Supplementary Material of "InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image"

Gyeongsik Moon[1], Shoou-I Yu[2], He Wen[2], Takaaki Shiratori[2], and Kyoung Mu Lee[1]

[1] ECE & ASRI, Seoul National University, Korea
[2] Facebook Reality Labs
{mks0601,kyoungmu}@snu.ac.kr, {shoou-i.yu,hewen,tshiratori}@fb.com

In this supplementary material, we present more experimental results that could not be included in the main manuscript due to the lack of space.

1 Comparison of human and machine annotation

To show how human and machine annotations are different, we visualize t-SNE of human and machine annotation in Figure 1. Each vector in t-SNE is a 20-dimensional hand pose vector. For this, we pre-defined 20 degrees of freedom for each hand. Two degrees of freedom are defined for each finger root (*i.e.*, T1, I1, M1, R1, and P1) as pitch and yaw angles. The other degree of freedom is defined for T2, T3, I2, I3, M2, M3, R2, R3, P2, and P3 as pitch angle. To calculate the angles, we assume I1, M1, R1, P1, and wrist joints are on the same plane s. As the figure shows, the machine-generated annotations have broader hand pose coverage than human-generated ones.

2 Effect of number of the available views

Our InterHand2.6M dataset has a large number of views. To show how the available number of views affect the 3D hand pose estimation accuracy, we report MPJPE of a model trained only from four widely used views (*i.e.*, top, frontal, right, and left views) in Table 1. For the fair comparison, we increased the number of iterations when using four views for training to make the total number of iterations in the training stage the same. Also, the same RootNet [1] trained on all views is used for both settings. As the table shows, our large number of views improves the performance significantly. This comparison shows a large number of views in our InterHand2.6M dataset is beneficial.

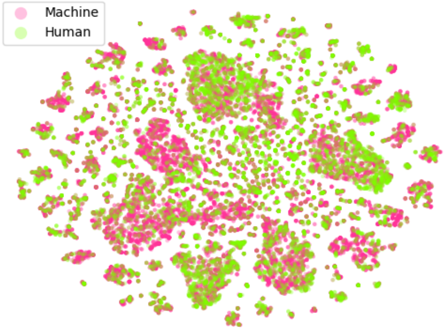


Fig. 1: Visualized t-SNE of human and machine-generated 3D hand pose annotation.

num. of training views	AP_h	MRRPE	MPJPE
4	0.95	61.90	26.80
all (ours)	**0.99**	**32.57**	**14.22**

Table 1: AP_h, MRRPE, and MPJPE on Test (H+M) of InterHand2.6M dataset using different number of training views.

3 InterHand2.6M sequence descriptions

We provide detailed descriptions and visualizations of the sequences in the proposed InterHand2.6M dataset.

Single hand sequences. Figure 2, 3, and 4 show PP of single hand sequences. Below are detailed descriptions of each sequence.

- neutral relaxed: the neutral hand pose. Hands in front of the chest, fingers do not touch, and palms face the side.
- neutral rigid: the neutral hand pose with maximally extended fingers, muscles tense.
- good luck: hand sign language with crossed index and middle fingers.
- fake gun: hand gesture mimicking the gun.
- star trek: hand gesture popularized by the television series Star Trek.
- star trek extended thumb: “star trek” with extended thumb.
- thumb up relaxed: hand sign language that means “good”, hand muscles relaxed.
- thumb up normal: “thumb up”, hand muscles average tenseness.
- thumb up rigid: “thumb up”, hand muscles very tense.
- thumb tuck normal: similar to fist, but the thumb is hidden by other fingers.
- thumb tuck rigid: “thumb tuck”, hand muscles very tense.
- aokay: hand sign language that means “okay”, where palm faces the side.
- aokay upright: “aokay” where palm faces the front.
- surfer: the SHAKA sign.
- rocker: hand gesture that represents rock and roll, where palm faces the side.
- rocker front: the “rocker” where palm faces the front.

