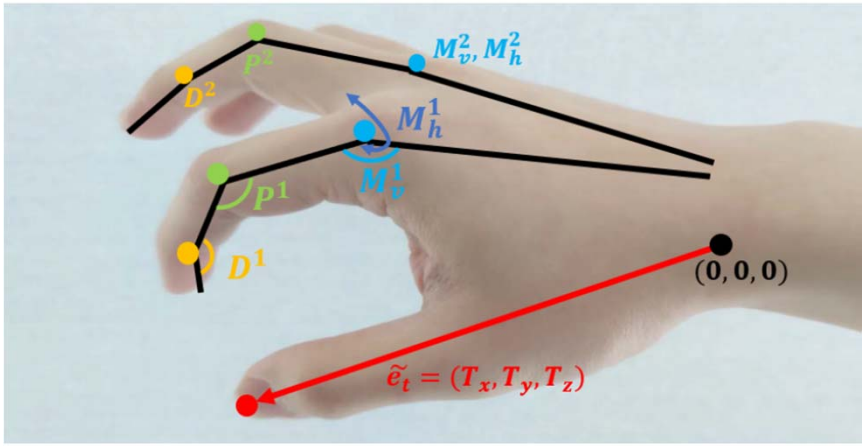
- Subject: 5 Male (age 25-32)
- Pose: 10 static gestures (ASL number)
  5 dynamic tapping (each finger)
- Length: 20FPS@30s * 5 sessions / gesture
- Number: 225,000 frames in total

Study: 3D Hand Pose Estimation Accuracy

Evaluation on each joint and finger:

Joint\Finger	Index (1)		Middle (2)		Ring (3)		Pinky(4)		Joint Avg.	Thumb (0)	
	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	MAE	SD
MCPv	7.05°	±0.40	6.32°	±0.54	6.3°	±0.39	6.92°	±1.21	**6.65°**		
MCPh	7.94°	±0.75	7.87°	±0.62	7.17°	±0.64	9.78°	±1.99	**8.14°**		
DIP	6.92°	±0.59	6.78°	±0.76	6.70°	±0.70	9.80°	±1.73	**7.55°**	12.69°	±2.26
PIP	8.47°	±0.92	7.85°	±0.98	7.66°	±0.87	11.11°	±1.33	**8.77°**		
Finger Avg.	**7.60°**		**7.20°**		**6.96°**		**9.40°**		\		

Table 1. Average result of the individual model of each joint/finger (metrics: MAE(SD) unit: degree).

