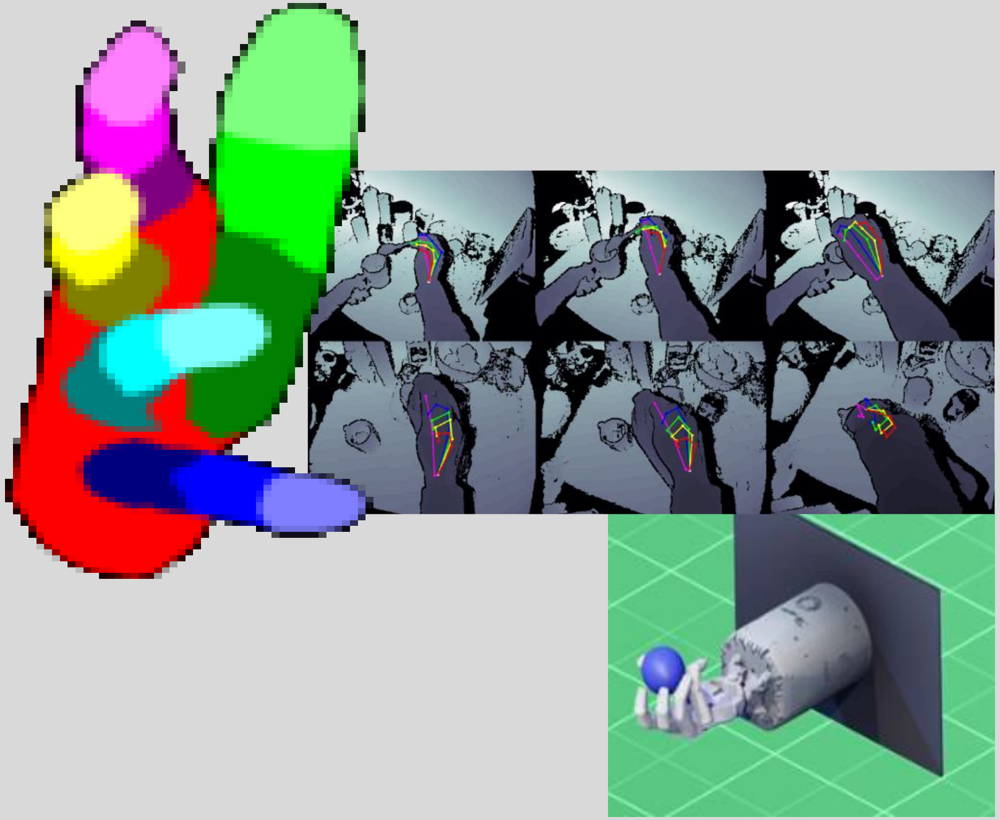
Data Augmentation for 3D Computer Vision



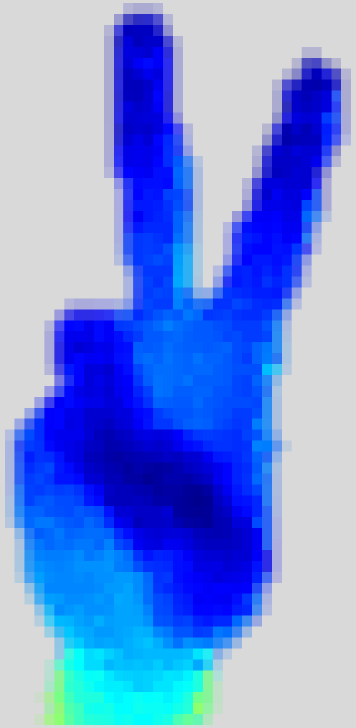
Tae-Kyun (T-K) Kim
Computer Vision and Learning Lab

KAIST Imperial College London

https://labicvl.github.io/
https://sites.google.com/view/tkkim/

3D Pose Estimation (hand or body)

Input Depth Image



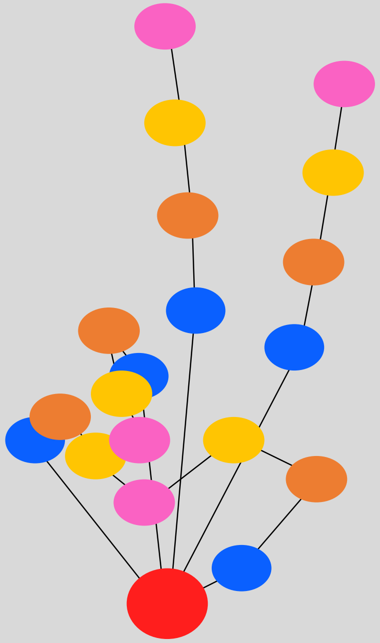
Z

Extract joint angles

$\theta \in \mathbb{R}^d$

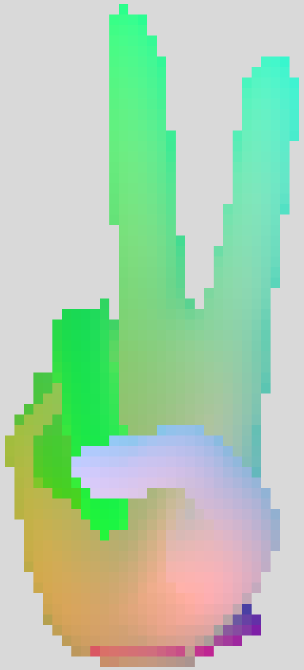
for current frame

Skeleton



θ

Rendered depth



R_θ

Challenges:

- High degree of freedom (d=26)
- Viewpoint changes and self occlusions
- Fast movement
- Annotation difficulty
- Shape variation

Dense Pose Estimation

- HANDS19 Challenge @ ICCV includes: Hand-object interaction, depth and colour modalities, extrapolation capabilities, the use of synthetic data (MANO).
- Fitted mesh models to BigHand2.2M, F-PHAB, HO-3D datasets, are provided.

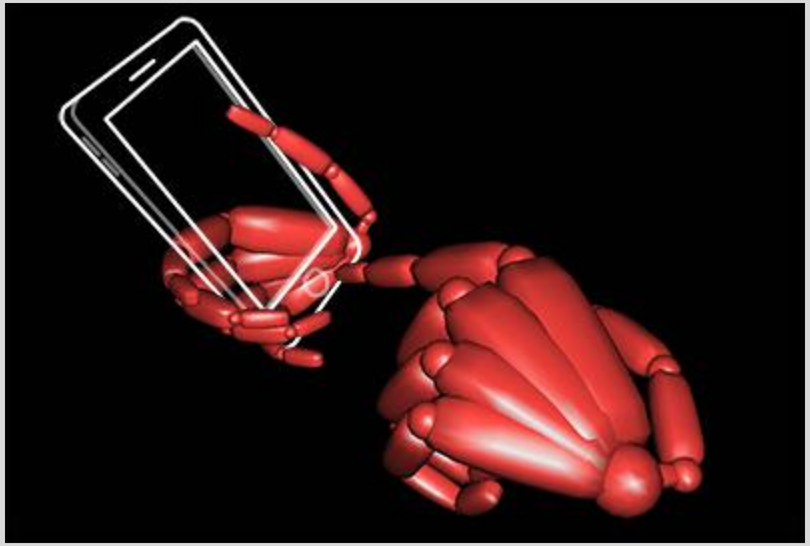
Interaction with AR/VR environment

[Oculus]

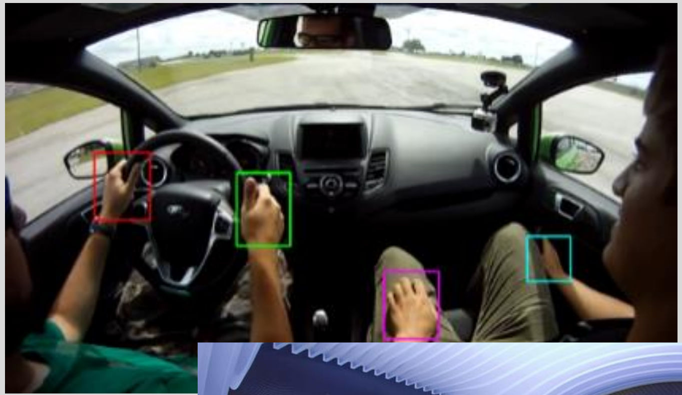
[Upload VR]

[Leap Motion]

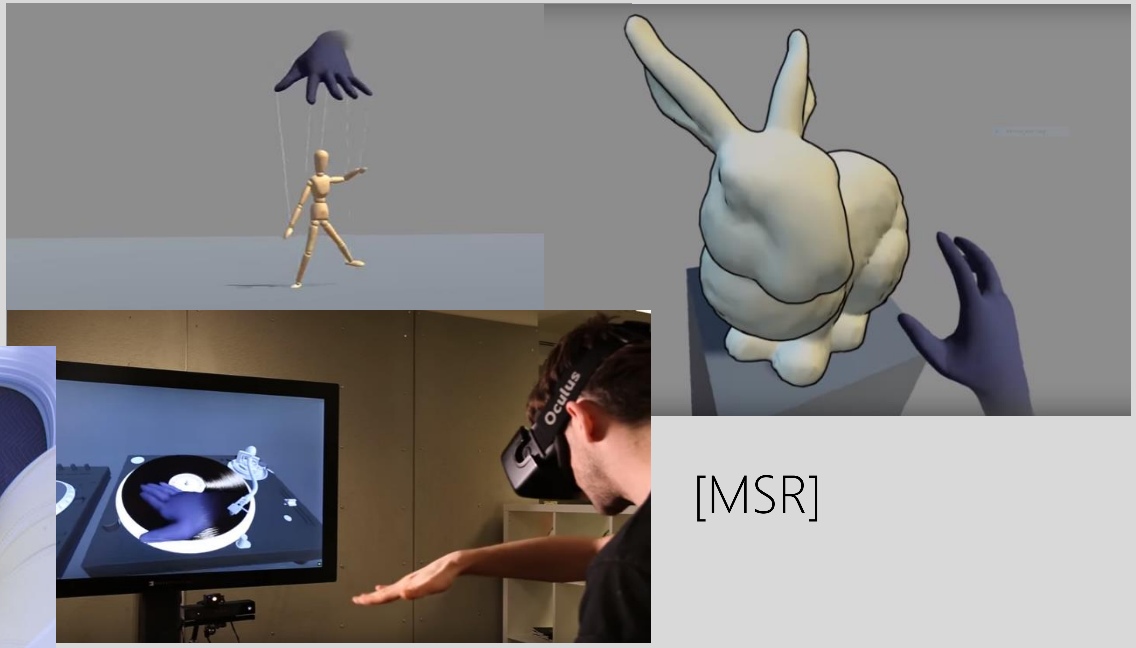
[NANSENSE]



More examples: AR/VR in autonomous cars



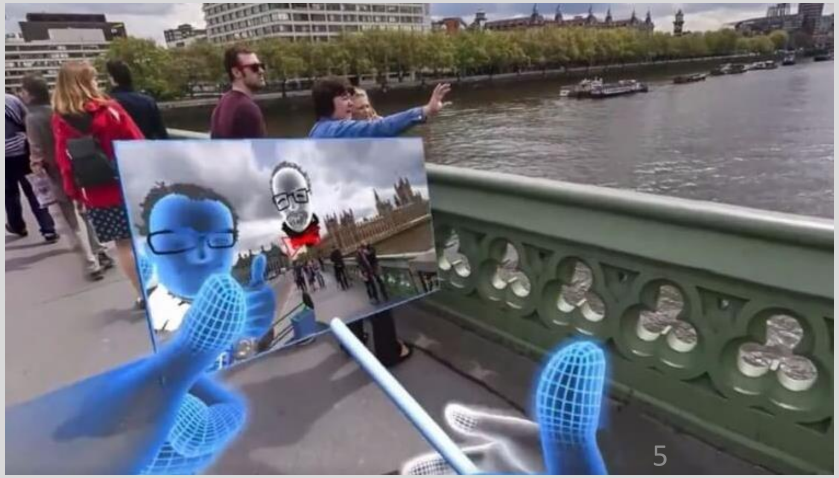
[UCSD]



[MSR]

[Upload VR]

Driver-vehicle interaction

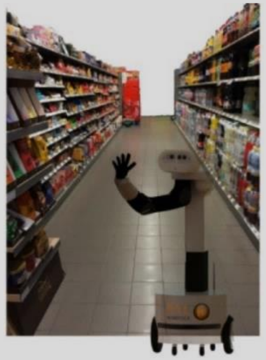


[Oculus]

6D Object Pose and Active Vision

Problem: Estimating objects' 3D location and pose

Application: E.g. picking and placing for logistics



Our methods deliver state-of-the-art performance:

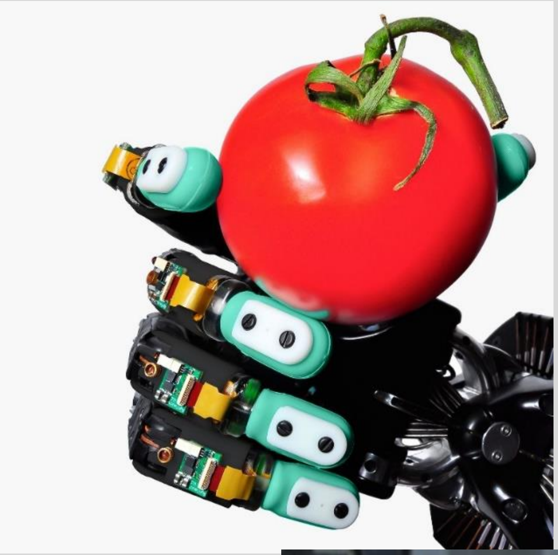
- Autonomous unfolding clothes (ICRA14, best paper award): regression RF, probabilistic active planning
- Latent Hough Forest (ECCV14) : template-matching splitting, one-class learning
- Active Forest (ECCV14) : multi-task learning, next-best view in RF
- Object pose in the crowd (CVPR16) : deep f eatures, next-best-view

Sponsored by:

OMRON

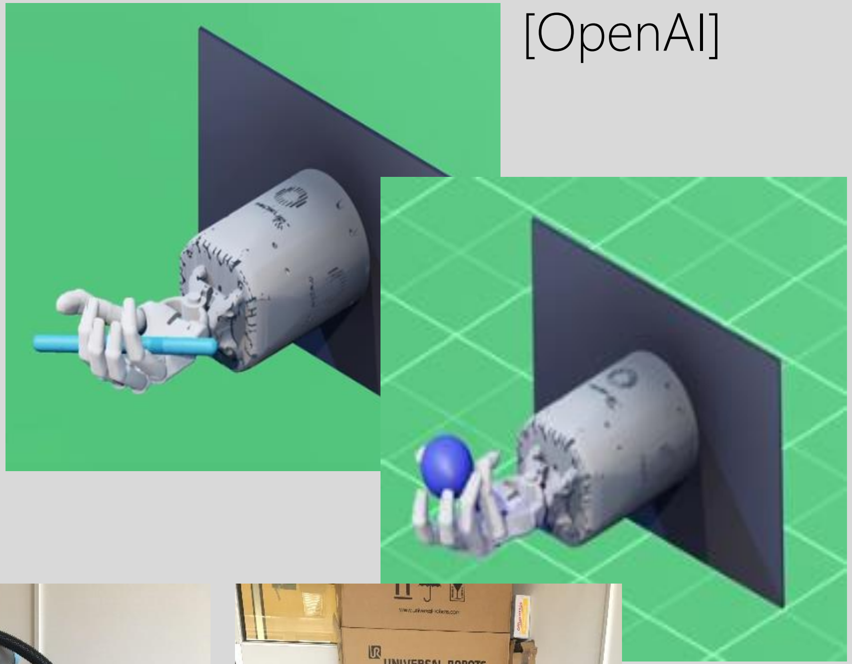
amazon

Physical interactions and robotics



[SynTouch]

[Spread]

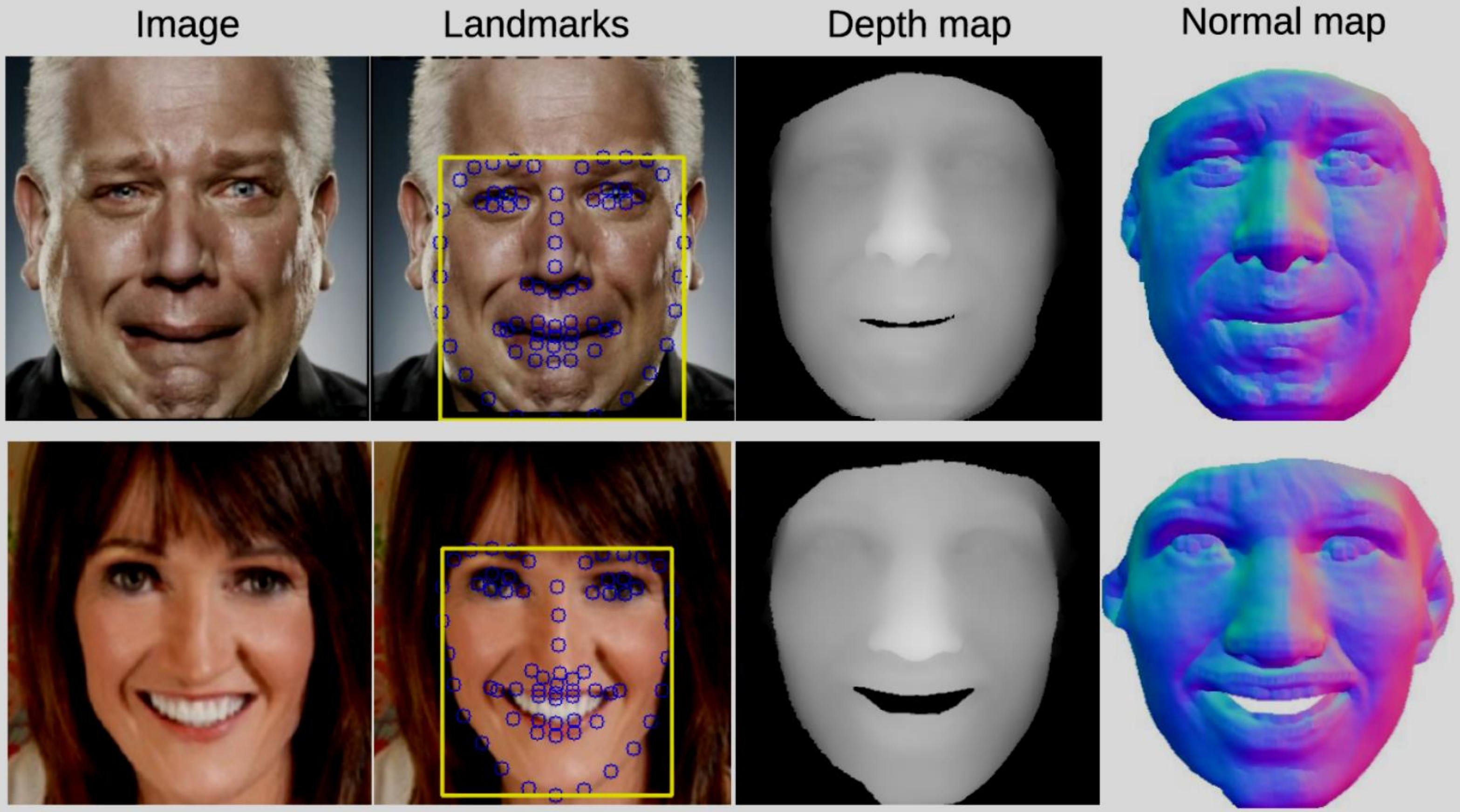


[OpenAI]

Robot-human interaction

[ICL & Samsung Research]

3D Facial Landmarking and Image Generation



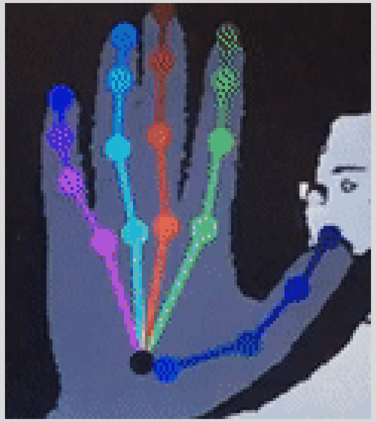
Progressive GANS (Karras et al, ICLR18)

Challenges

Challenges

- A training dataset that spans all data variations is hard to obtain:

[Viewpoint]



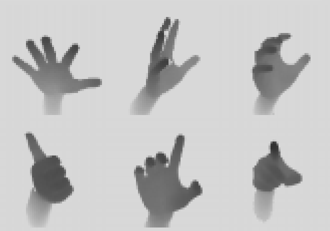
[Shape]

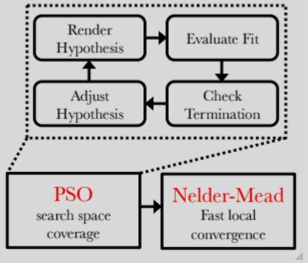
[Articulation]

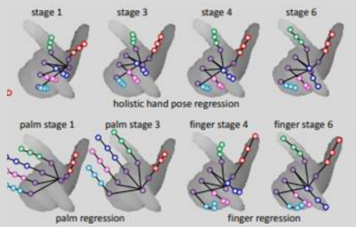
- The data space needs to be densely covered:
  - Depth images change a lot by slight hand pose variation due to self-occlusions etc.