- rocker back: the “rocker” where palm faces the back.
- fist: fist hand pose.
- fist rigid: fist with very tense hand muscles.
- alligator closed: hand gesture mimicking the alligator with a closed mouth.
- one count: hand sign language that represents “one.”
- two count: hand sign language that represents “two.”
- three count: hand sign language that represents “three.”
- four count: hand sign language that represents “four.”
- five count: hand sign language that represents “five.”
- indextip: thumb and index fingertip are touching.
- middletip: thumb and middle fingertip are touching.
- ringtip: thumb and ring fingertip are touching.
- pinkytip: thumb and pinky fingertip are touching.
- palm up: has palm facing up.
- finger spread relaxed: spread all fingers, hand muscles relaxed.
- finger spread normal: spread all fingers, hand muscles average tenseness.
- finger spread rigid: spread all fingers, hand muscles very tense.
- capisce: hand sign language that represents “got it” in Italian.
- claws: hand pose mimicking claws of animals.
- peacock: hand pose mimicking peacock.
- cup: hand pose mimicking a cup.
- shakespeareyorick: hand pose from Yorick from Shakespeare’s play Hamlet.
- dinosaur: hand pose mimicking a dinosaur.
- middle finger: hand sign language that has an offensive meaning.

Figure 5 shows ROM of single hand sequences. Below are detailed descriptions of each sequence.

- five count: count from one to five.
- five countdown: count from five to one.
- fingertip touch: thumb touch each fingertip.
- relaxed wave: wrist relaxed, fingertips facing down and relaxed, wave.
- fist wave: rotate wrist while hand in a fist shape.
- prom wave: wave with fingers together.
- palm down wave: wave hand with the palm facing down.
- index finger wave: hand gesture that represents “no” sign.
- palmer wave: palm down, scoop towards you, like petting an animal.
- snap: snap middle finger and thumb.
- finger wave: palm down, move fingers like playing the piano.
- finger walk: mimicking a walking person by index and middle finger.
- cash money: rub thumb on the index and middle fingertips.
- snap all: snap each finger on the thumb.

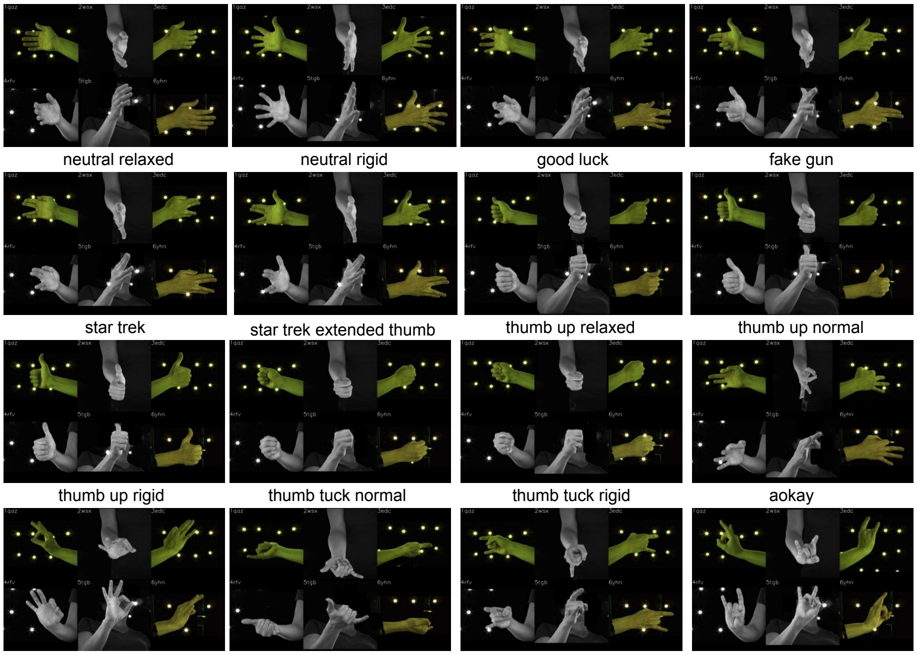
Interacting hand sequences. Figure 6 shows PP of interacting hand sequences. Below are detailed descriptions of each sequence.

- right clasp left: two hands clasp each other, right hand is on top of the left hand.

- left clasp right: two hands clasp each other, left hand is on top of the right hand.
- fire gun: mimicking a gun with two hands together.
- right fist cover left: right fist completely covers the left hand.
- left fist cover right: left fist completely covers the right hand.
- interlocked fingers: fingers of the right and left hands are interlocked.
- pray: hand sign language that represents praying.
- right fist over left: right fist is on top of the left fist.
- left fist over right: left fist is on top of the right fist.
- right babybird: mimicking caring a babybird with two hands, the right hand is placed at the bottom.
- left babybird: mimicking caring a babybird with two hands, the left hand is placed at the bottom.
- interlocked finger spread: fingers of the right and left hands are interlocked yet spread.
- finger squeeze: squeeze all five fingers with the other hand.