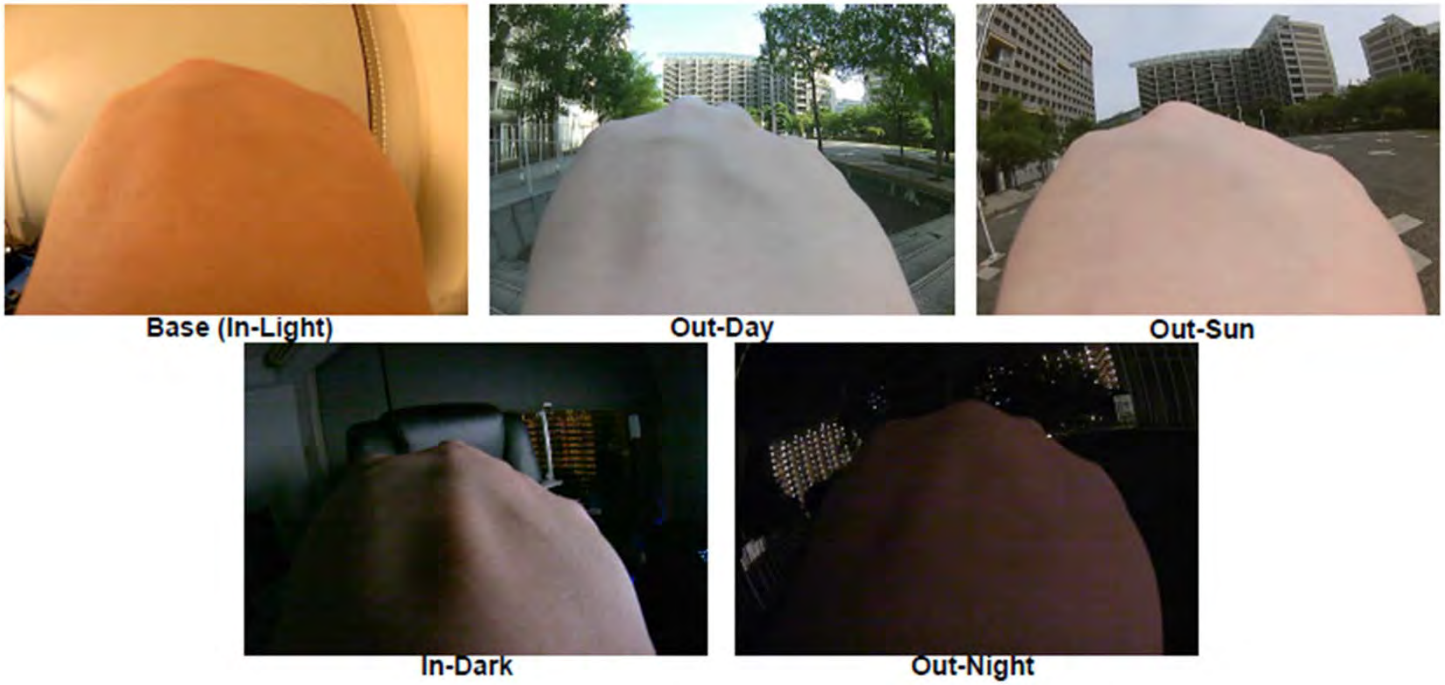


Study: 3D Hand Pose Estimation Accuracy

Method	Individual	General	Leave-1-user
Nearest N.[13]	18.44°	21.78°	20.89°
Direct(ResNet18)	18.39°	22.09°	29.11°
Yuan et al. [55]	12.48°	14.40 °	14.53°
Yeo et al. [53]	16.67°	18.52°	20.24°
Zhou et al. [59]	15.95°	20.06°	21.06°
Ours (w/o KF)	9.28°	10.33°	10.71°
Ours (w/ KF)	**8.81°**	**9.77°**	**9.72°**

Table 2. Comparison with baseline methods, Our methods are divided into with/without Kalman filter (KF).

Study: Different Lighting Condition



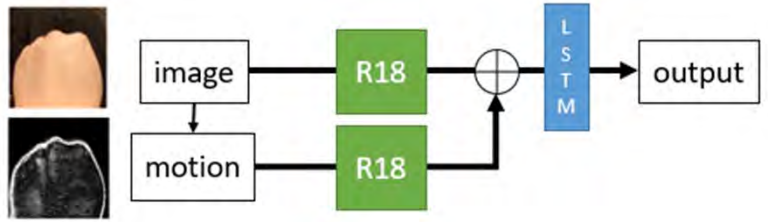
	Base(In-Light)	Out-Day	In-Dark	Out-Night
MAE	7.93°	7.77°	8.46°	8.21°

Table 3. Comparing the accuracy of our method in different lighting condition (Out-Sun removed due to lack of ground truth).

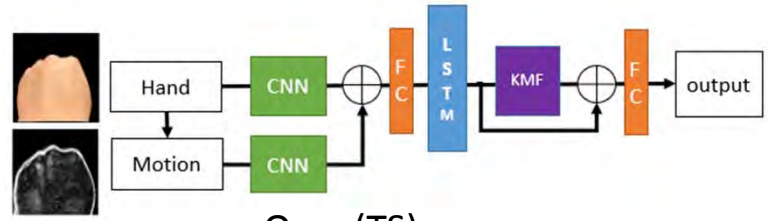
Study: Ablation Study

VGG16(RGB)

ResN18+LSTM (RGB)



ResN18+LSTM (TS)