Real vs synthetic data collection

Real Data



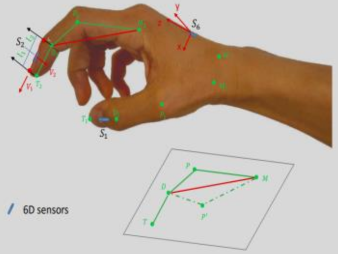


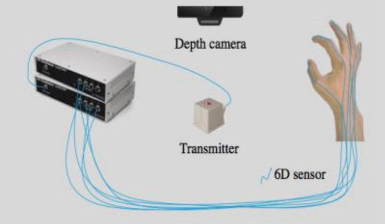


ICVL, NYU, MSRA: Use of tracking & refinement.

[Tang *et al*. CVPR'14, Tompson *et al*. TOG'14, Sun *et al*. CVPR'15]

→ Still, lacking in combination of viewpoint/shape/articulation.

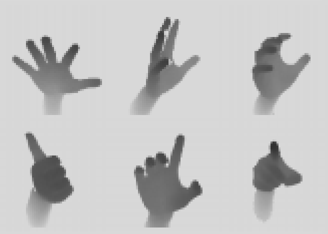


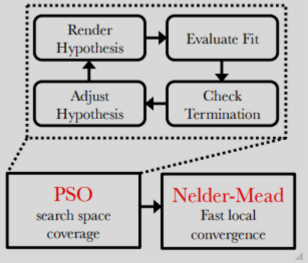


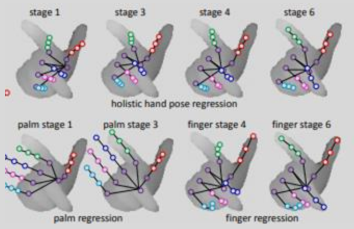
Big Hand 2.2M: Use of sensors/inverse kinematics. [Yuan *et al*. CVPR'17]

Real vs synthetic data collection

Real Data



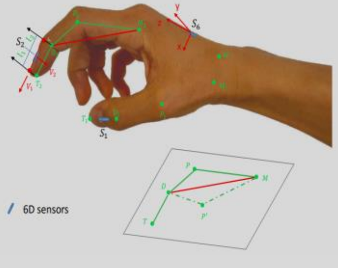


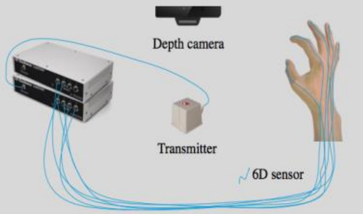


ICVL, NYU, MSRA: Use of tracking & refinement.

[Tang *et al.* CVPR'14, Tompson *et al.* TOG'14, Sun *et al.* CVPR'15]

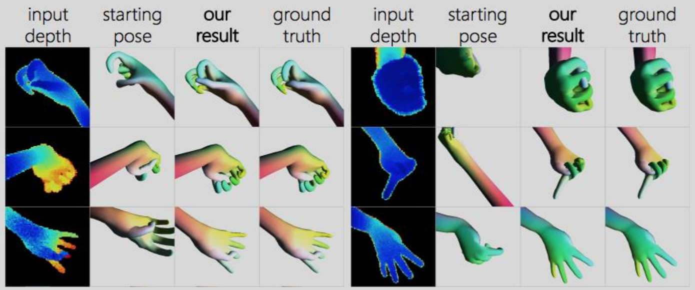
→ Still, lacking in combination of viewpoint/shape/articulation.





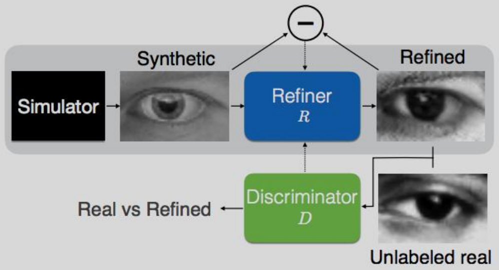
Big Hand 2.2M: Use of sensors/inverse kinematics. [Yuan *et al.* CVPR'17]

Synt. Data



Data collection by using a synthetic 3D model.

[Sharp *et al.* CHI'15]



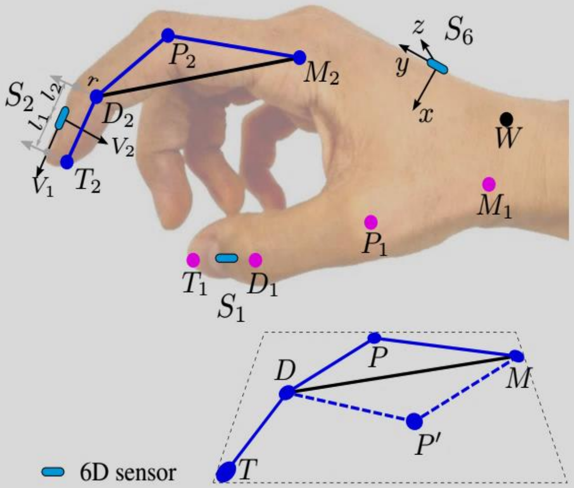
Reduce the gap between synthetic and real data.

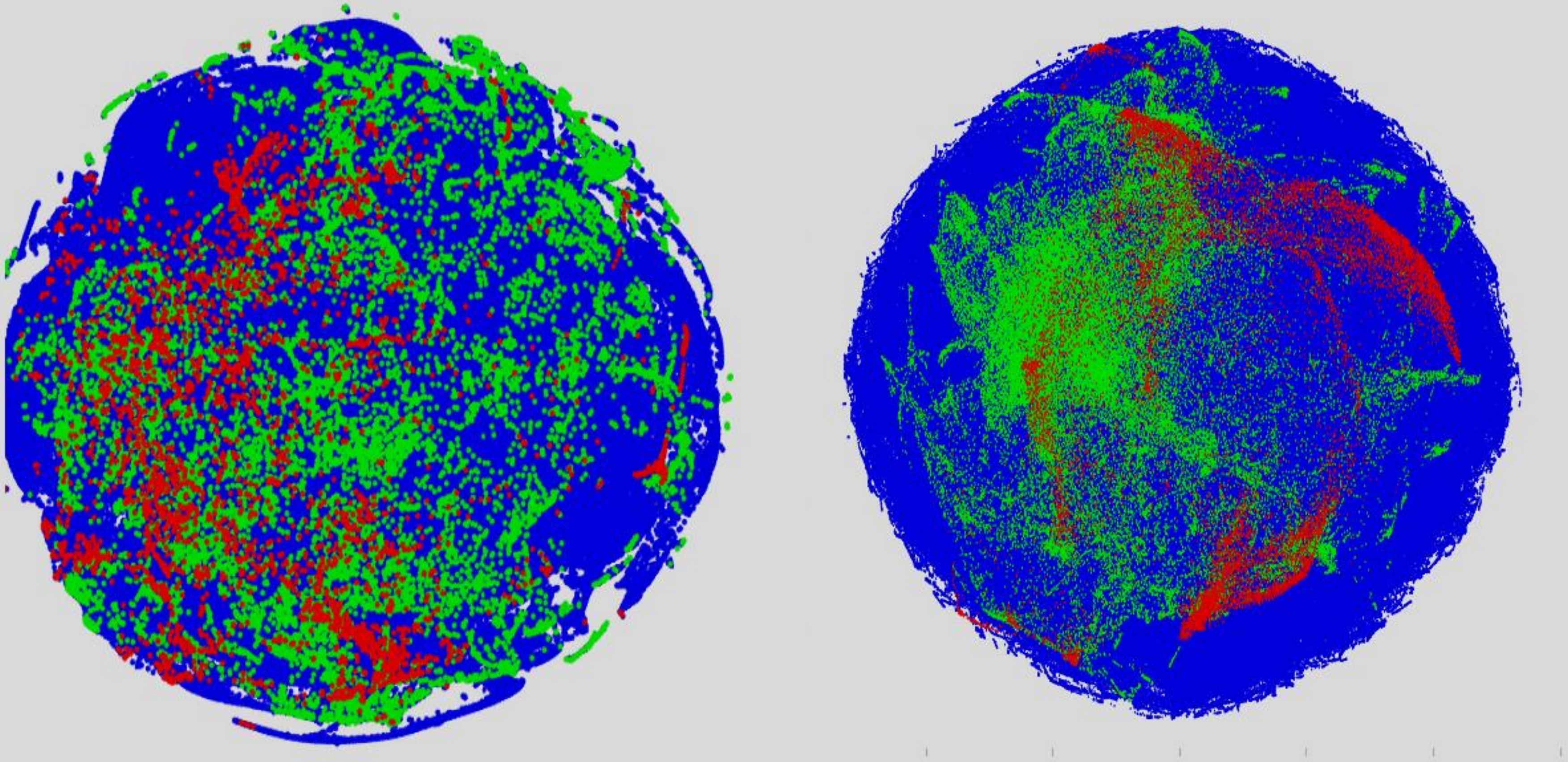
[Shrivastava *et al.* CVPR'17]

→ Synthetic-real data discrepancy

→ No interaction between the data generator and hand pose estimator.

BigHand2.2M benchmark CVPR17 [used by 116 unique institutions, 491 downloads]



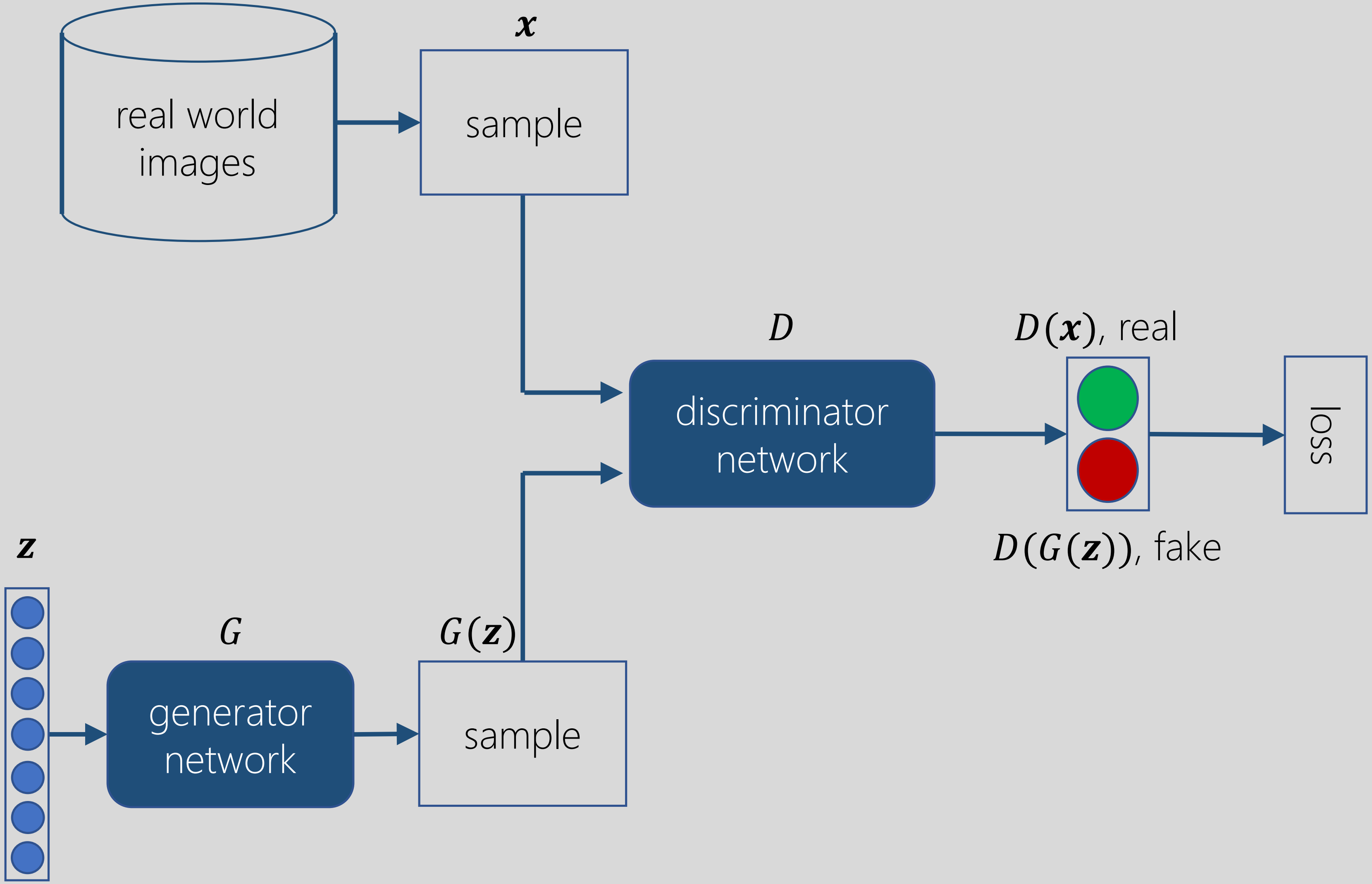


t-SNE embedding. *BigHand2.2M* (blue), ICVL (red), and NYU (green).
global view point (left), articulation space in 25D (right)

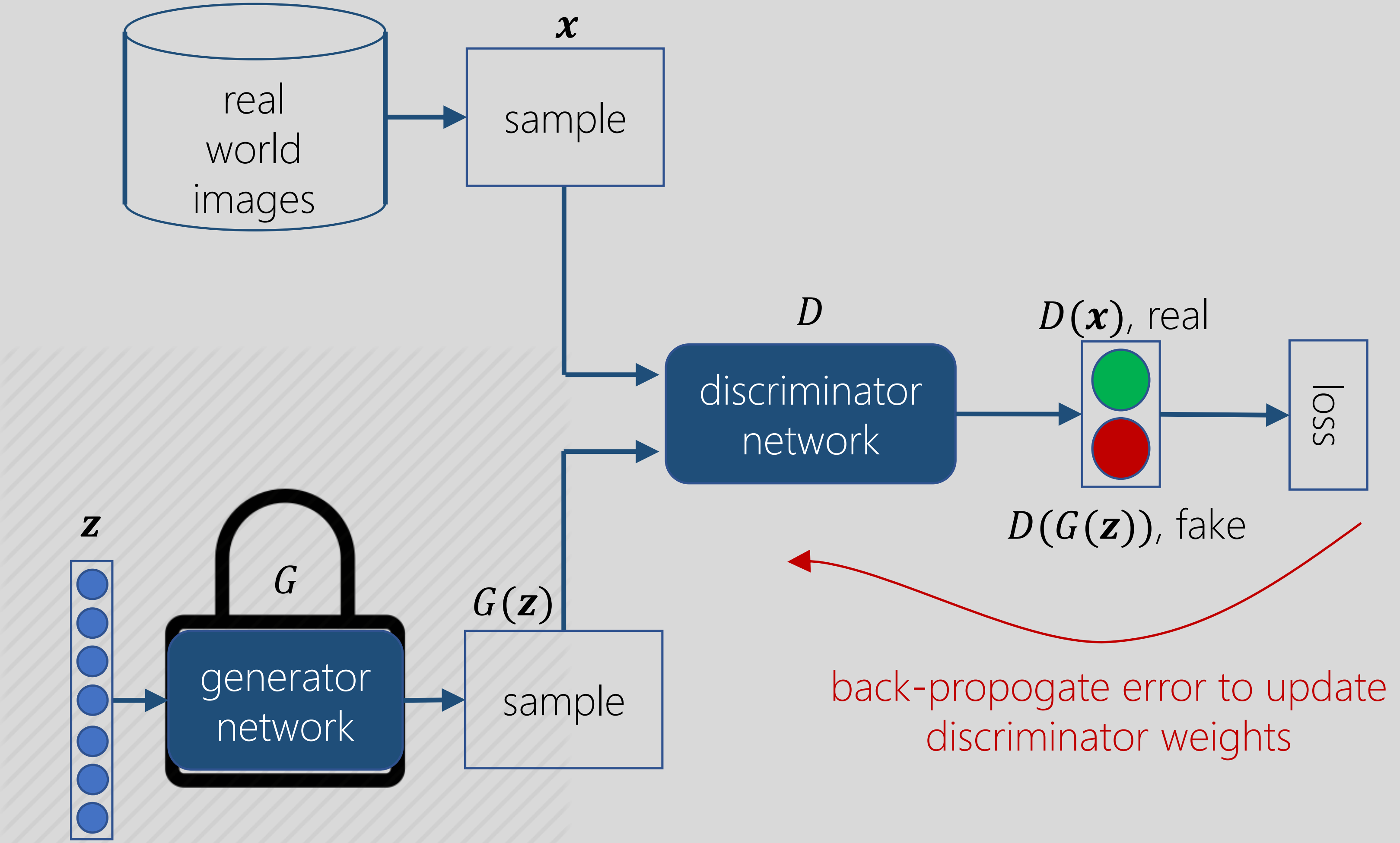
Data augmentation

GAN Architecture

- GAN composes of two networks: the **generator** and the **discriminator**.

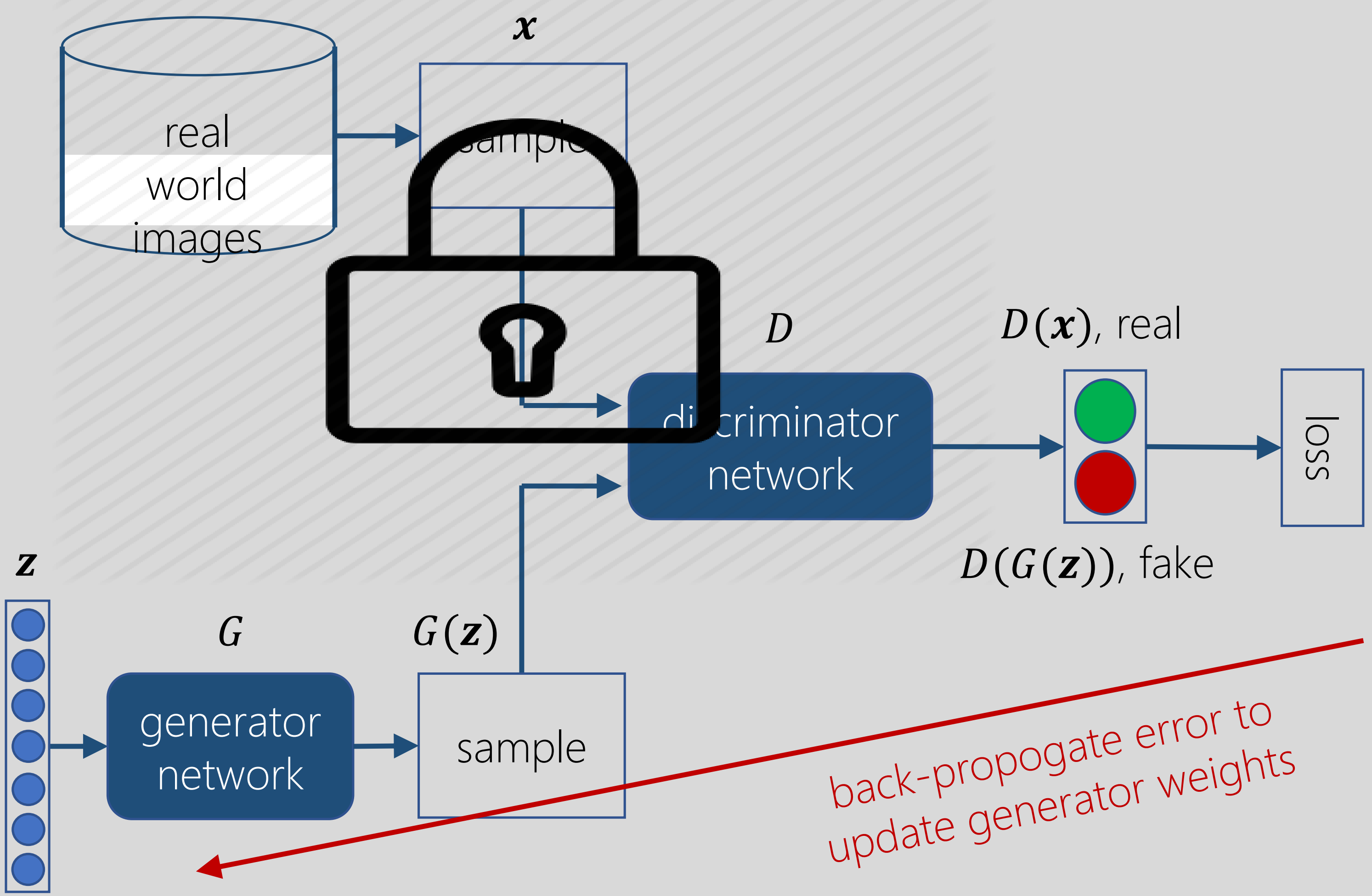


GAN Training (Discriminator)



The training process of the "Discriminator Network". Error is back-propagated over the discriminator network only, in order to update discriminator weights, while the Generator Network is locked.