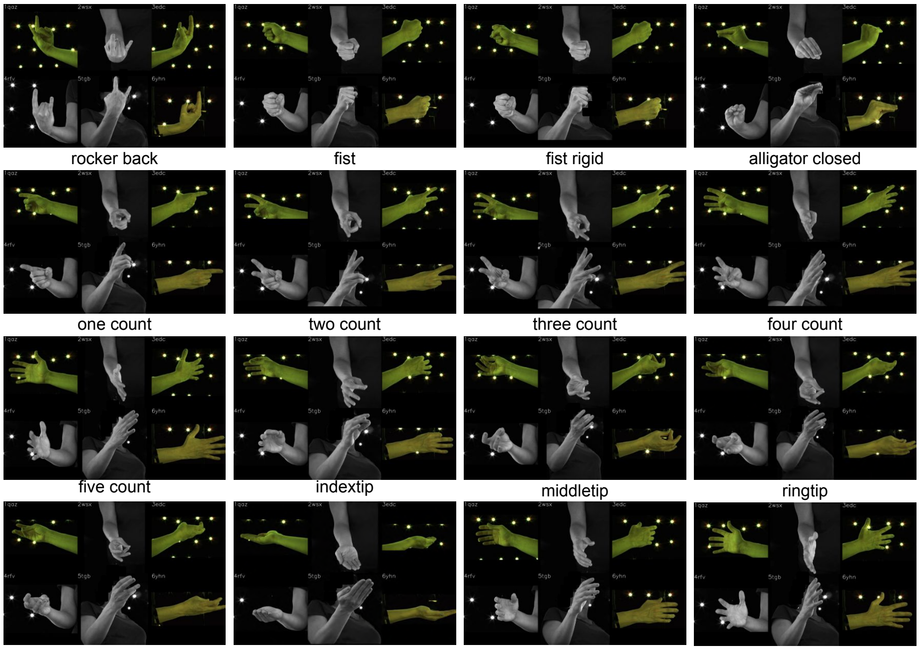
Finally, Figure 7 shows ROM of interacting hand sequences. Below are detailed descriptions of each sequence.

- palmerrub: rub palm of one hand with opposite hand's thumb.
- knuckle crack: crack each finger by having the opposite hand compress a bent finger.
- golf claplor: light clap, left over right.
- itsy bitsy spider: finger motion used when singing the children song "itsy bitsy spider", like this (link).
- finger noodle: fingers interlocked, palms facing opposite directions, wiggle middle fingers.
- nontouch: two hands random motion, hands do not touch.
- sarcastic clap: exaggerated, slow clap.
- golf claprol: light clap, right over left.
- evil thinker: wrist together, tap fingers together one at a time.
- rock paper scissors: hold rock, then paper, then scissors.
- hand scratch: using the opposite hand, lightly scratch palm then top of hand; switch and do the same with the other hand.
- touch: two hands interacting randomly in a frame, touching.
- pointing towards features: using the opposite index finger, point out features on the palm and back of the hand (trace lifelines/wrinkles, etc.).
- interlocked thumb tiddle: interlock fingers, rotate thumbs around each other.
- right finger count index point: using the right pointer finger, count up to five on the left hand, starting with the pinky.
- left finger count index point: using left pointer finger, count up to five on the right hand, starting with the pinky.
- single relaxed finger: this consists of a series of actions: (1) touch each fingertip to the center of the palm for the same hand, do this for both hands, (2) interlock fingers and press palms out, (3) with the opposite hand, hold wrist, (4) with the opposite hand, bend wrist down and back, (5) point at watch on both wrists, (6) circle wrists, (7) look at nails, and (8) point at yourself with thumbs then with index fingers.



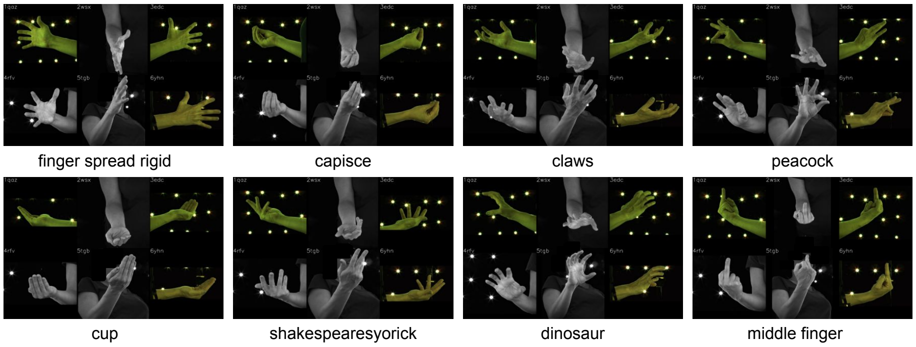
neutral relaxed
neutral rigid
good luck
fake gun
star trek
star trek extended thumb
thumb up relaxed
thumb up normal
thumb up rigid
thumb tuck normal
thumb tuck rigid
aokay
aokay upright
surfer
rocker
rocker front