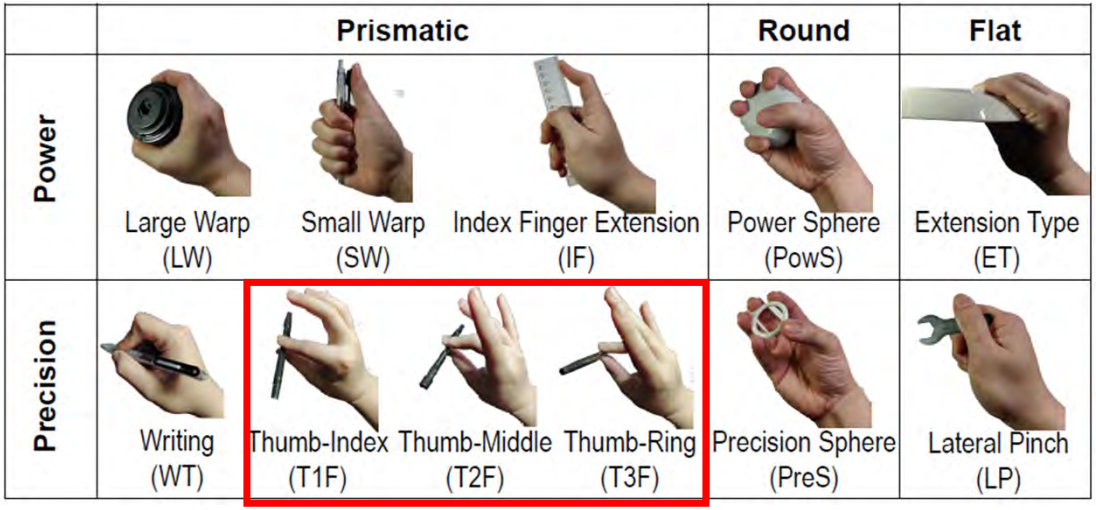


Ours (TS)

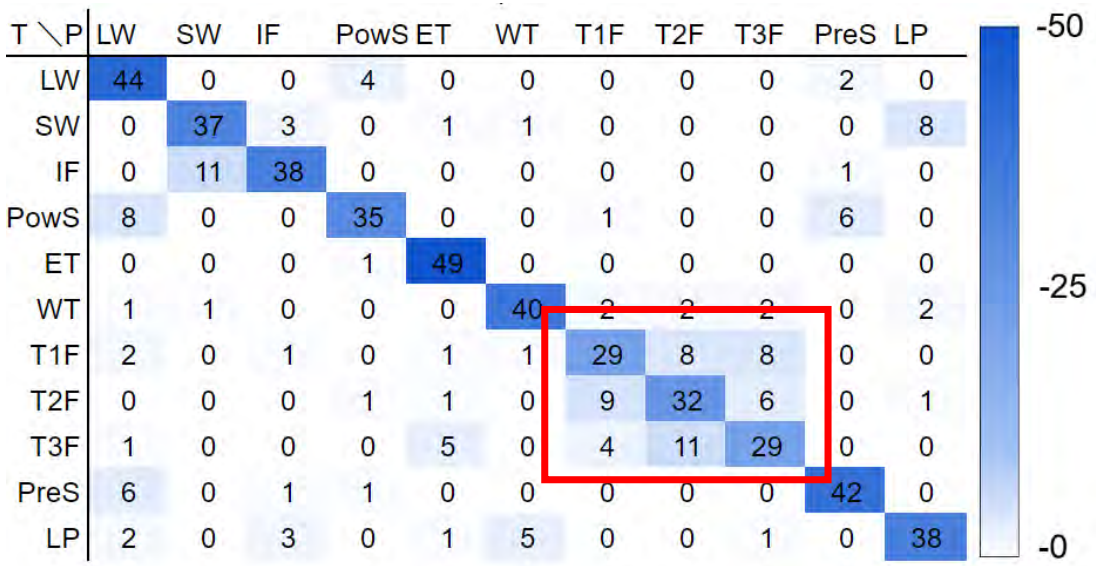
Architecture (Input)	Angle Error		Inference
	Individ.	General	Time (ms)
VGG16 (RGB)	16.07	18.19	54
ResN18 (RGB)	16.11	18.70	17
ResN18+LSTM (RGB)	11.95	14.01	35
ResN18+LSTM (Motion)	9.29	10.69	33
ResN18+LSTM (TS)	9.28	10.13	40
Ours (w/o Data Aug.)	9.35	11.11	40
Ours (TS)	**8.81**	**9.77**	41

Table 4. Results of ablation study of different network architecture and input data. The metrics of Angle Error is MAE (degree), TS stands for two-stream input, 'Ours' stands for ResN18+LSTM+KF (TS).

Study: Grasp Recognition



Different Grasp Types



T \ P	LW	SW	IF	PowS	ET	WT	T1F	T2F	T3F	PreS	LP
LW	44	0	0	4	0	0	0	0	0	2	0
SW	0	37	3	0	1	1	0	0	0	0	8
IF	0	11	38	0	0	0	0	0	0	1	0
PowS	8	0	0	35	0	0	1	0	0	6	0
ET	0	0	0	1	49	0	0	0	0	0	0
WT	1	1	0	0	0	40	2	2	2	0	2
T1F	2	0	1	0	1	1	29	8	8	0	0
T2F	0	0	0	1	1	0	9	32	6	0	1
T3F	1	0	0	0	5	0	4	11	29	0	0
PreS	6	0	1	1	0	0	0	0	0	42	0
LP	2	0	3	0	1	5	0	0	1	0	38