GAN Training (Generator)



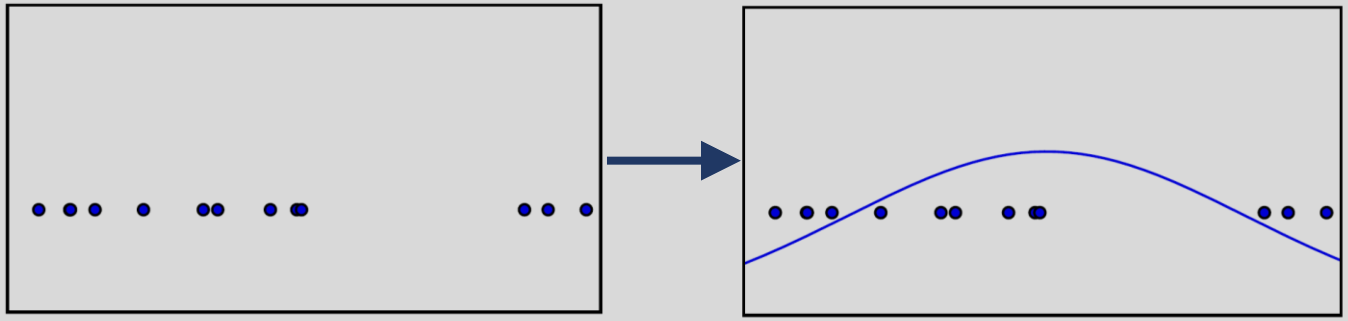
The training process of the "Generator Network". Error is back-propagated over the generator network only, in order to update generator weights, while the Discriminator Network is locked.

Generative Models

- Generative Adversarial Networks (GANs) are example of generative models.

- Generative models take a training set (samples drawn from a data-generating distribution p_{data}), and learn to represent an estimate of that distribution. The result is a probability distribution p_{model}.

- In some cases, the model estimates p_{model} explicitly. Generative model performing density estimation takes training data, which are of an unknown data-generating distribution p_{data}, and return an estimate of that distribution. The estimate p_{model} can be evaluated for a particular value of $\boldsymbol{x}$ to obtain an estimate $p_{model}(\boldsymbol{x})$ of the true density $p_{model}(\boldsymbol{x})$:



Generative Models

- In other cases, the model is only able to generate samples from p_{model}. Some generative models are able to generate samples from the model distribution p_{model}. Ideally, a generative model would be able to train on examples (left), and then create more examples from the same distribution (right):

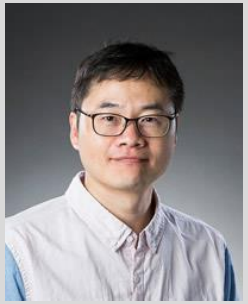
training examples

model samples

Weakly-supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects

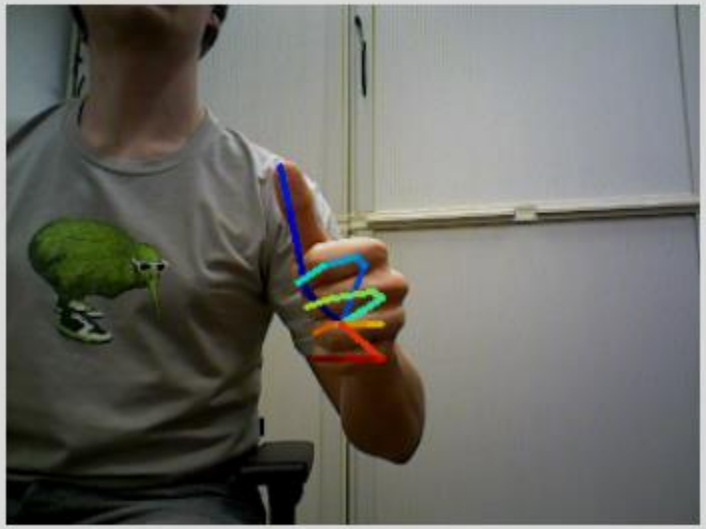
(CVPR20 oral, best paper finalist)


Seungryul Baek
Imperial College London


Kwang In Kim
UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

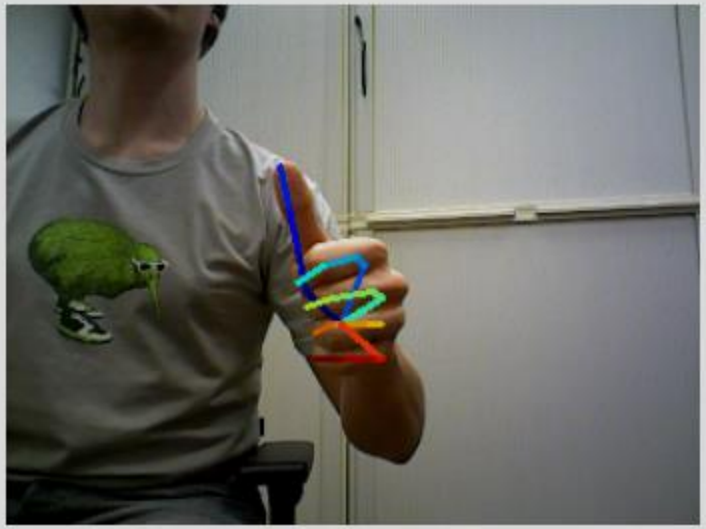

Tae-Kyun Kim
Imperial College London

Objective

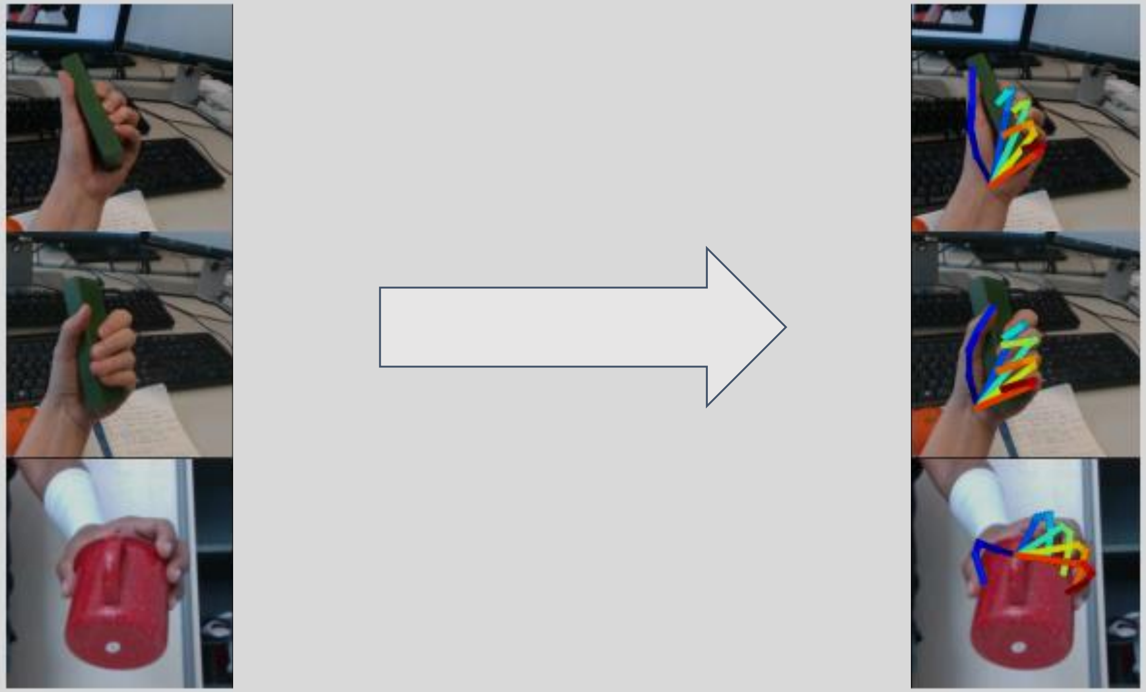


Hand pose estimation for Hand-only scenario.

Objective



Hand pose estimation for Hand-only scenario.



Hand pose estimation from single RGB images under hand object interaction (HOI) scenario.

Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV'16)

EgoDexter (ICCV'17)

[**Real** dataset – Few in quantity, inaccurate/insufficient 3D annotation]

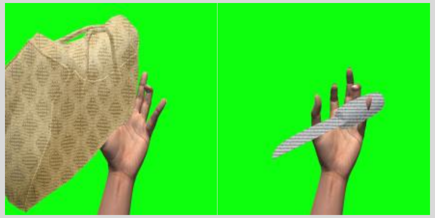
Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV’16)

Obman (CVPR’19)

EgoDexter (ICCV’17)



SynthHands (ICCV’17)

[**Real** dataset]

[**Synthetic** dataset – gap to real dataset]

Related works

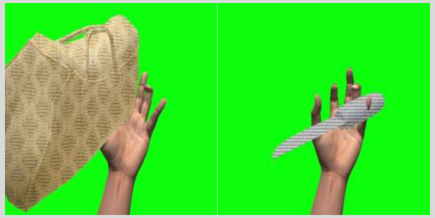
Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV'16)

Obman (CVPR'19)

FPHA (CVPR'18)

EgoDexter (ICCV'17)



SynthHands (ICCV'17)

GANerated (CVPR'18)

[**Real** dataset]

[**Synthetic** dataset]

[Using **GAN/Sensors** – Still limited]

Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

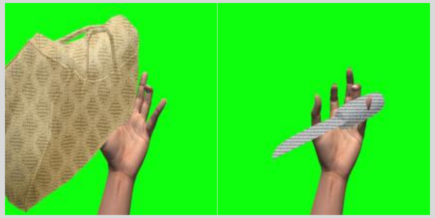
Dexter+Object (ECCV'16)

Obman (CVPR'19)

FPHA (CVPR'18)

HO3D (ArXiv'19)

EgoDexter (ICCV'17)



SynthHands (ICCV'17)

GANerated (CVPR'18)

FreiHand (ICCV'19)

[**Real** dataset]

[**Synthetic** dataset]

[Using **GAN/Sensors**]

[Iterative **3D model** fitting – #sample]

Challenges

STB (ICIP'17)

RHD (ICCV'17)

[Real and synthetic Hand-only data]

[Diverse objects]

Challenges

STB (ICIP'17)

RHD (ICCV'17)

[Real and synthetic Hand-only data]

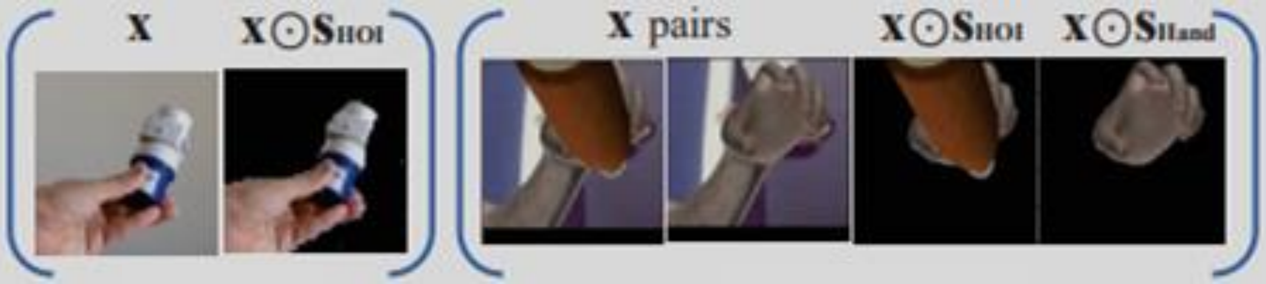
[Diverse objects]

HO3D (ArXiv'19)
Real, <15000 frame, 6 objects.

FreiHand (ICCV'19)
Real, <3000 frame, <30 objects.

Key Idea

Image-level supervision with **HOI images**:

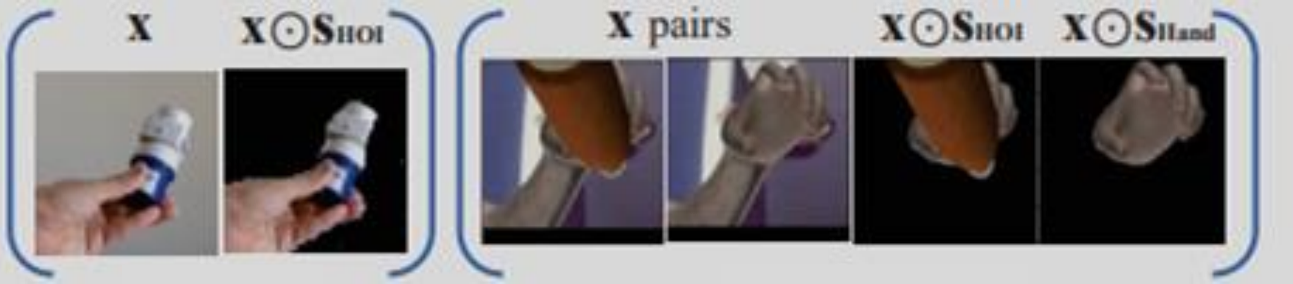


Real HOI Synthetic HOI

We exploit only easily available **Real and synthetic hand-only data**, **Real HOI images with segmentation masks**, **Synthetic hand-only and HOI image pairs.**

Key Idea

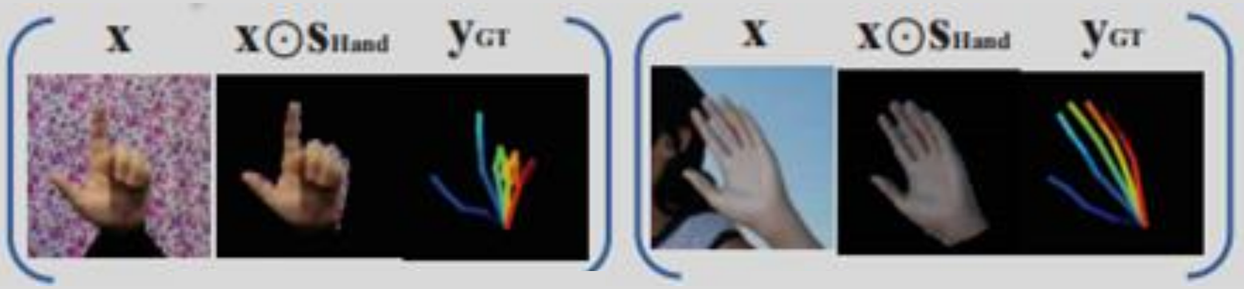
Image-level supervision with **HOI images**:



Real HOI

Synthetic HOI

3D supervision with **Hand-only data**:



Real Hand-only

Synthetic Hand-only

We exploit only easily available **Real and synthetic hand-only data**, **Real HOI images with segmentation masks**, **Synthetic hand-only and HOI image pairs.**

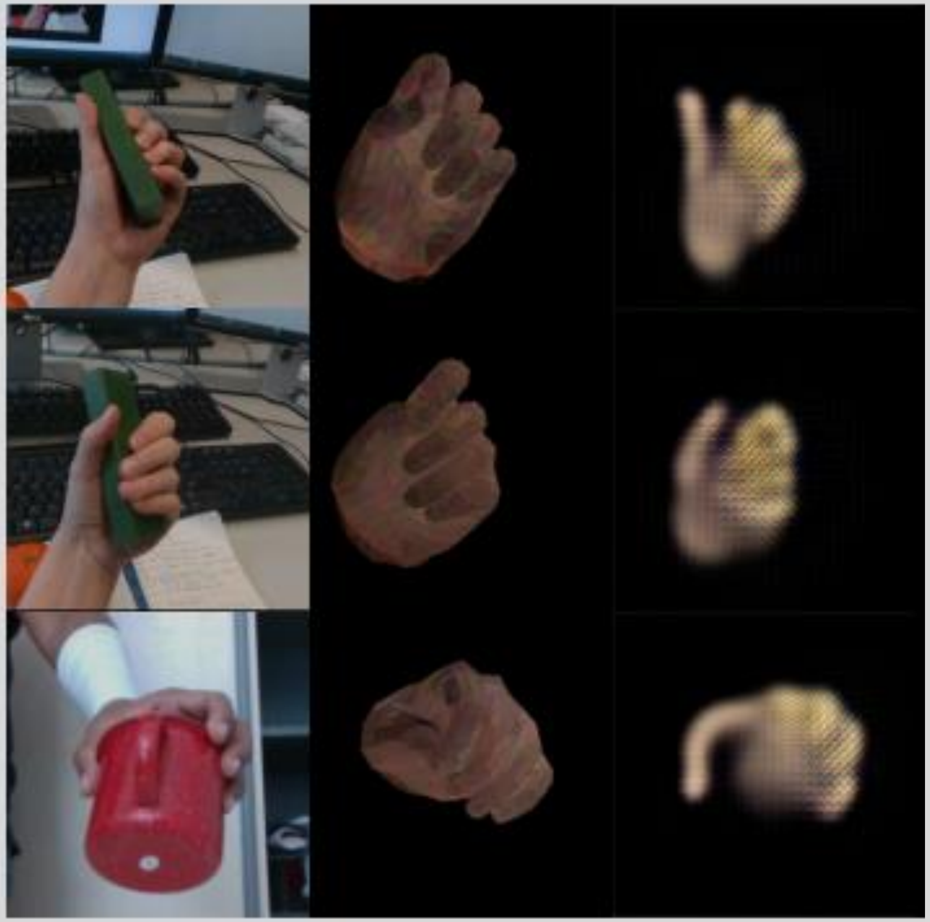
Key Idea

We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea

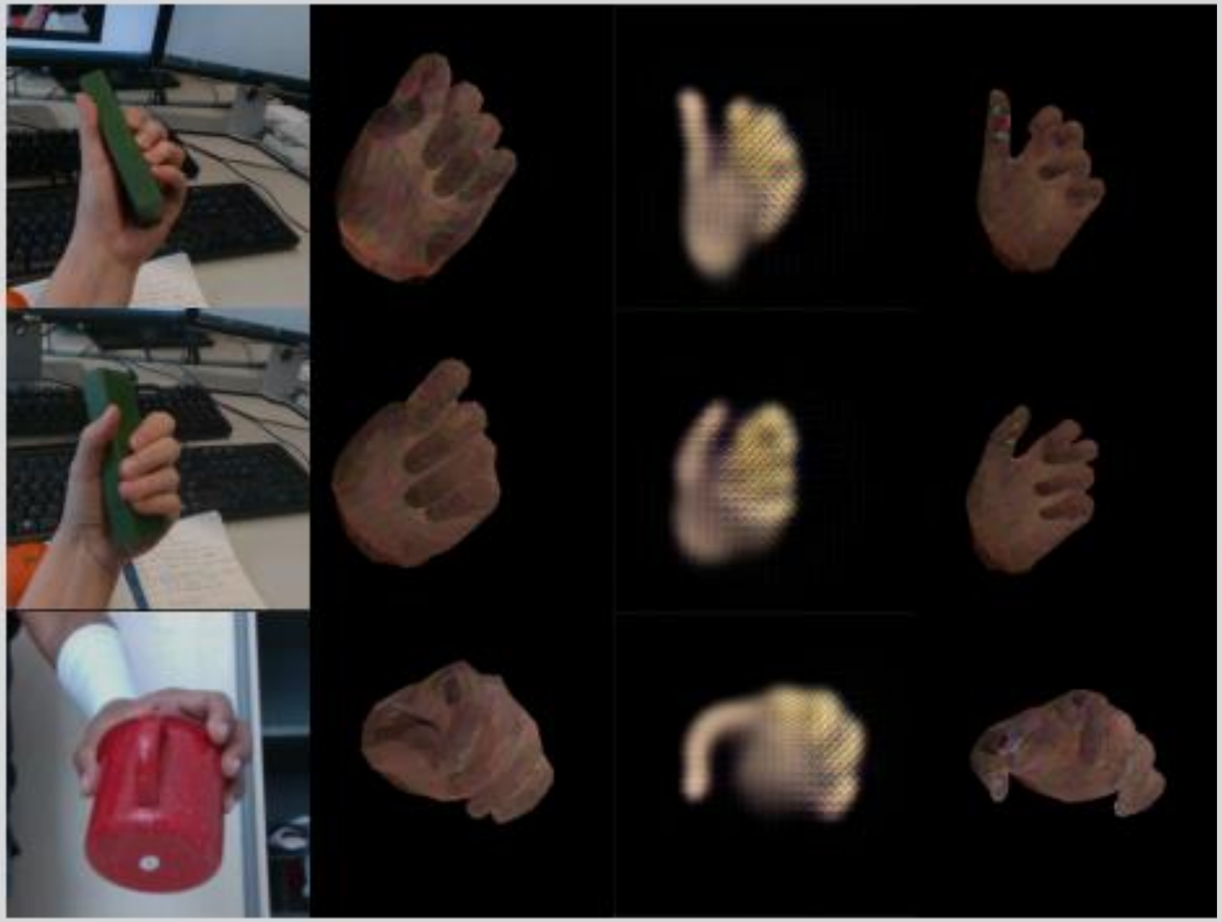
We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea



We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea

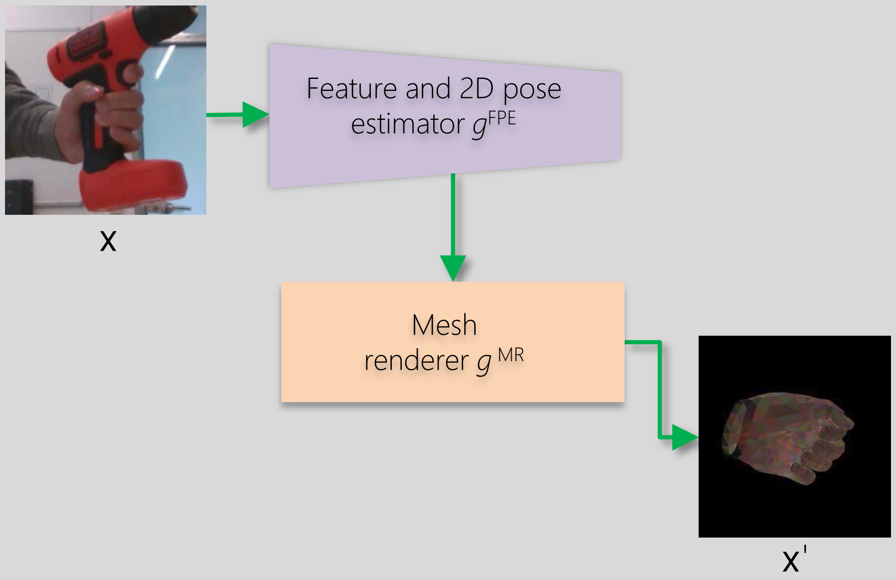


Then, we learn the hand mesh estimation using the translated images.

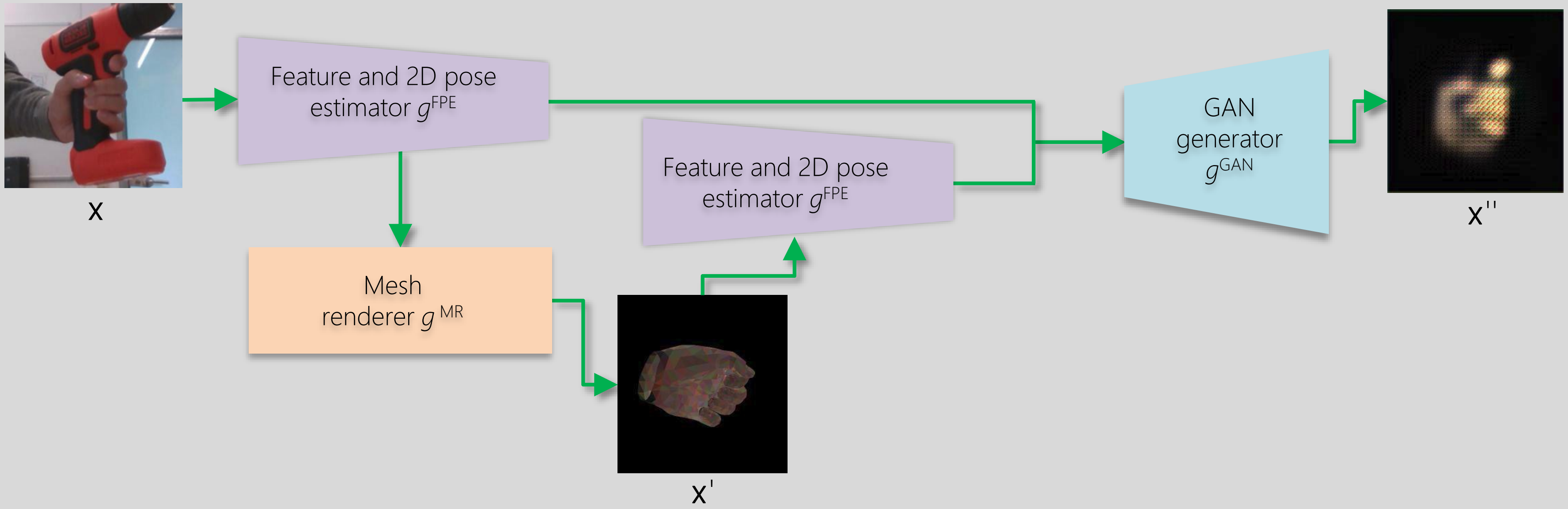
Key Idea

Finally, we obtain the skeletons from the mesh.

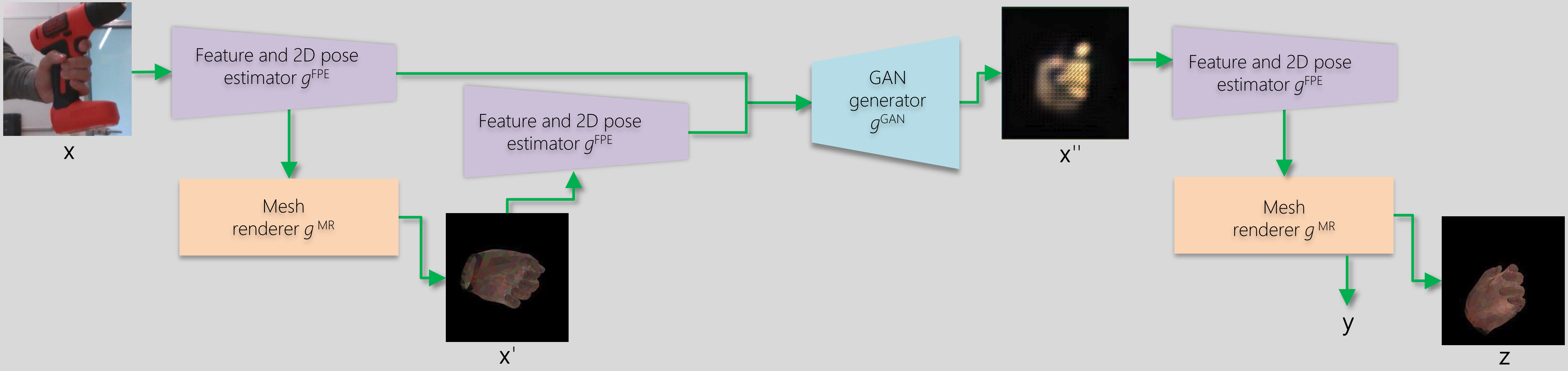
Pipeline



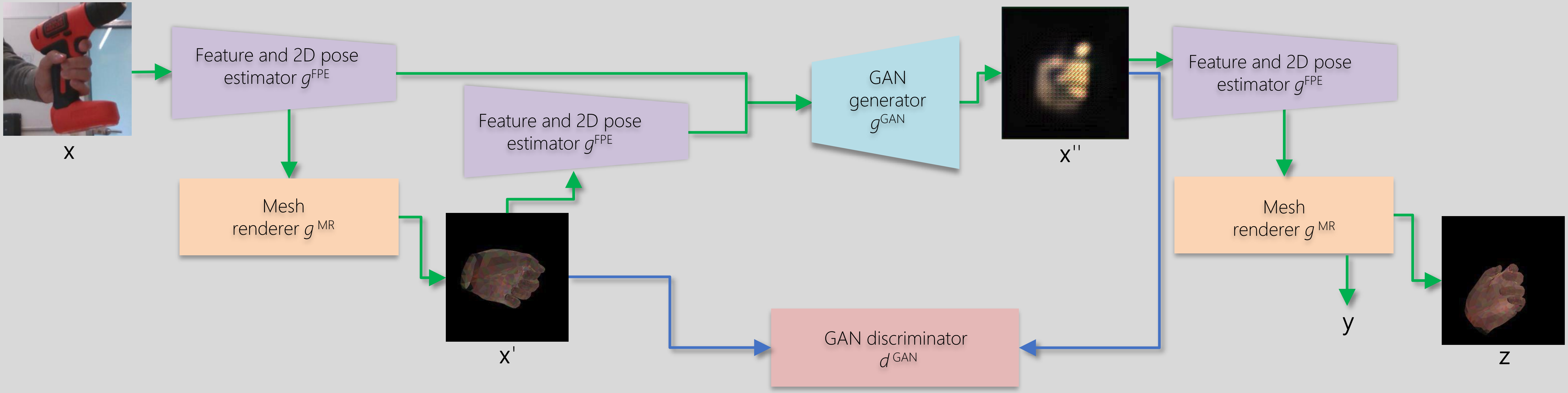
Pipeline


Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''

Pipeline

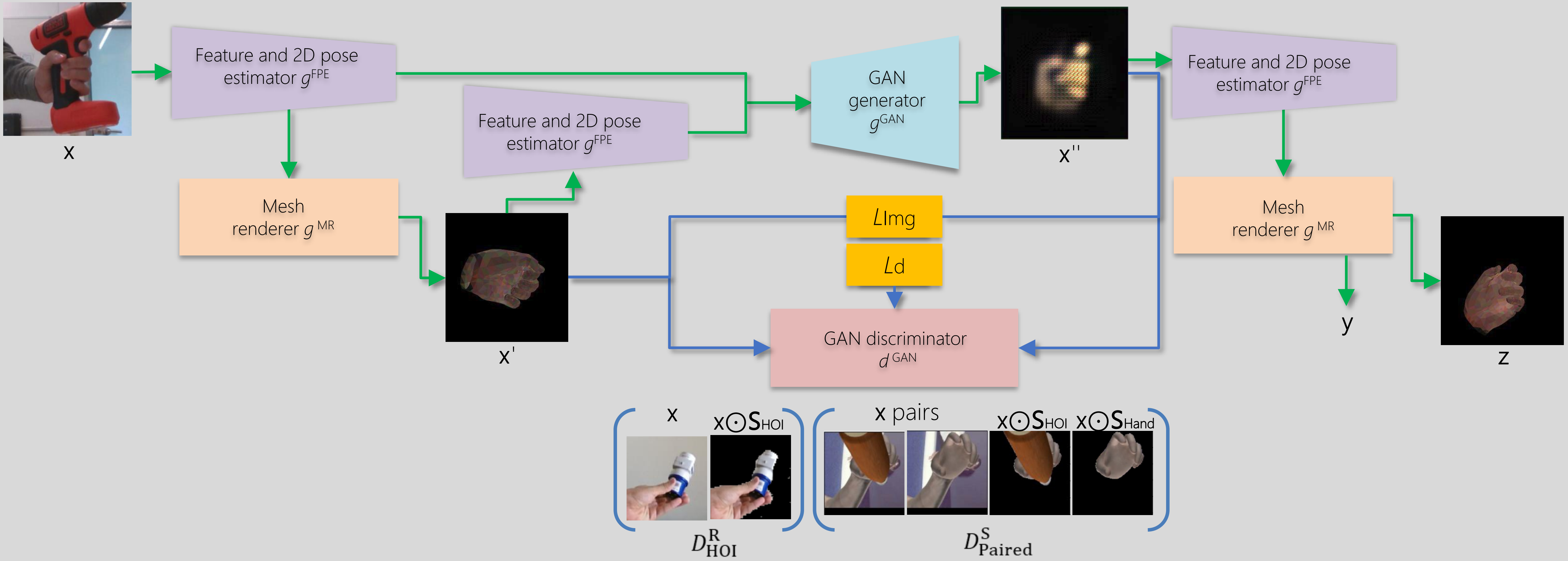

x
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
x'
Feature and 2D pose estimator g^{FPE}
GAN generator g^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
y
z

Pipeline



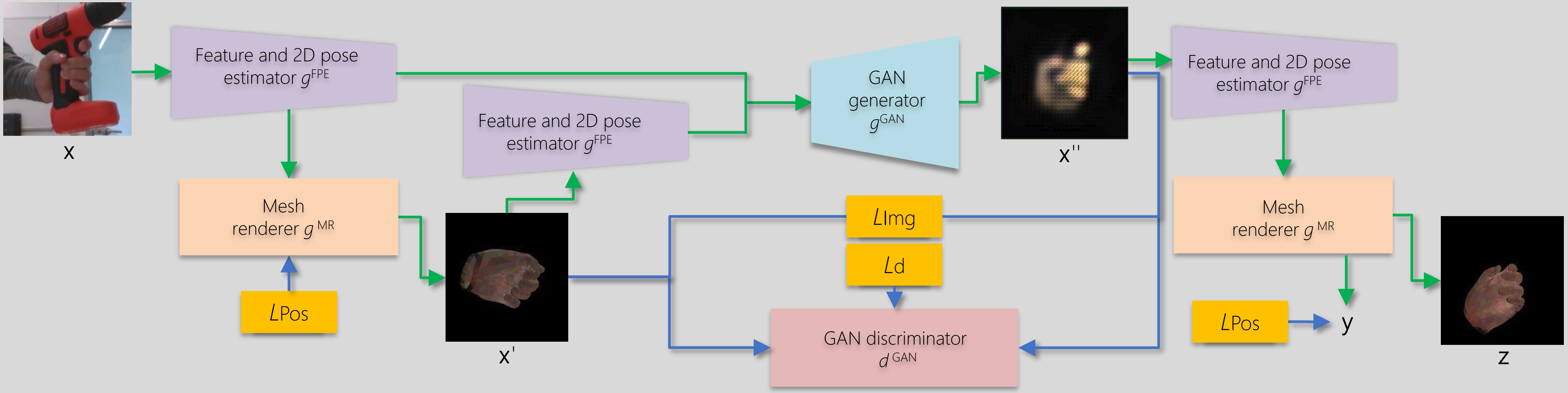
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
GAN discriminator d^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
y
z

Pipeline



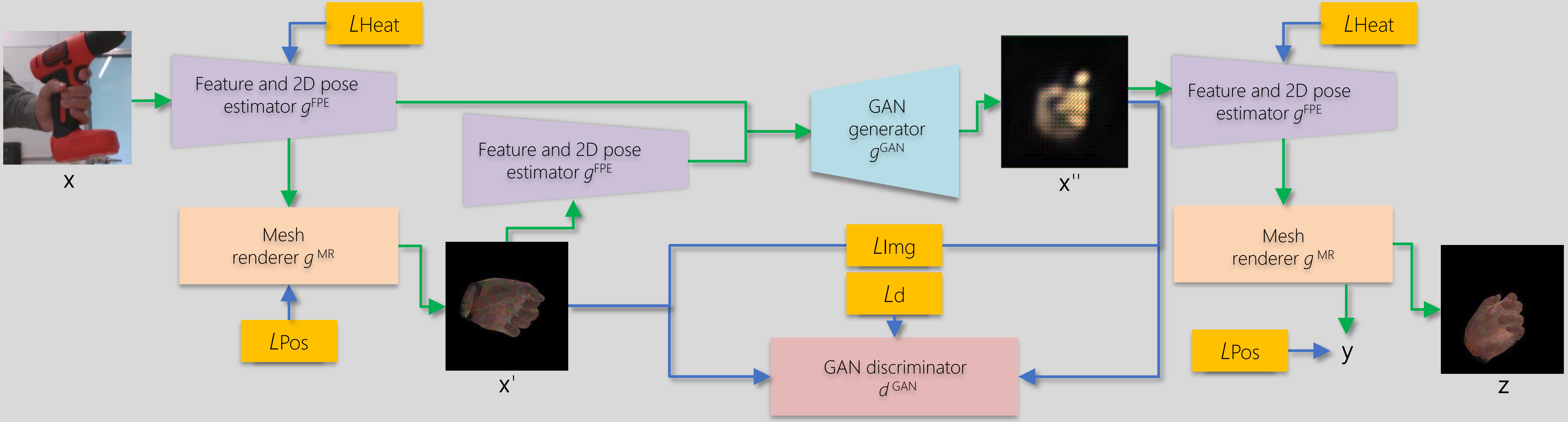
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LImg
Ld
GAN discriminator d^{GAN}
y
z
x
x⊙S_{HOI}
D^{R}_{HOI}
x pairs
x⊙S_{HOI}
x⊙S_{Hand}
D^{S}_{Paired}

Pipeline



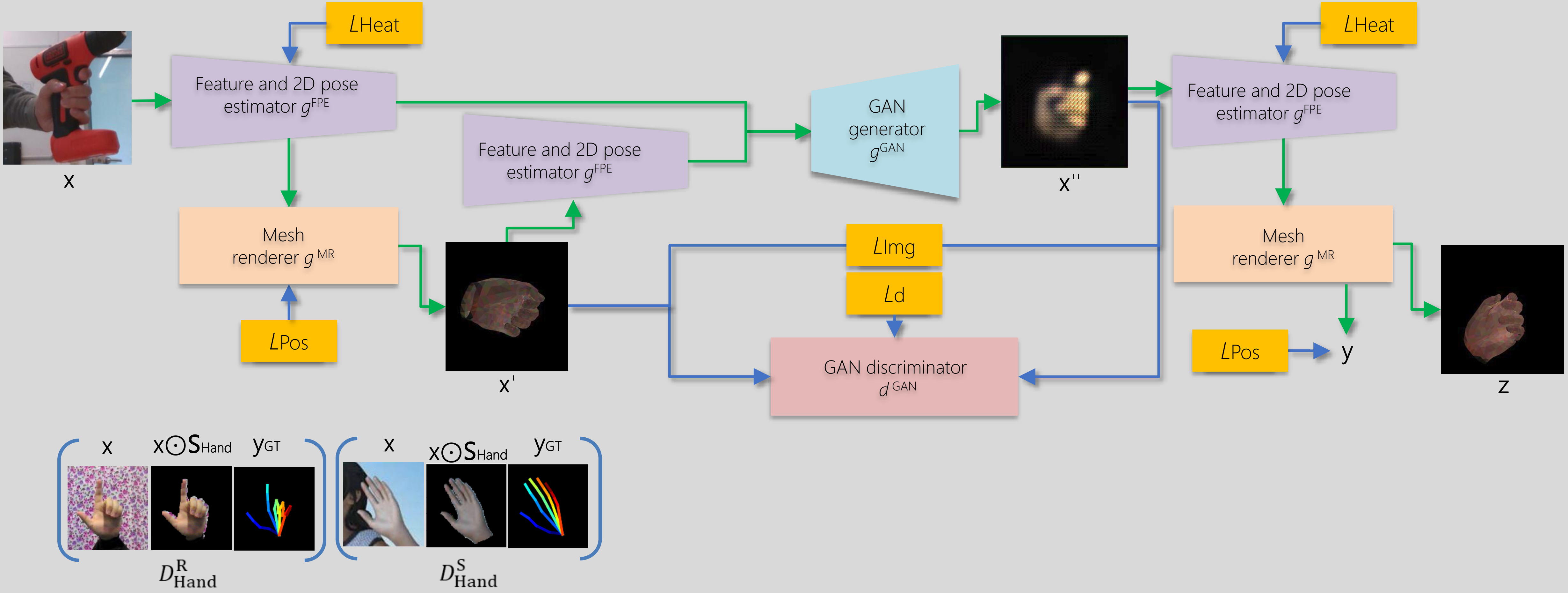
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z