Fig. 2: Visualization of the single hand PP sequences.



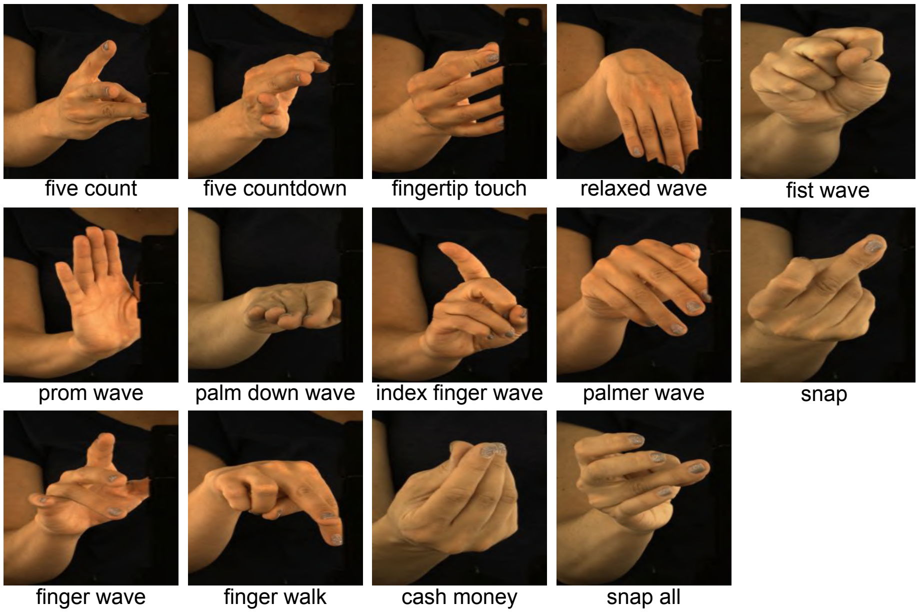
rocker back
fist
fist rigid
alligator closed
one count
two count
three count
four count
five count
indextip
middletip
ringtip
pinkytip
palm up
finger spread relaxed
finger spread normal

Fig. 3: Visualization of the single hand PP sequences.



finger spread rigid, capisce, claws, peacock, cup, shakespearesyorick, dinosaur, middle finger

Fig. 4: Visualization of the single hand PP sequences.



five count, five countdown, fingertip touch, relaxed wave, fist wave, prom wave, palm down wave, index finger wave, palmer wave, snap, finger wave, finger walk, cash money, snap all

Fig. 5: Visualization of the single hand ROM sequences.