Confusion Matrix of 50 test result

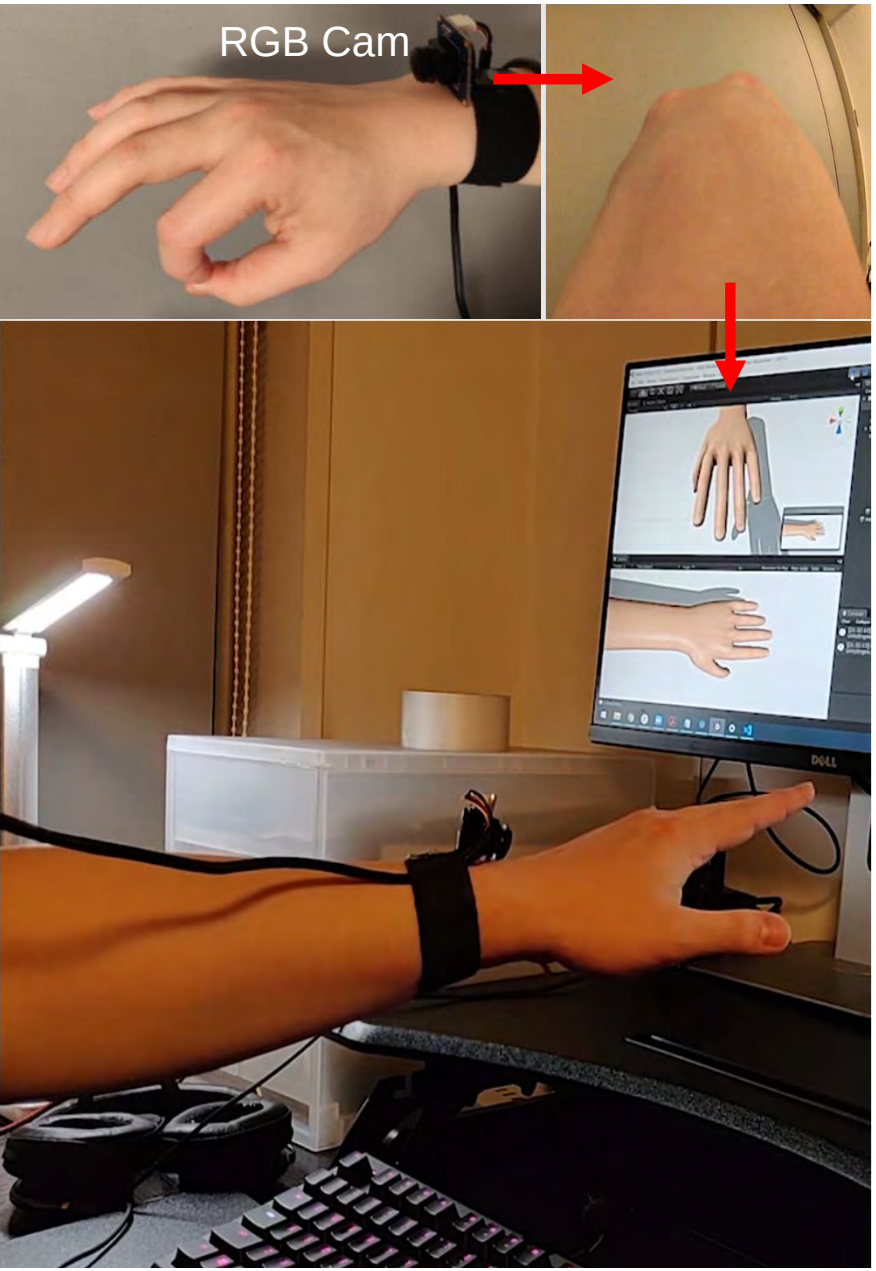
Application

Summary

- Proposed a novel DorsalNet to reconstruct angle-based hand pose from images of the back of the hand.
- Evaluations on 3D hand pose by comparing with several baselines and an ablation study.
- Evaluations on gesture and grasp recognition.
- Demonstrated several types of applications to show use-cases.

Future Work

- Implementation for full wrist actuation & rotation
- Solutions for completely dark environment
- More various subjects to study the generalization
- Embedded in real watches



THANK YOU!

CONTACT: wu.e.aa@m.titech.ac.jp