Pipeline



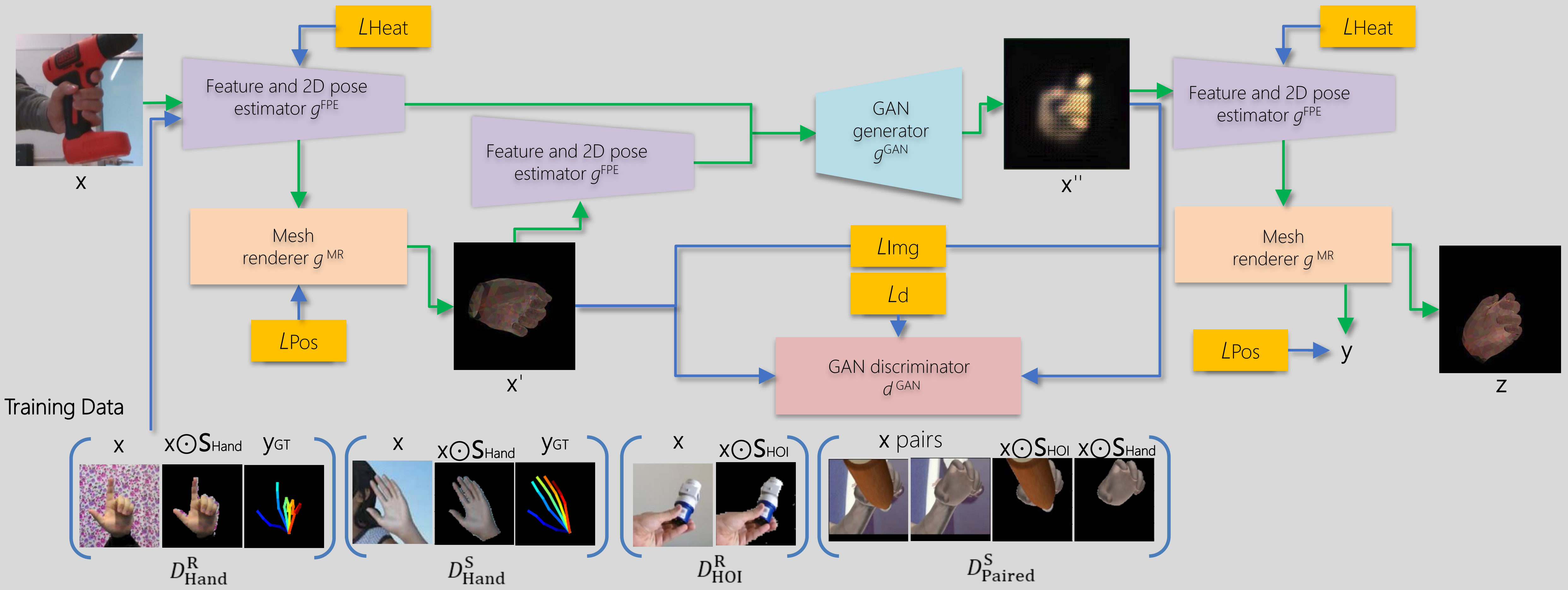
LHeat
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z

Pipeline



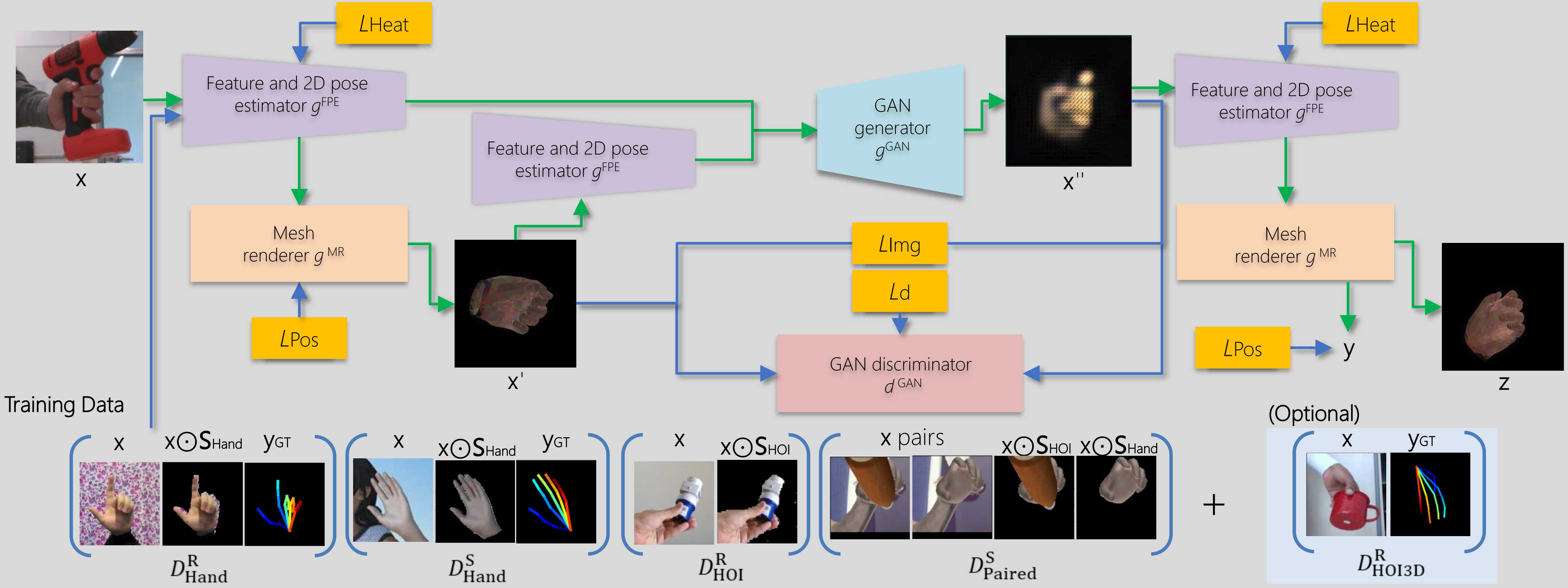
LHeat
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z
x
x⊙SHand
yGT
D^{R}_{Hand}
x
x⊙SHand
yGT
D^{S}_{Hand}

Pipeline



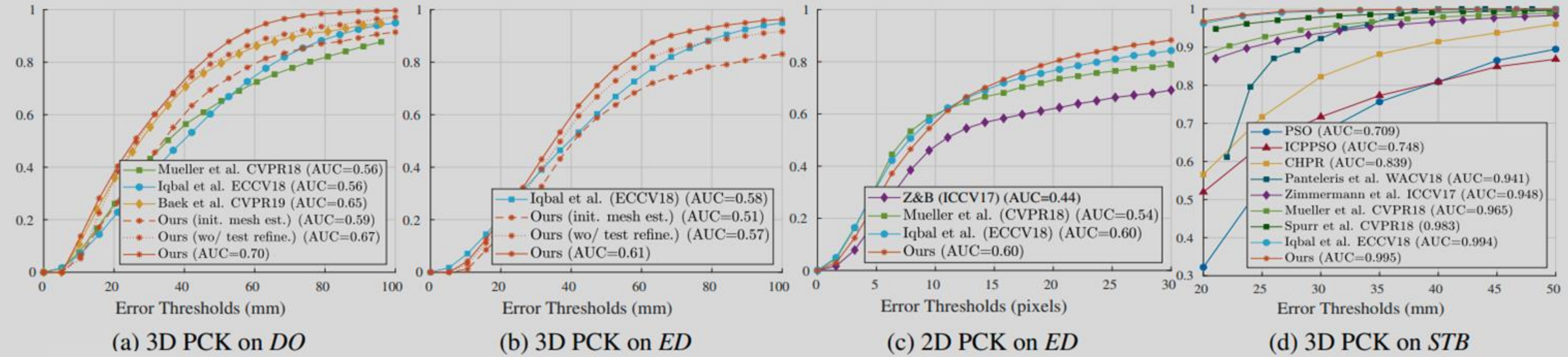
x
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
x'
Feature and 2D pose estimator g^{FPE}
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z
Training Data
x
x⊙SHand
yGT
D^R_{Hand}
x
x⊙SHand
yGT
D^S_{Hand}
x
x⊙SHOI
D^R_{HOI}
x pairs
x⊙SHOI
x⊙SHand
D^S_{Paired}

Pipeline



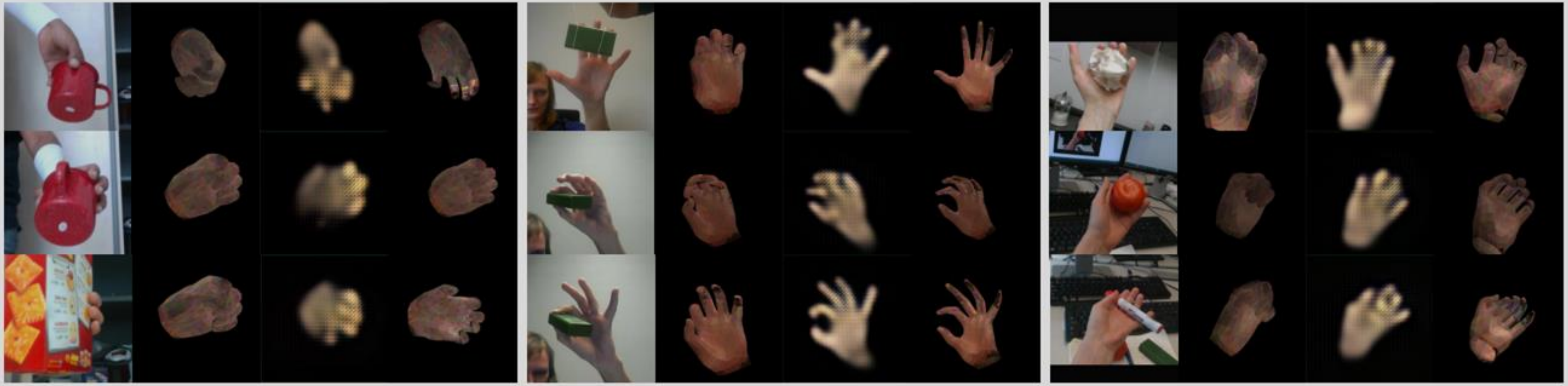
LHeat
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z
Training Data
x
x⊙SHand
yGT
D^R_{Hand}
x
x⊙SHand
yGT
D^S_{Hand}
x
x⊙SHOI
D^R_{HOI}
x pairs
x⊙SHOI
x⊙SHand
D^S_{Paired}
+
(Optional)
x
yGT
D^R_{HOI3D}

Results

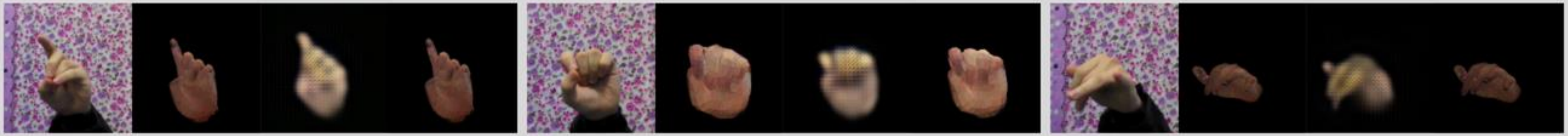


- We obtained state-of-the-art performance, with our weakly supervised approach in challenging HOI datasets (DO, ED).
- We maintained the state-of-the-art performance in hand-only dataset (STB).

Results



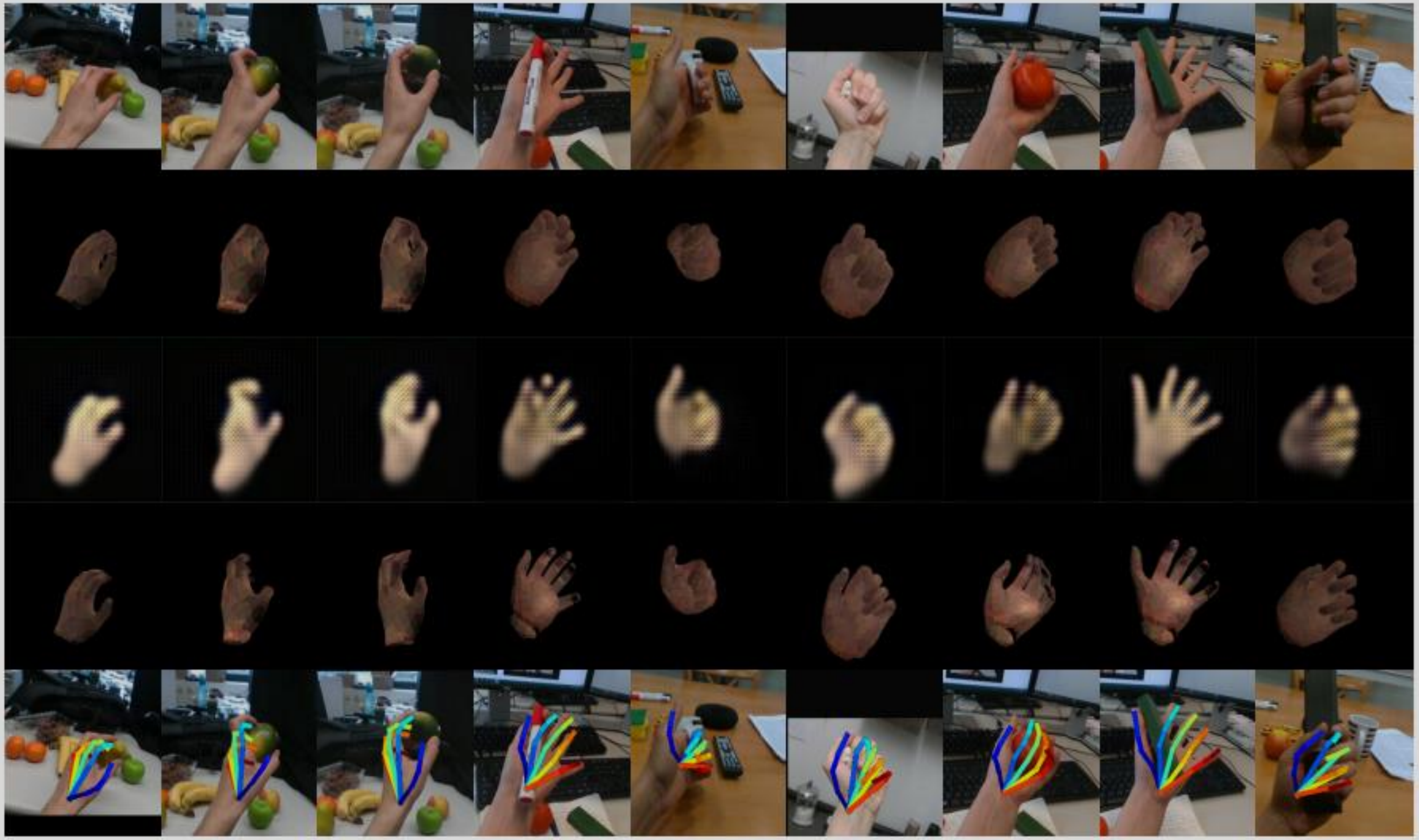
Hand-Object Interaction examples (Input/Init. Mesh/GAN output/2nd Mesh).



Hand-only examples (Input/Init. Mesh/GAN output/2nd Mesh).

Results

Results

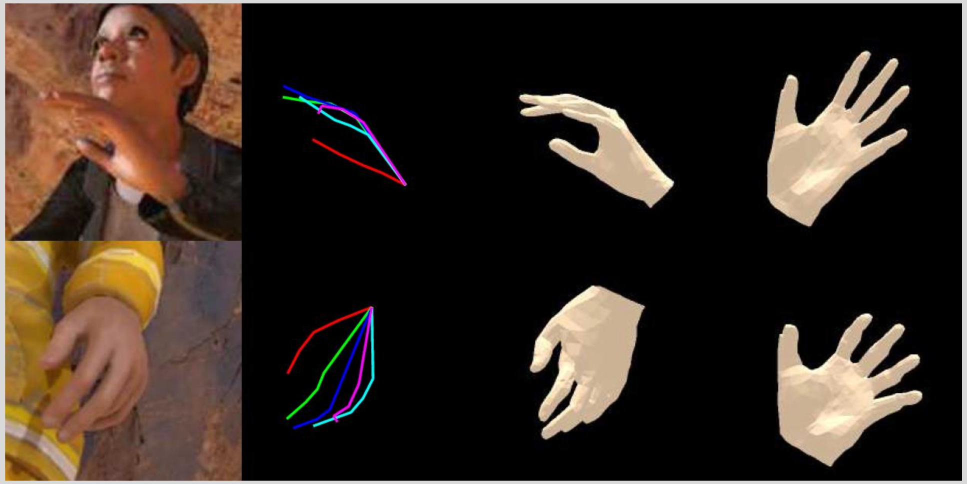


Results

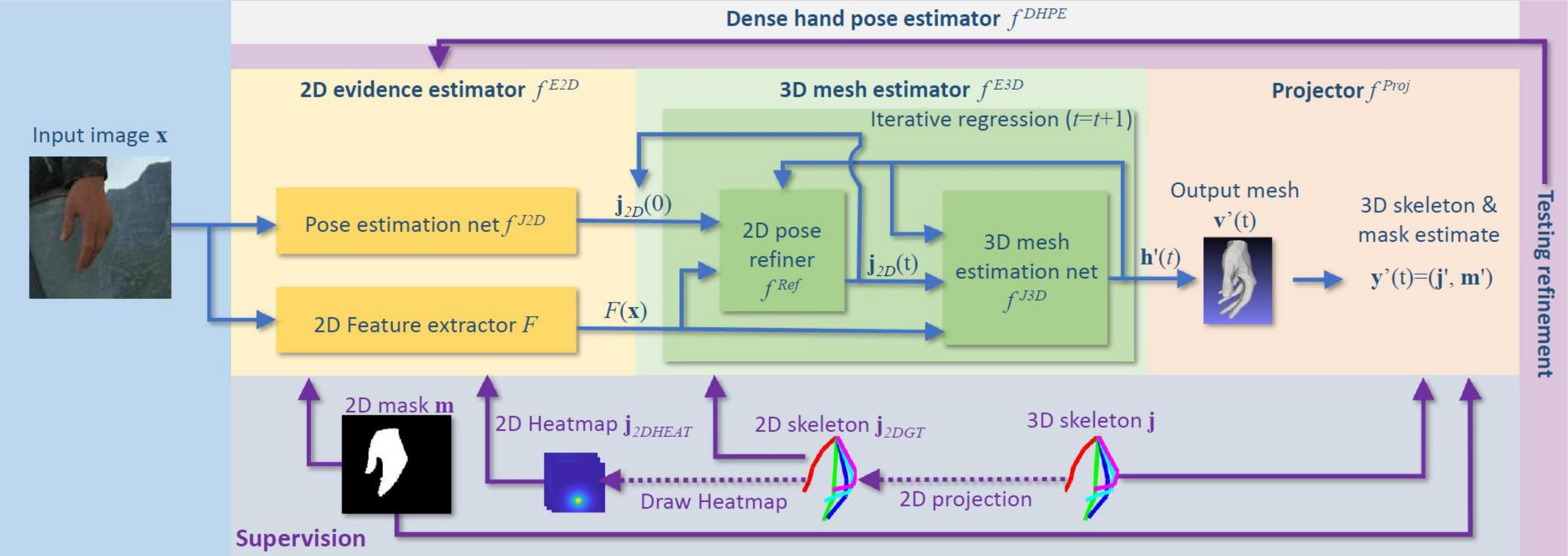
RGB-based Dense 3D Hand Pose via Neural Rendering