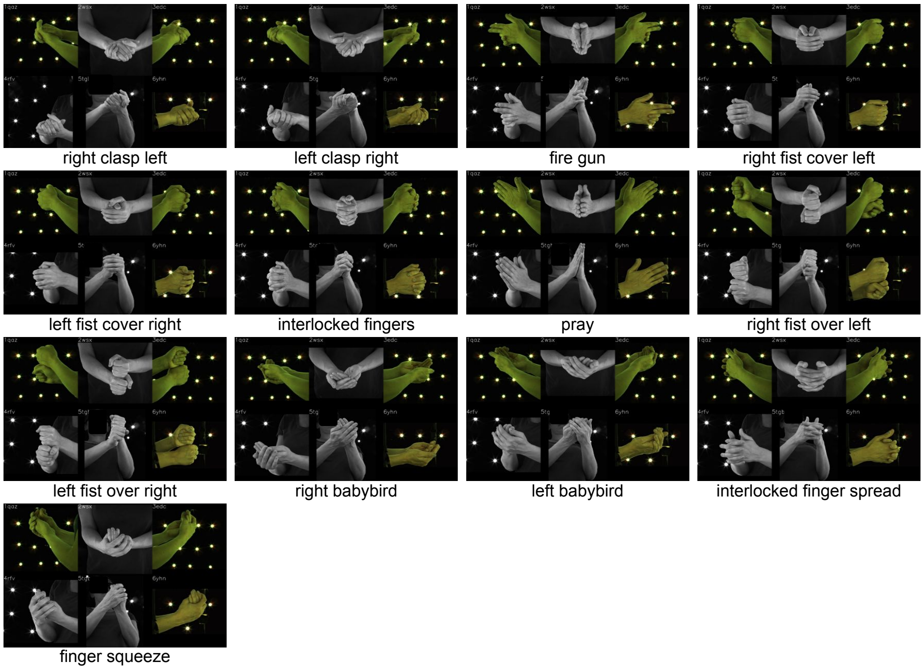

right clasp left
left clasp right
fire gun
right fist cover left
left fist cover right
interlocked fingers
pray
right fist over left
left fist over right
right babybird
left babybird
interlocked finger spread
finger squeeze

Fig. 6: Visualization of the interacting hand PP sequences.

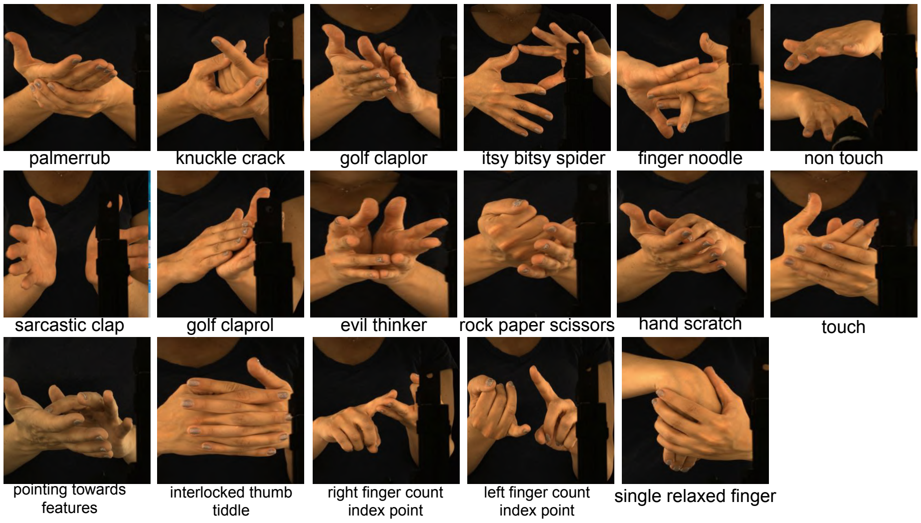

palmerrub
knuckle crack
golf claplor
itsy bitsy spider
finger noodle
non touch
sarcastic clap
golf claprol
evil thinker
rock paper scissors
hand scratch
touch
pointing towards features
interlocked thumb tiddle
right finger count index point
left finger count index point
single relaxed finger

Fig. 7: Visualization of the interacting hand ROM sequences.

4 InterHand2.6M human annotation procedure

Figure 8 shows human annotation procedure of InterHand2.6M. In the left figure, an annotator clicks hand joint positions at the easiest view (red circle). Then, the annotator clicks the positions of the same hand joints at another view (red circle). Our human annotation tool automatically triangulates human annotations from two views in the 3D space and projects the 3D point to remaining views (green circles).

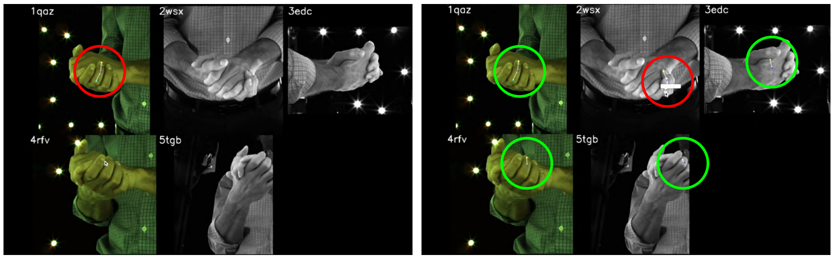


Fig. 8: The human annotation procedure of InterHand2.6M.

5 InterHand2.6M capture studio environment

Figure 9 shows a rendering of our constructed multi-view studio for the data capture.

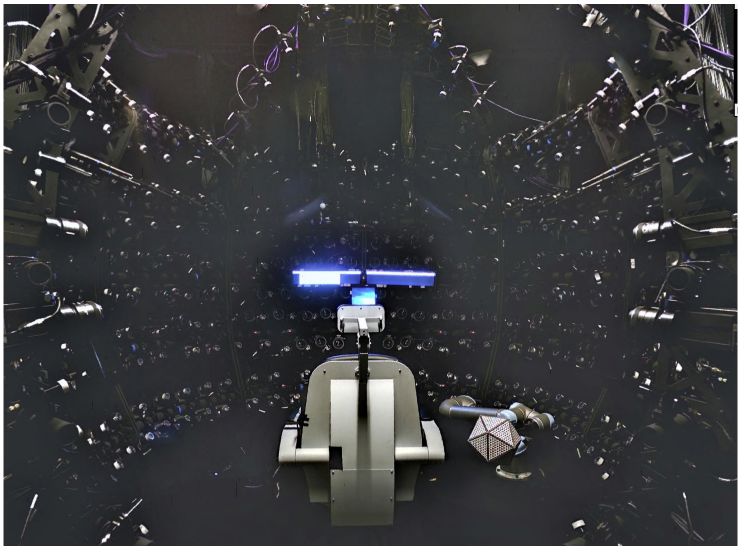


Fig. 9: Rendering of our constructed multi-view studio.

6 Evaluation on various test set using various training set

We provide more experimental results on various test set of the InterHand2.6M (*i.e.*, Val (M), Test (H), Test (M), and Test (H+M)) by training InterNet on various training set (*i.e.*, Train (H), Train (M), and Train (H+M)). Table 2 and Table 3 show AP_h and MRRPE on all various testing sets from models trained on different set. Table 4, 5, 6, and 7 show MPJPE on Val (M), Test (H), Test (M), and Test (H+M), respectively.

training set	Val (M)	Test (H)	Test (M)	Test (H+M)
Train (H)	**99.10**	**99.79**	98.95	99.01
Train (M)	97.79	99.07	98.85	98.87
Train (H+M)	98.14	99.77	**99.03**	**99.09**

Table 2: AP_h comparison from models trained with different training set.

training set	Val (M)	Test (H)	Test (M)	Test (H+M)
Train (H)	40.06	21.80	38.50	36.80
Train (M)	40.50	23.21	38.84	37.25
Train (H+M)	**35.72**	**20.26**	**33.97**	**32.57**

Table 3: MRRPE comparison from models trained with different training set.

training set	T4	T3	T2	T1	I4	I3	I2	I1	M4	M3	M2	M1	R4	R3	R2	R1	P4	P3	P2	P1	avg.
results on single hand sequences																					
Train (H)	18.5	14.5	11.1	7.7	20.6	17.7	15.3	12.0	22.9	19.7	16.7	12.1	21.5	18.0	**15.0**	11.2	**19.6**	**16.7**	**14.4**	**10.3**	15.02
Train (M)	18.2	14.5	11.1	**7.2**	20.8	18.0	15.9	12.3	23.0	20.4	17.5	12.3	21.7	18.8	15.7	11.4	20.5	17.7	15.1	10.5	15.36
Train (H+M)	**17.6**	**14.0**	**10.7**	**7.2**	**19.7**	**17.2**	**15.1**	**11.7**	**21.5**	**19.0**	**16.4**	**11.8**	**20.4**	**17.8**	**15.0**	**10.9**	**19.6**	17.1	14.7	**10.3**	**14.65**
results on interacting hand sequences																					
Train (H)	25.9	18.8	15.3	10.6	28.8	24.3	20.7	15.8	29.7	24.2	21.3	15.8	26.5	22.0	19.3	14.9	25.2	21.2	18.8	14.4	19.70
Train (M)	25.9	18.8	15.4	10.1	32.2	26.1	21.6	15.8	29.7	24.7	21.6	15.9	27.0	22.3	19.5	14.9	26.0	21.7	19.1	14.3	20.13
Train (H+M)	**23.8**	**17.5**	**14.2**	**9.7**	**28.0**	**23.4**	**19.6**	**14.5**	**27.8**	**22.9**	**20.1**	**14.6**	**24.9**	**20.8**	**18.2**	**13.9**	**24.2**	**20.4**	**18.0**	**13.5**	**18.58**

Table 4: MPJPE of our InterNet on the Val (M) of InterHand2.6M dataset using various training set.