CVPR19

Dense hand poses

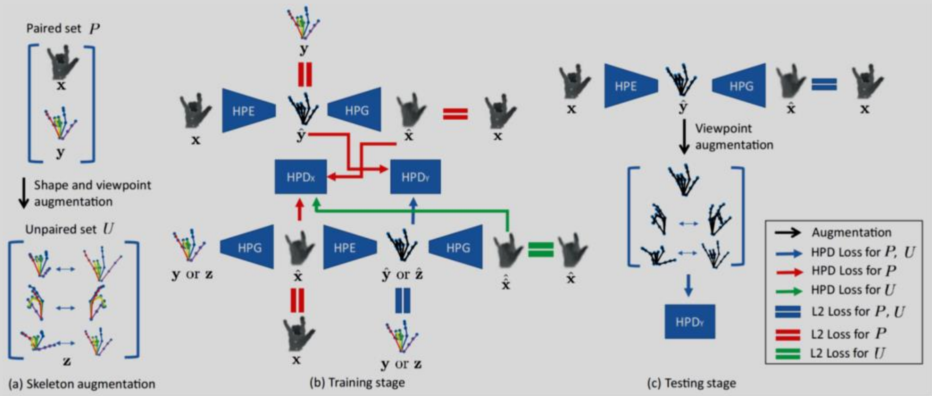


Learning 3D shapes by weak supervision via neural renderer



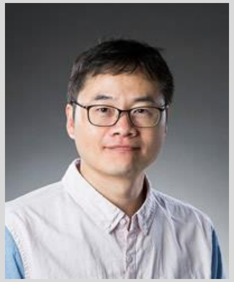
Augmented Skeleton Space Transfer for Depth-based Hand Pose Estimation

(CVPR18 oral)



Seungryul Baek

Imperial College London



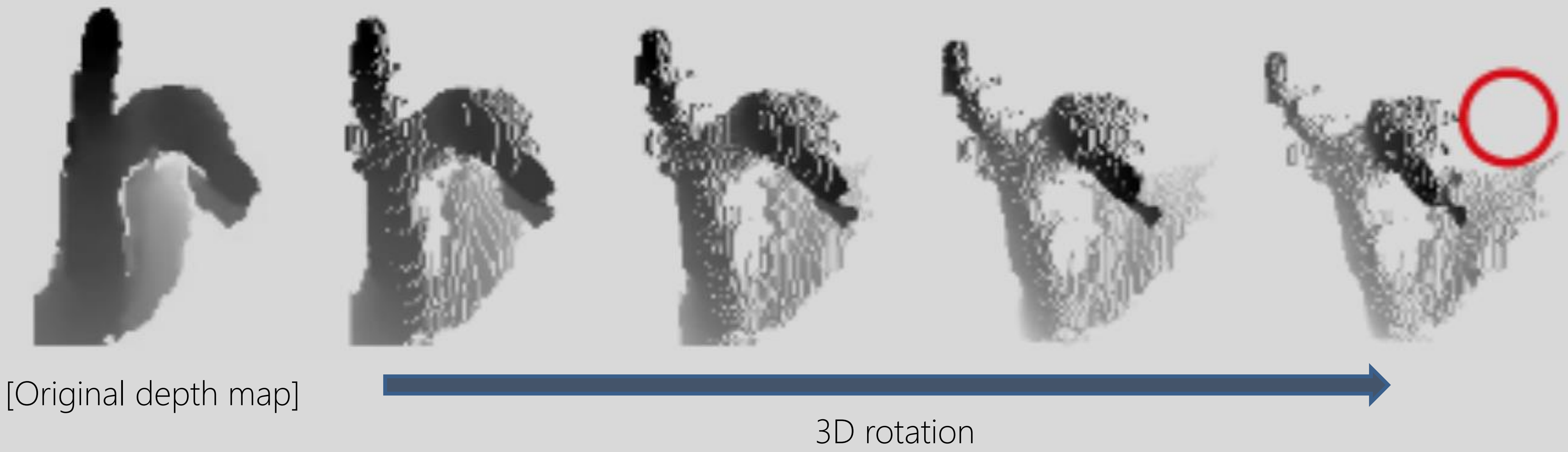
Kwang In Kim

University of Bath

Tae-Kyun Kim

Imperial College London

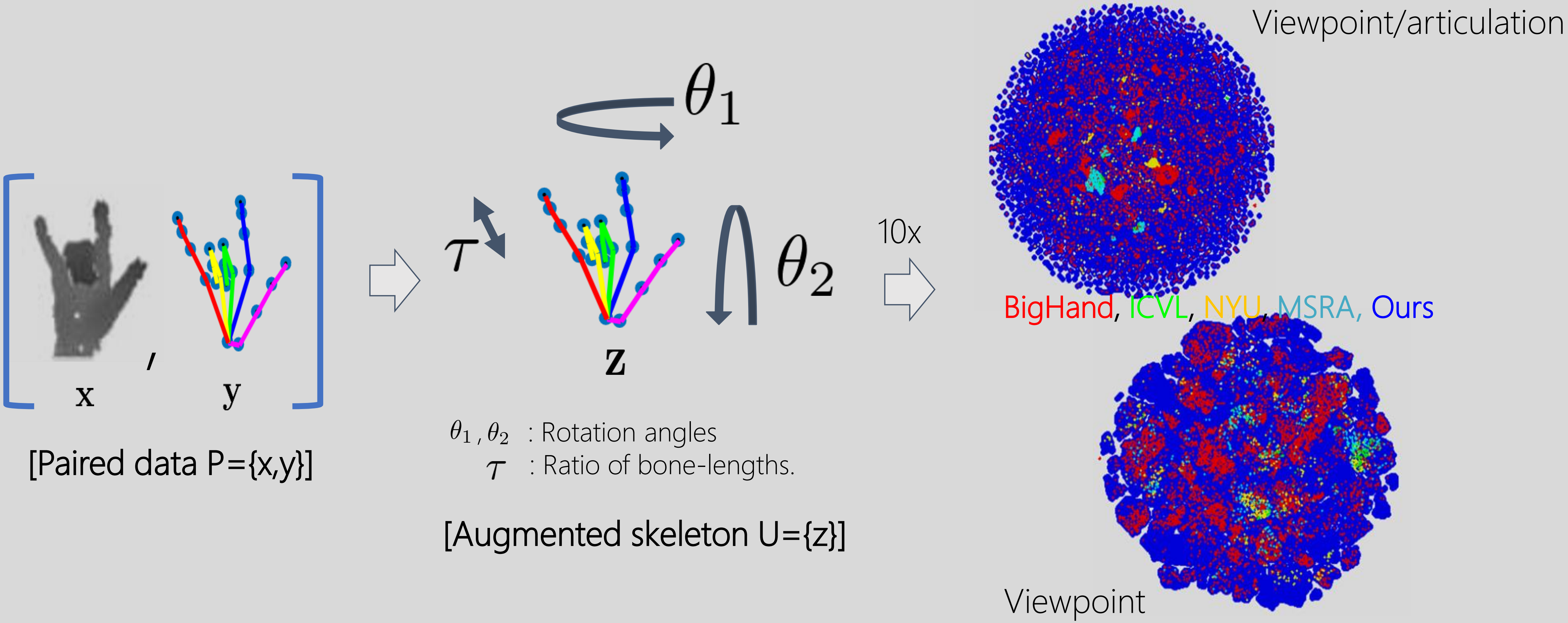
Directly augmenting depth images is difficult



→ Rotating a 2.5D depth map in 3D results in missing pixels.
→ Non-trivial is to change hand shapes (long-slim/fat, small/big etc).

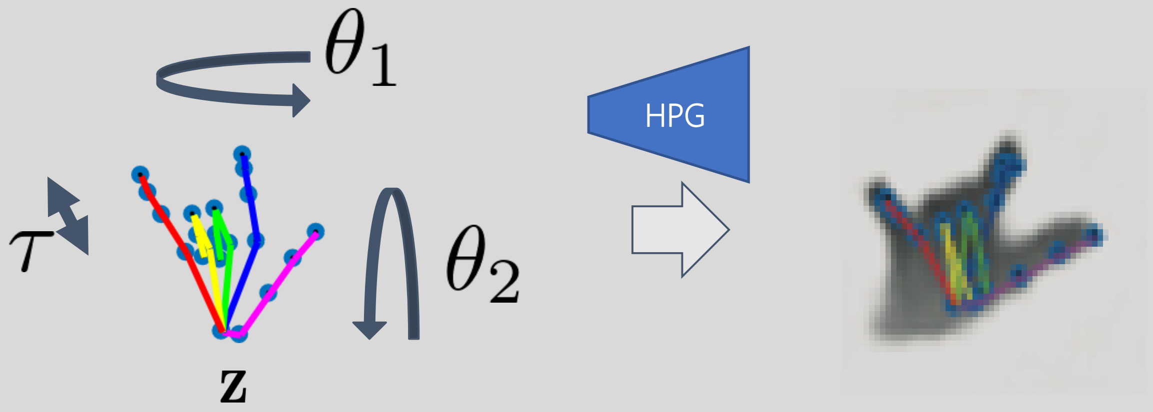
Data augmentation in skeleton space

- Generate data with unseen shapes/viewpoints from paired data.



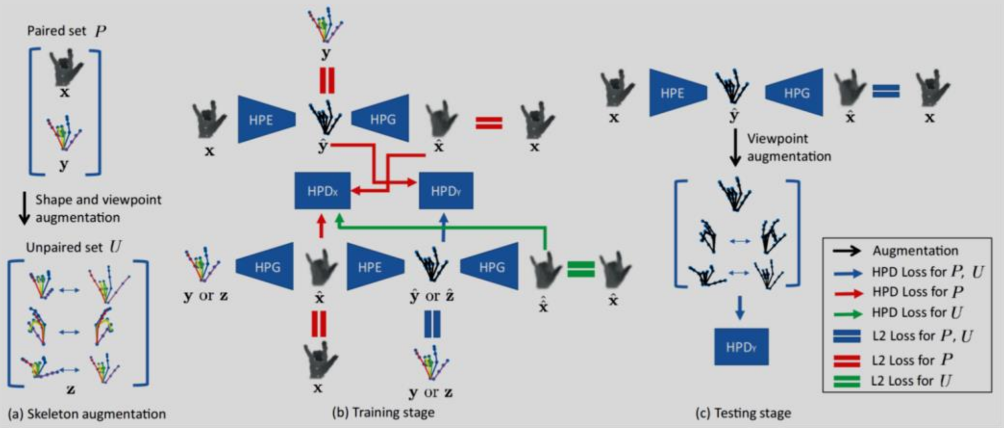
Augmented skeleton space transfer to depth

- Joint learning of 4 networks (HPE, HPG, HPDx, HPDy) to transfer augmented skeletons to depth images.

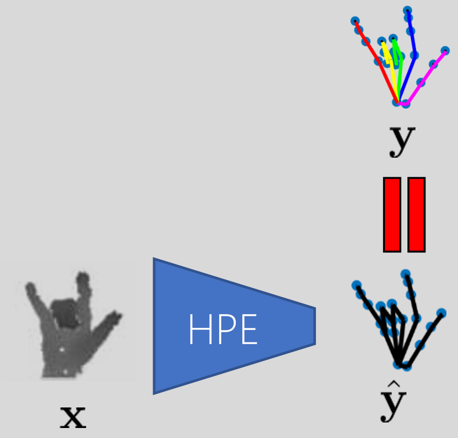


[Augmented skeleton U={z}]

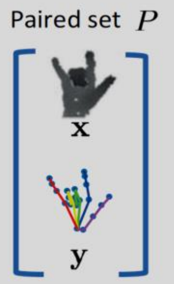
[Synthesized depth image]



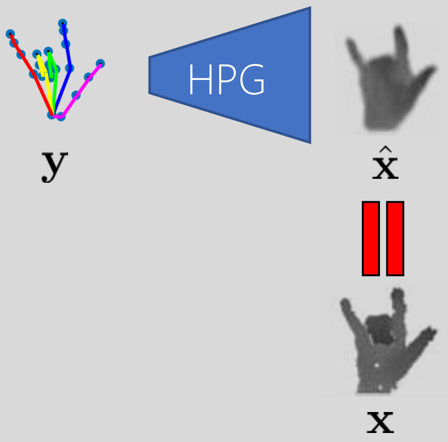
Joint Learning of HPG/HPE/HPD



- HPE and HPG are trained by paired data P={x, y}.

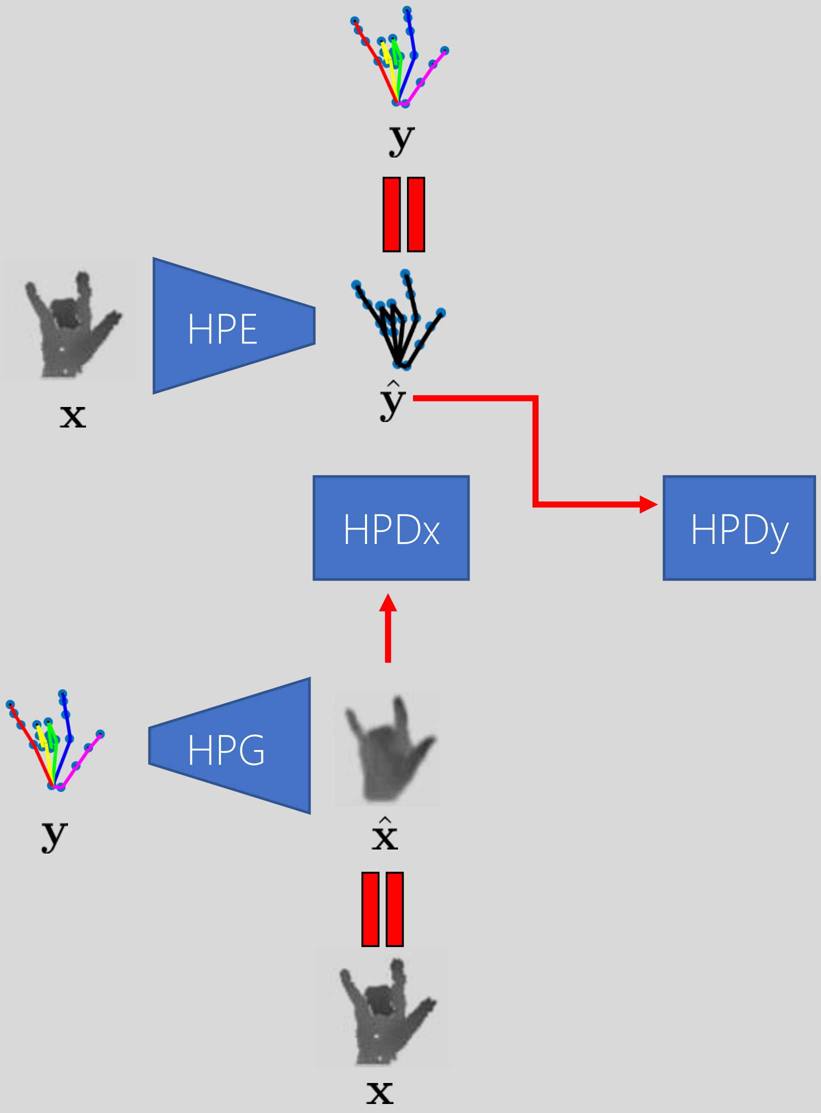


$$\mathcal{L}_E(f^E, f^{D_Y}) = ||f^E(\mathbf{x}) - \mathbf{y}||_2^2$$



$$\mathcal{L}_G(f^G, f^{D_X}) = ||f^G(\mathbf{y}) - \mathbf{x}||_2^2$$

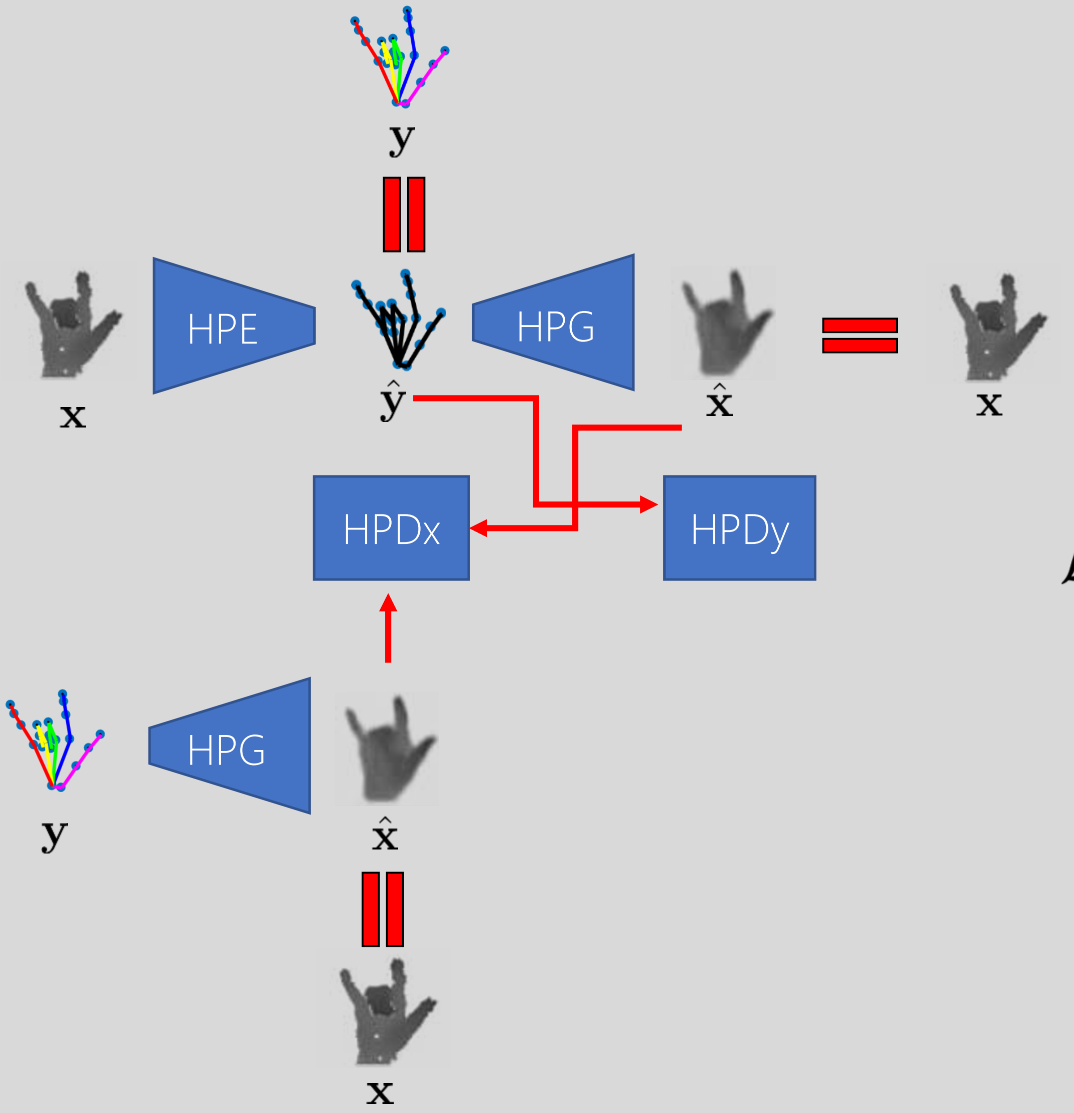
Joint Learning of HPG/HPE/HPD



- Adversarial loss is added.