training set	T4	T3	T2	T1	I4	I3	I2	I1	M4	M3	M2	M1	R4	R3	R2	R1	P4	P3	P2	P1	avg.
results on single hand sequences																					
Train (H)	13.4	11.1	9.3	7.7	13.1	11.8	10.8	9.1	13.9	11.7	10.7	9.1	14.0	11.2	10.2	9.0	13.2	11.1	10.0	8.5	10.42
Train (M)	13.1	11.1	9.2	7.4	13.5	12.3	11.3	9.3	14.2	12.4	11.2	9.4	13.9	11.5	10.6	9.4	13.1	11.2	10.4	8.8	10.64
Train (H+M)	**12.1**	**10.4**	**8.9**	**7.3**	**12.1**	**11.2**	**10.2**	**8.6**	**13.0**	**11.2**	**10.2**	**8.7**	**12.8**	**10.6**	**9.7**	**8.7**	**12.4**	**10.6**	**9.8**	**8.2**	**9.85**
results on interacting hand sequences																					
Train (H)	18.4	14.5	11.7	10.0	17.2	14.7	13.3	11.7	17.1	14.6	13.0	11.3	16.4	14.1	12.4	10.7	15.7	13.7	12.5	11.0	13.05
Train (M)	20.2	15.7	12.3	10.2	19.7	16.6	14.7	12.4	19.4	16.3	14.2	12.0	18.3	15.4	13.5	11.5	17.0	14.9	13.4	11.7	14.26
Train (H+M)	**17.1**	**13.6**	**11.0**	**9.6**	**16.1**	**13.8**	**12.5**	**11.0**	**16.0**	**13.8**	**12.2**	**10.7**	**15.3**	**13.2**	**11.7**	**10.2**	**14.9**	**13.1**	**11.8**	**10.4**	**12.29**

Table 5: MPJPE of our InterNet on the Test (H) of InterHand2.6M dataset using various training set.

training set	T4	T3	T2	T1	I4	I3	I2	I1	M4	M3	M2	M1	R4	R3	R2	R1	P4	P3	P2	P1	avg.
results on single hand sequences																					
Train (H)	16.6	12.8	9.7	6.9	17.2	14.7	13.0	11.0	18.6	15.8	13.7	10.6	17.7	14.7	**12.9**	10.0	16.2	**13.6**	**12.0**	9.7	12.74
Train (M)	**15.7**	**12.1**	**9.3**	**6.5**	16.6	14.5	13.1	10.8	18.0	15.7	13.8	10.4	17.4	14.8	13.1	10.1	16.2	13.8	12.2	9.6	12.56
Train (H+M)	**15.7**	**12.1**	**9.3**	6.6	**16.1**	**14.2**	**12.8**	**10.6**	**17.4**	**15.3**	**13.5**	**10.3**	**16.9**	**14.5**	**12.9**	**9.8**	**15.8**	**13.6**	**12.0**	**9.4**	**12.32**
results on interacting hand sequences																					
Train (H)	23.3	17.8	13.8	9.5	25.9	21.8	18.8	15.0	29.5	22.3	18.7	14.2	25.8	19.8	17.5	13.7	23.5	18.9	16.7	13.6	18.10
Train (M)	23.9	18.0	13.8	9.0	28.8	23.3	19.5	15.2	29.4	22.9	19.1	14.5	25.6	20.4	17.9	14.0	24.3	19.6	17.3	13.9	18.59
Train (H+M)	**21.5**	**16.4**	**12.6**	**8.5**	**24.4**	**20.5**	**17.4**	**13.7**	**27.7**	**20.9**	**17.5**	**13.1**	**24.0**	**18.6**	**16.4**	**12.7**	**22.3**	**17.9**	**15.8**	**12.6**	**16.88**

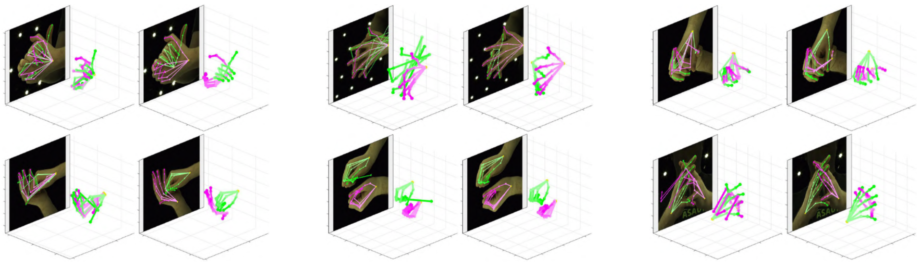
Table 6: MPJPE of our InterNet on the Test (M) of InterHand2.6M dataset using various training set.