$$
\begin{aligned}
\mathcal{L}_E(f^E, f^{D_Y}) &= ||f^E(\mathbf{x}) - \mathbf{y}||_2^2 \\
&\quad + \mathbb{E}_{\mathbf{y}}[\log f^{D_Y}(\mathbf{y})] \\
&\quad + \mathbb{E}_{\mathbf{x}}\left[\log(1 - f^{D_Y}(f^E(\mathbf{x}))\right] \\
\mathcal{L}_G(f^G, f^{D_X}) &= ||f^G(\mathbf{y}) - \mathbf{x}||_2^2 \\
&\quad + \mathbb{E}_{\mathbf{x}}[\log f^{D_X}(\mathbf{x})] \\
&\quad + \mathbb{E}_{\mathbf{y}}\left[\log(1 - f^{D_X}(f^G(\mathbf{y}))\right]
\end{aligned}
$$

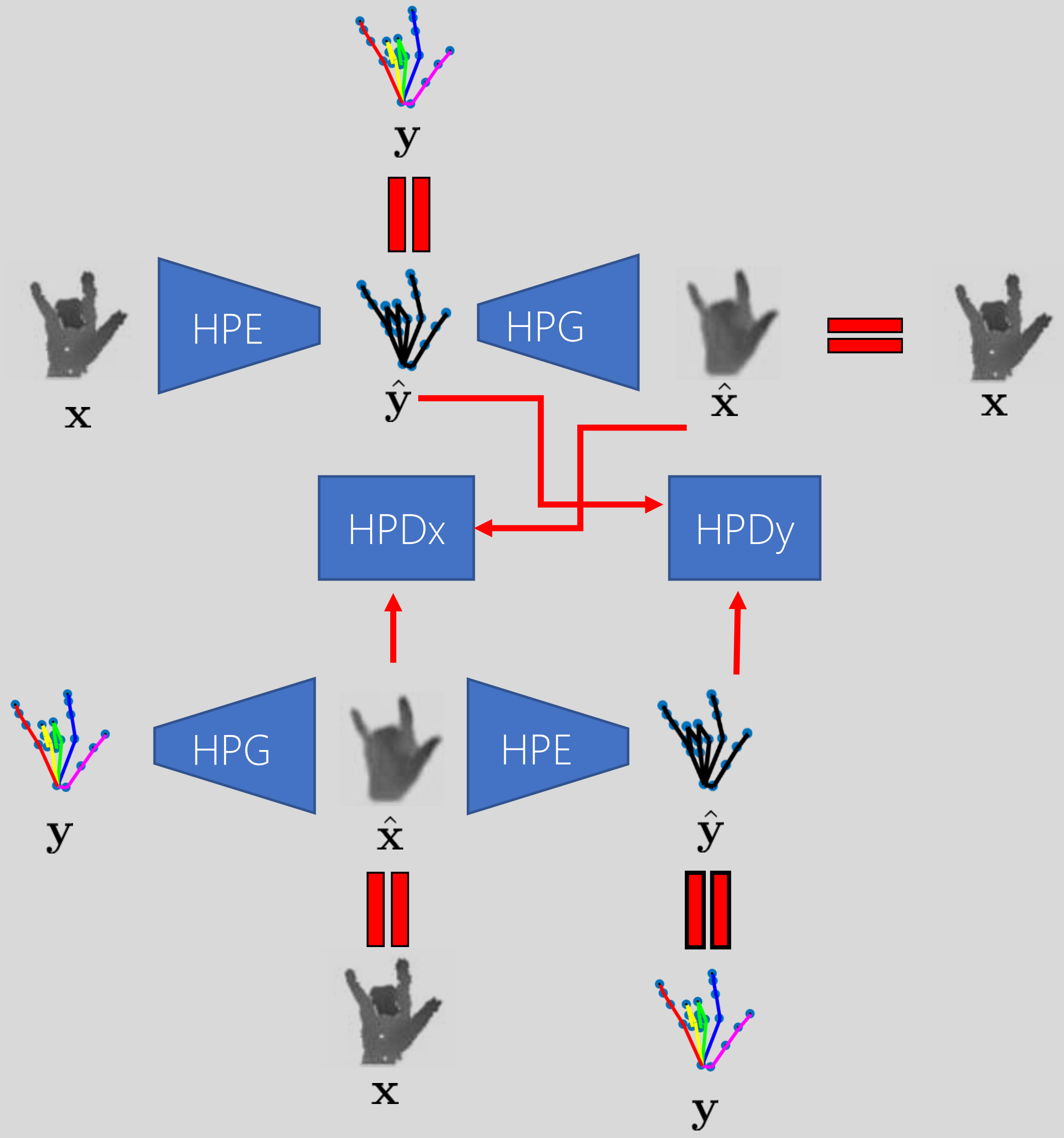
Joint Learning of HPG/HPE/HPD



- Cyclic consistency on x.

$$\mathcal{L}_P(f^E, f^G) = ||f^G(f^E(\mathbf{x})) - \mathbf{x}||_2^2 + \mathbb{E}_{\mathbf{x}}\left[\log(1 - f^{Dx}(f^G(f^E(\mathbf{x}))))\right] + \mathbb{E}_{\mathbf{y}}\left[\log f^{Dx}(\mathbf{x})\right]$$

Joint Learning of HPG/HPE/HPD

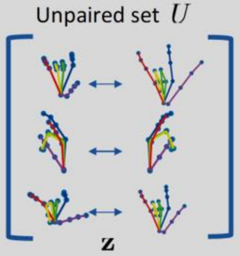


- Cyclic consistency on y.

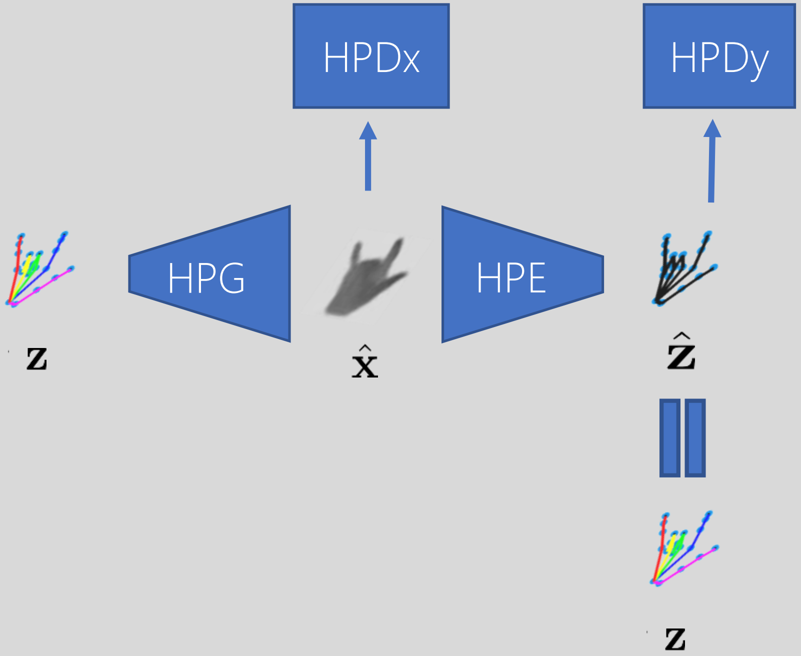
$$
\begin{aligned}
\mathcal{L}_P(f^E, f^G) \quad = \quad & ||f^G(f^E(\mathbf{x})) - \mathbf{x}||_2^2 \\
& + \mathbb{E}_{\mathbf{x}}\left[\log(1 - f^{D_X}(f^G(f^E(\mathbf{x}))))\right] \\
& + \mathbb{E}_{\mathbf{y}}\left[\log f^{D_X}(\mathbf{x})\right] \\
& + ||f^E(f^G(\mathbf{y})) - \mathbf{y}||_2^2 \\
& + \mathbb{E}_{\mathbf{y}}\left[\log(1 - f^{D_Y}(f^E(f^G(\mathbf{y}))))\right] \\
& + \mathbb{E}_{\mathbf{x}}\left[\log f^{D_Y}(\mathbf{y})\right]
\end{aligned}
$$

Joint Learning of HPG/HPE/HPD

- Cyclic consistency for unpaired data U={z}.

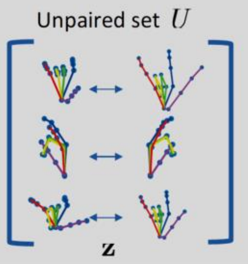


$$\mathcal{L}_U(f^E, f^G) = \|f^E(f^G(\mathbf{z})) - \mathbf{z}\|_2^2 + \mathbb{E}_{\mathbf{z}}\Big[\log f^{D_Y}(\mathbf{z}) + \log(1 - f^{D_Y}(f^E(f^G(\mathbf{z}))))\Big]$$

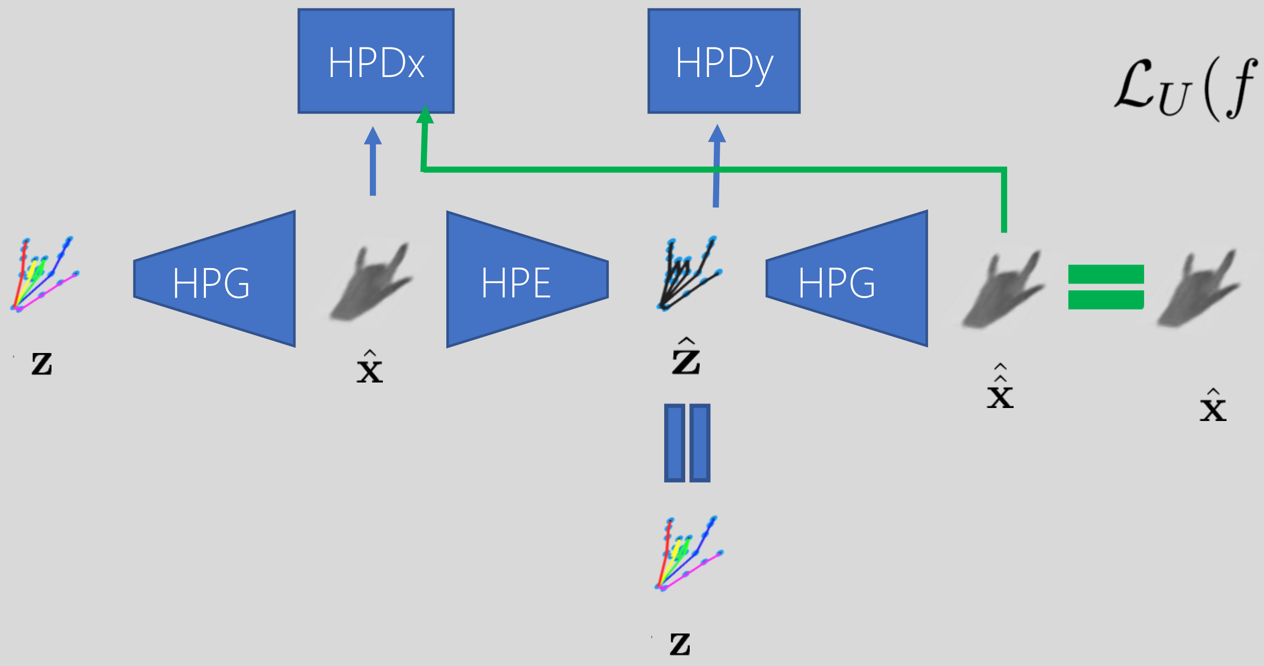


Joint Learning of HPG/HPE/HPD

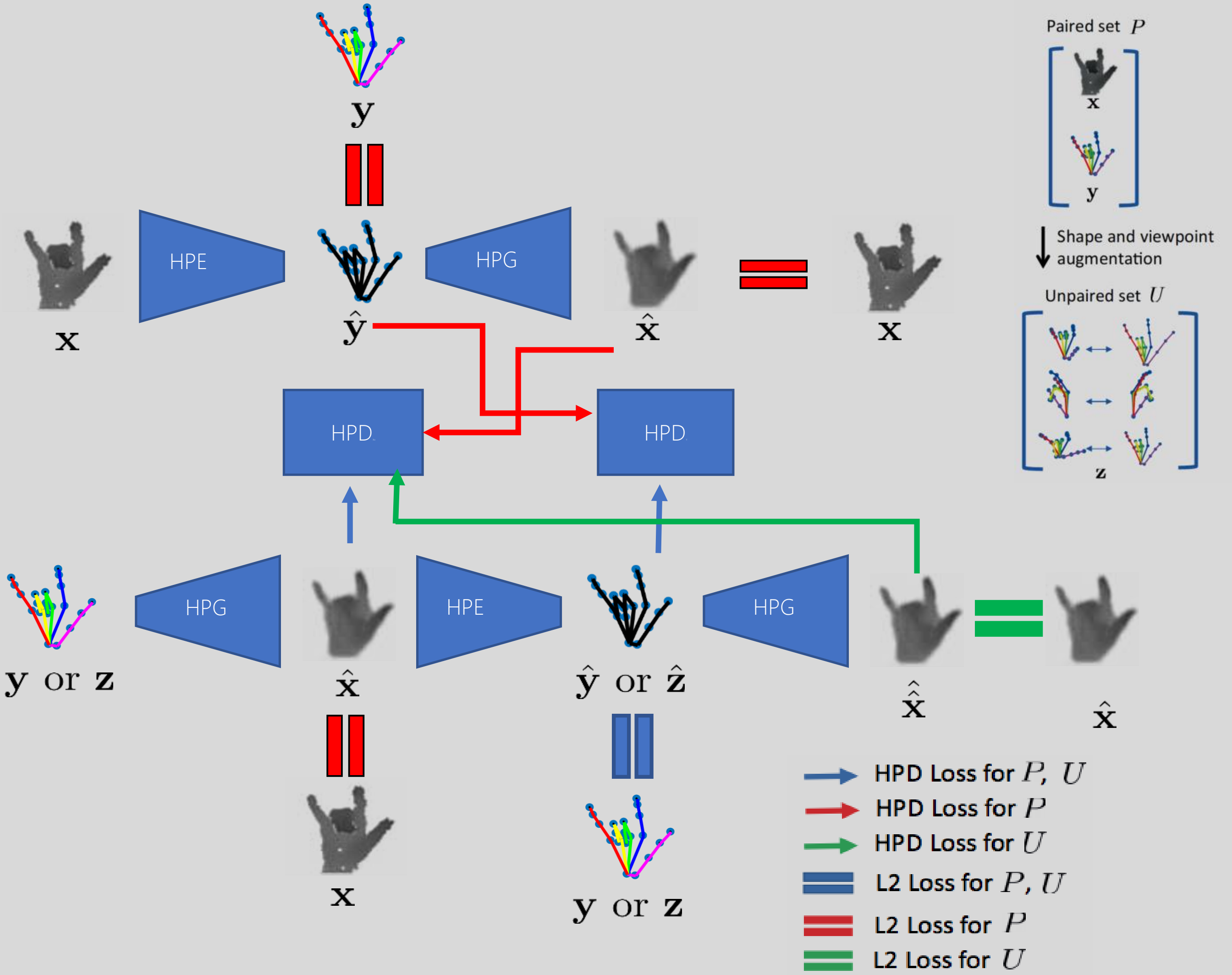
- Cyclic consistency for unpaired data U={z}.



$$
\begin{aligned}
\mathcal{L}_U(f^E, f^G) \quad = \quad & \|f^E(f^G(\mathbf{z})) - \mathbf{z}\|_2^2 \\
& + \mathbb{E}_{\mathbf{z}}\big[\log f^{D_Y}(\mathbf{z}) \\
& + \log(1 - f^{D_Y}(f^E(f^G(\mathbf{z}))))\big] \\
& + \|f^G(f^E(f^G(\mathbf{z}))) - f^G(\mathbf{z})\|_2^2 \\
& + \log(1 - f^{D_X}(f^G(\mathbf{z}))) \\
& + \log(1 - f^{D_X}(f^G(f^E(f^G(\mathbf{z})))))\big]
\end{aligned}
$$



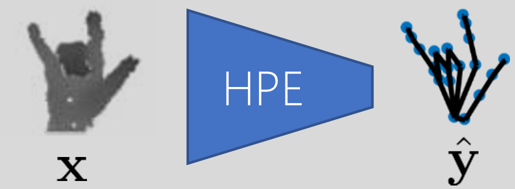
Joint Learning of HPG/HPE/HPD



- Final loss:

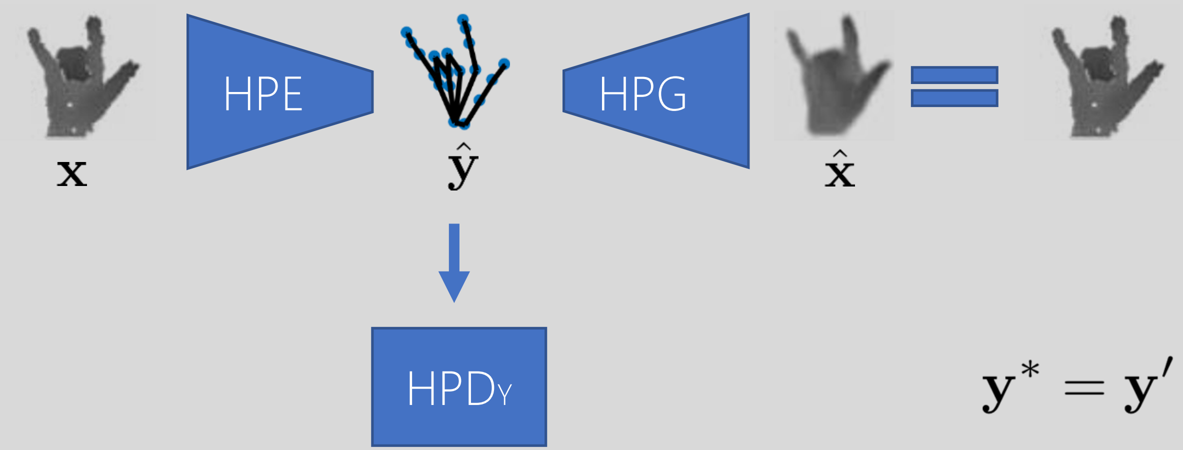
$$\mathcal{L}(f^G, f^E, f^{D_X}, f^{D_Y})$$
$$= \mathcal{L}_G(f^G, f^{D_X}) + \mathcal{L}_E(f^E, f^{D_Y})$$
$$+ \lambda(\mathcal{L}_P(f^E, f^G) + \mathcal{L}_U(f^E, f^G))$$