training set	T4	T3	T2	T1	I4	I3	I2	I1	M4	M3	M2	M1	R4	R3	R2	R1	P4	P3	P2	P1	avg.
results on single hand sequences																					
Train (H)	16.4	12.7	9.7	7.0	16.9	14.5	12.9	10.8	18.3	15.5	13.5	10.5	17.5	14.5	**12.7**	9.9	16.0	**13.4**	**11.9**	9.6	12.58
Train (M)	15.5	**12.0**	**9.3**	**6.5**	16.4	14.4	13.0	10.7	17.7	15.5	13.7	10.3	17.2	14.6	12.9	10.1	16.0	13.7	12.1	9.6	12.43
Train (H+M)	**15.4**	**12.0**	**9.3**	6.7	**15.8**	**14.0**	**12.6**	**10.4**	**17.1**	**15.0**	**13.3**	**10.2**	**16.6**	**14.3**	**12.7**	**9.7**	**15.6**	**13.4**	**11.9**	**9.4**	**12.16**
results on interacting hand sequences																					
Train (H)	22.4	17.1	13.4	9.7	24.4	20.5	17.8	14.4	27.1	20.9	17.7	13.7	23.8	18.8	16.6	13.1	22.0	18.0	15.9	13.1	17.16
Train (M)	23.2	17.6	13.6	9.3	27.2	22.1	18.6	14.7	27.5	21.7	18.2	14.0	24.0	19.5	17.1	13.5	22.9	18.8	16.5	13.5	17.79
Train (H+M)	**20.7**	**15.9**	**12.3**	**8.8**	**23.0**	**19.3**	**16.6**	**13.2**	**25.4**	**19.5**	**16.5**	**12.6**	**22.1**	**17.6**	**15.5**	**12.2**	**20.9**	**17.1**	**15.0**	**12.2**	**16.02**

Table 7: MPJPE of our InterNet on the Test (H+M) of InterHand2.6M dataset using various training set.

7 Qualitative results

We compare the qualitative results of our InterNet trained on (a) only single hand data and (b) both single and interacting hand data in Figure 10. As the figure shows, when InterNet is trained only on single hand data, it provides a reasonable 3D hand pose when the input image contained separated two hands (*i.e.*, bottom middle example). However, it totally fails for all interacting hand sequences. We provide more qualitative results and failure cases of our InterNet on Test (H+M) of the proposed InterHand2.6M dataset in Figure 11. As the figure shows, severe occlusions make 2.5D hand pose estimation and right hand-relative left hand depth estimation fail.



(a) trained on SH (b) trained on SH+IH (a) trained on SH (b) trained on SH+IH (a) trained on SH (b) trained on SH+IH

Fig. 10: Qualitative results comparison of our InterNet trained on (a) only single hand data and (b) both single and interacting hand data from the proposed InterHand2.6M dataset.

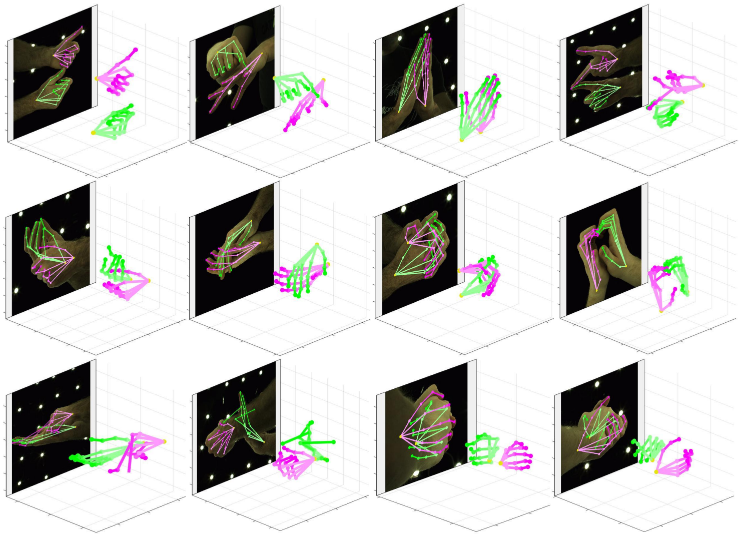


Fig. 11: Qualitative results (top two rows) and failure case (last row) of our InterNet on the proposed InterHand2.6M dataset.

References

1. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: ICCV (2019)