Inference with augmented skeletons



- Initial hand pose estimation

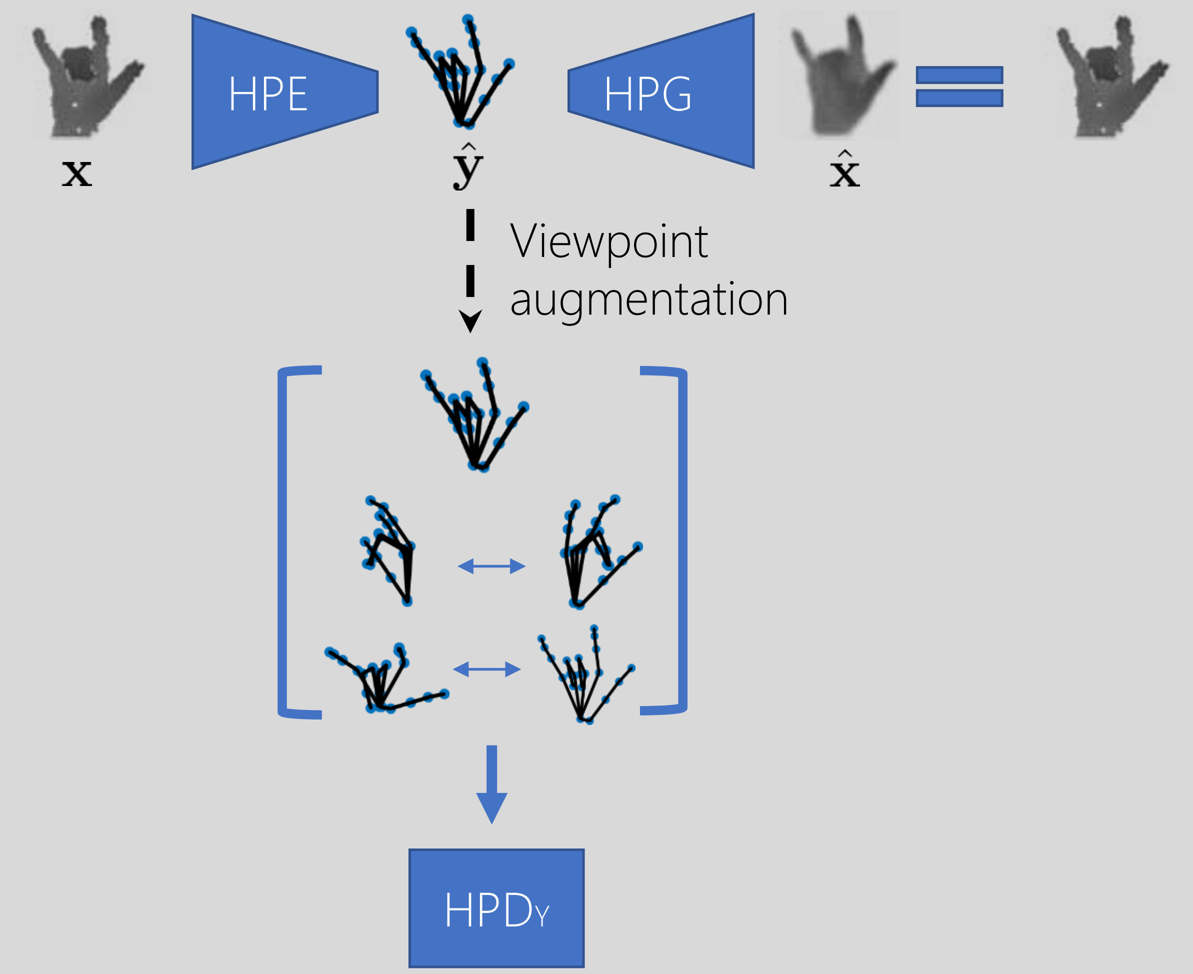
Inference with augmented skeletons



- The estimated pose is further refined by the gradients from both HPG and HPD_Y.

$$\mathbf{y}^* = \mathbf{y}' - \gamma \nabla \left(-f^{D_Y}(\mathbf{y}') + \lambda_{ref} || f^G(\mathbf{y}') - \mathbf{x}' ||_2^2 \right)$$

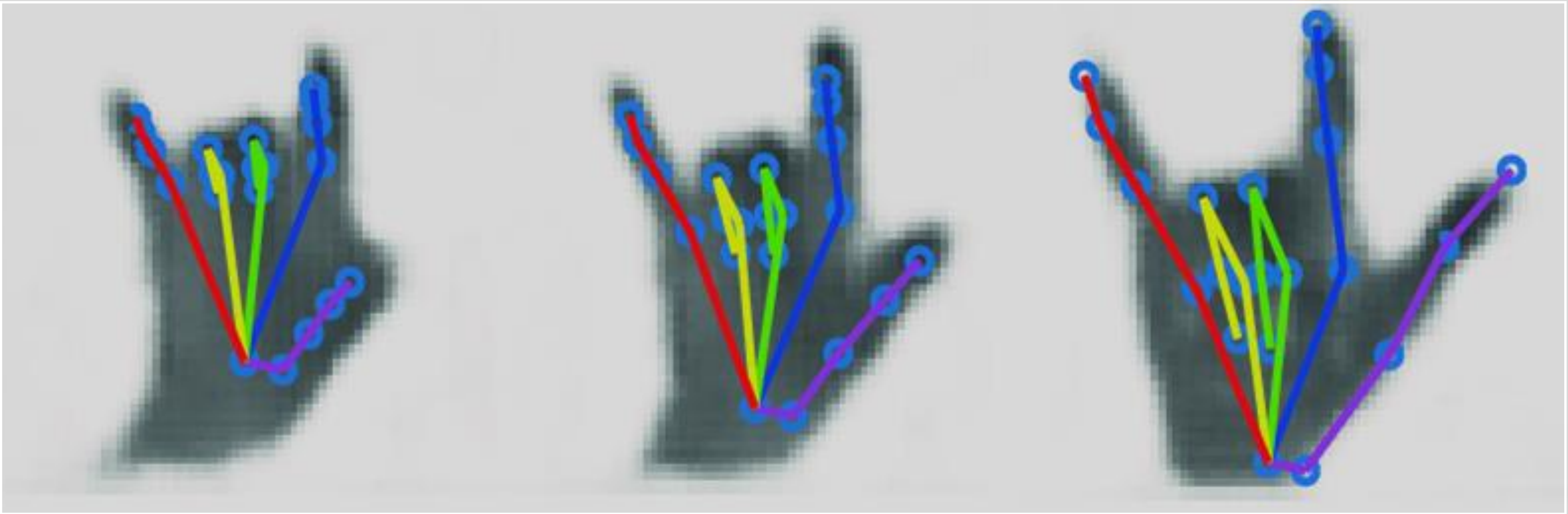
Inference with augmented skeletons



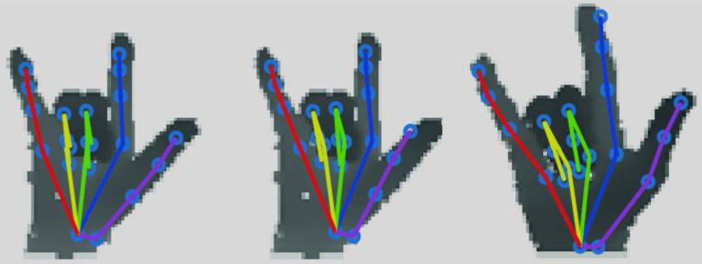
- Ensembling refinement:
  1) The estimated skeleton is randomly rotated.
  2) We receive gradients from multiple views by HPD_Y
  3) Then, we average them to the final result with the updates from HPG.

Transferred depth maps

HPG Output

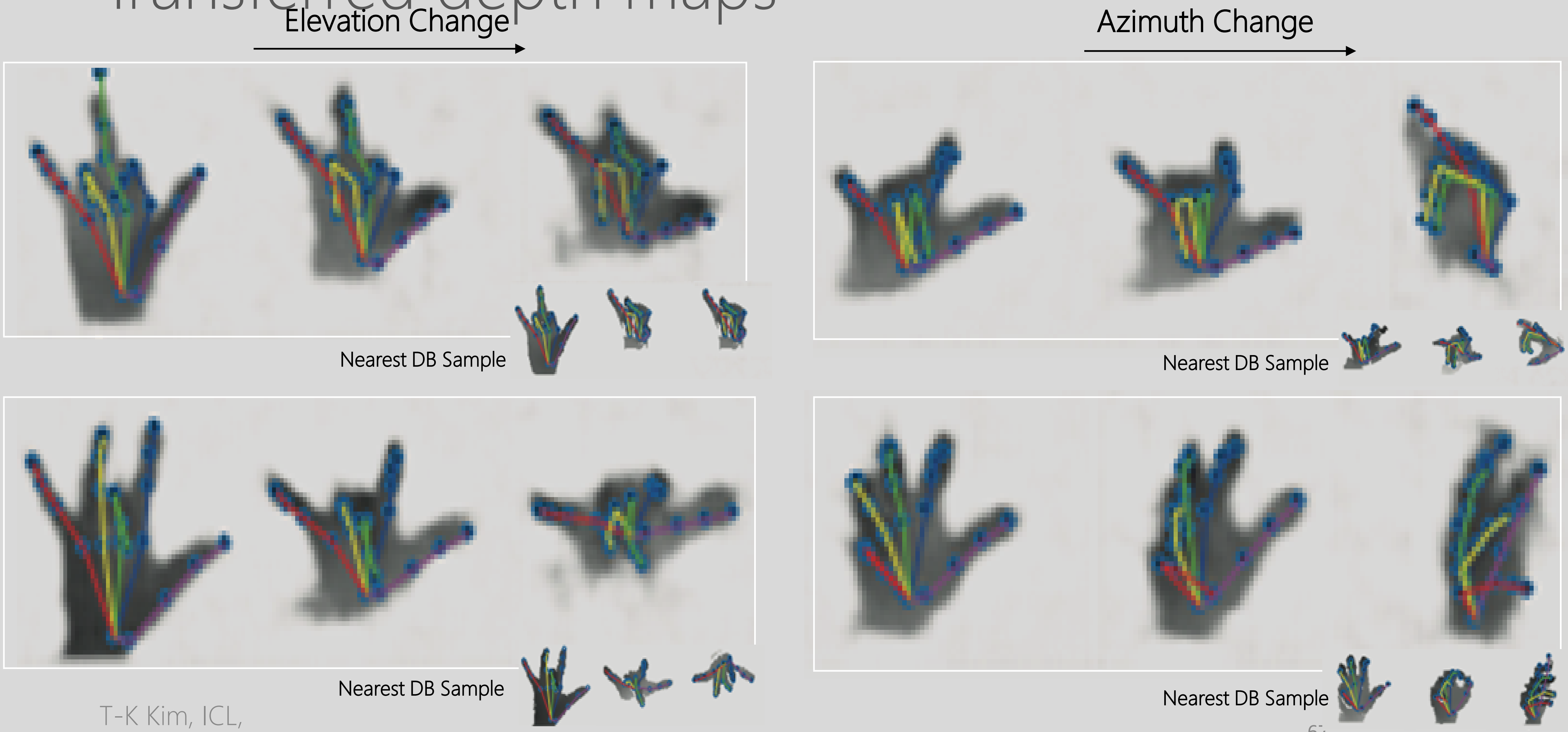


Nearest DB Sample

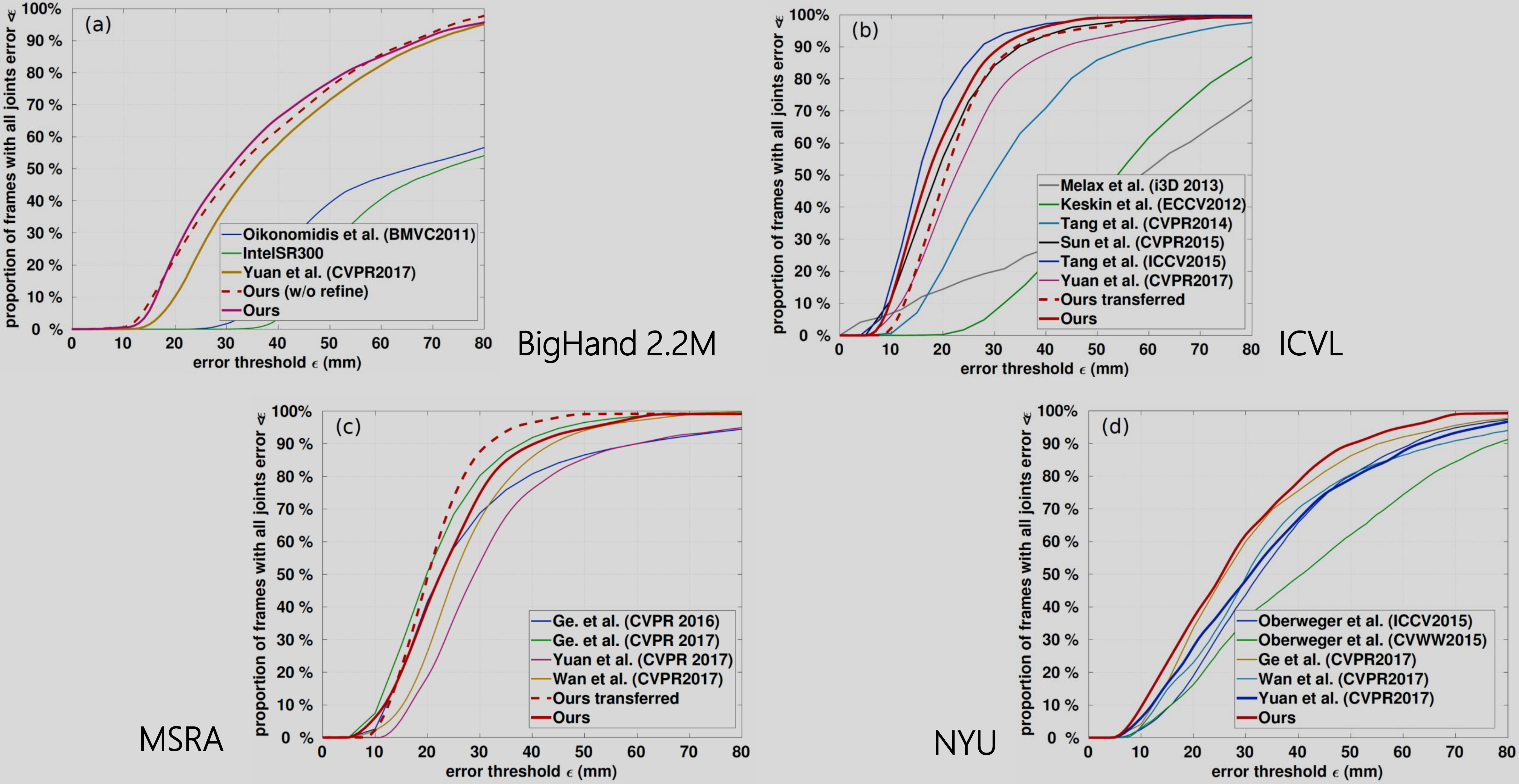


- HPG generates different shapes.

Transferred depth maps



Experiments



Experiments

(a)

Configuration	*Big Hand 2.2M*	*ICVL*	*MSRA*	*NYU*
f^G (baseline)	0.151	0.588	0.482	0.451
f^G (w/o aug.; refine)	0.124	0.516	0.470	0.415
f^G (w/o refine)	0.102	0.486	0.438	0.396
f^E (baseline)	17.1	12.1	16.3	17.3
f^E (w/o aug.; refine)	15.7	10.4	14.4	16.4
f^E (w/o refine)	14.1	9.1	13.1	14.9
f^E	13.7	8.5	12.5	14.1

(b)

Configuration	Error (mm)
HPE baseline	17.1
Ours (w/o aug.; refine)	15.7
Ours (w/ in-plane-rot 10x.; w/o aug.; refine)	14.9
Ours (5× aug.; w/o refine)	15.1
Ours (10× aug.; w/o refine)	14.1
Ours (20× aug.; w/o refine)	14.0
Ours (w/ in-plane-rot; 10x aug.; w/o refine)	12.5

f^G/f^E (baseline): HPG/HPE trained using $\mathcal{L}_G/\mathcal{L}_E$

f^G/f^E (w/o aug.; refine): HPG/HPE trained using $\mathcal{L}_G/\mathcal{L}_E+\mathcal{L}_P$,

f^G/f^E (w/o refine): HPG/HPE trained using $\mathcal{L}_G/\mathcal{L}_E+\mathcal{L}_P+\mathcal{L}_U$

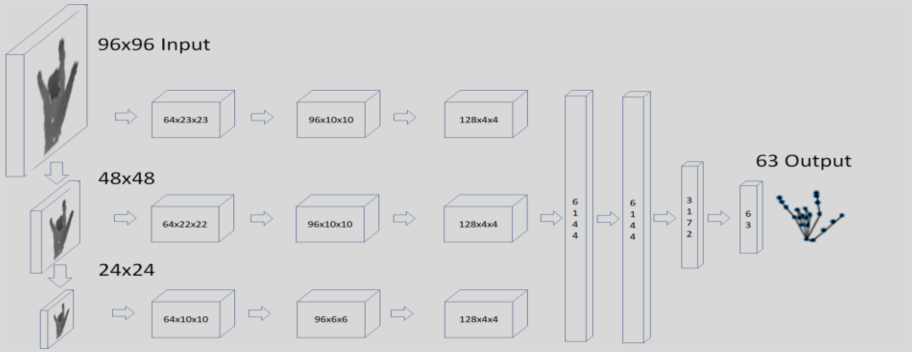
f^E: f^E trained using $\mathcal{L}_G/\mathcal{L}_E+\mathcal{L}_P+\mathcal{L}_U$+Testing refinement.

- HPG also improves its accuracy by seeing more data.
- Conventional augmentation (In-plane-rotation) is orthogonal to ours.
- We also augment 5x, 10x, 20x; 10x is the best considering time/accuracy.

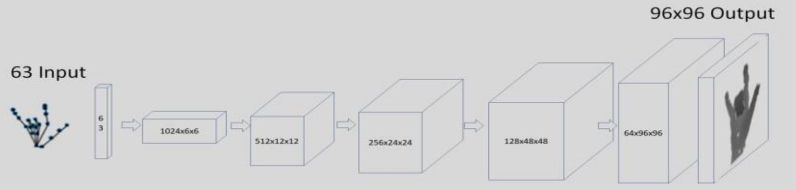
Network architecture and computation

- Implemented with the Torch library, on an Intel 3.40 GHz i7 machine with two NVIDIA GTX 1070 GPUs.
- Training: 3-4 days (100 epochs) on 10x augmentation.
- Testing: 300 FPS using the GPU

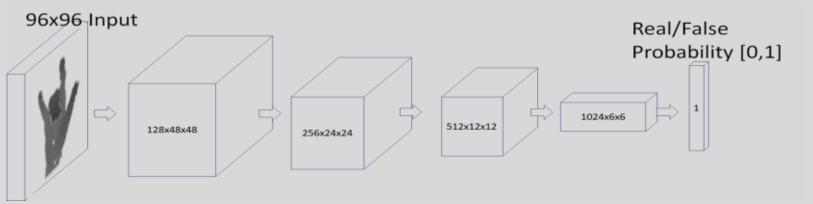
Also tried ResNet for HPE.



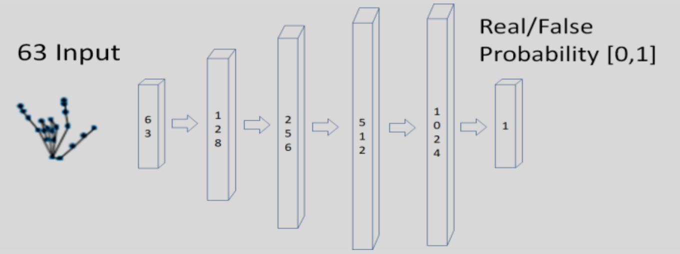
Hand Pose Estimator (HPE): f^E



Hand Pose Generator (HPG): f^G



Depth Discriminator (HPD_X): f^{D_X}



Skeleton Discriminator (HPD_Y): f^{D_Y}

ECCV'20 ONLINE 23-28 AUGUST 2020

FACER2VM

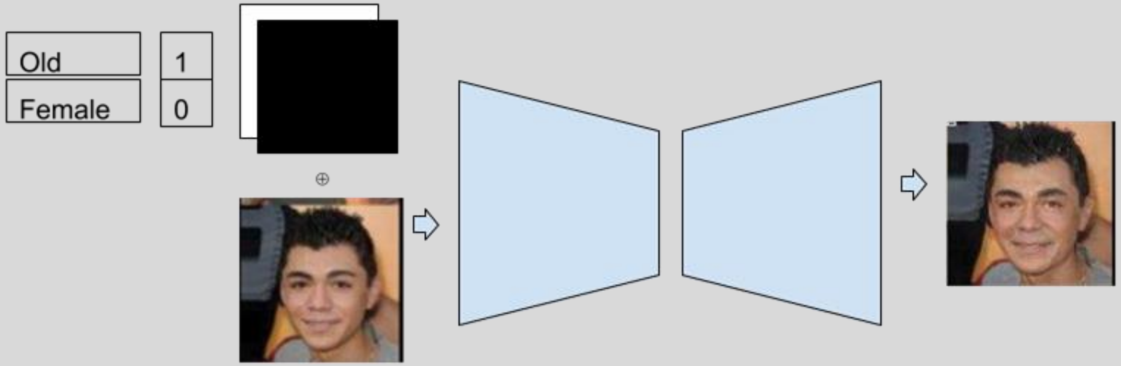
Imperial College London

Inducing Optimal Attributes Representations for Conditional GANs

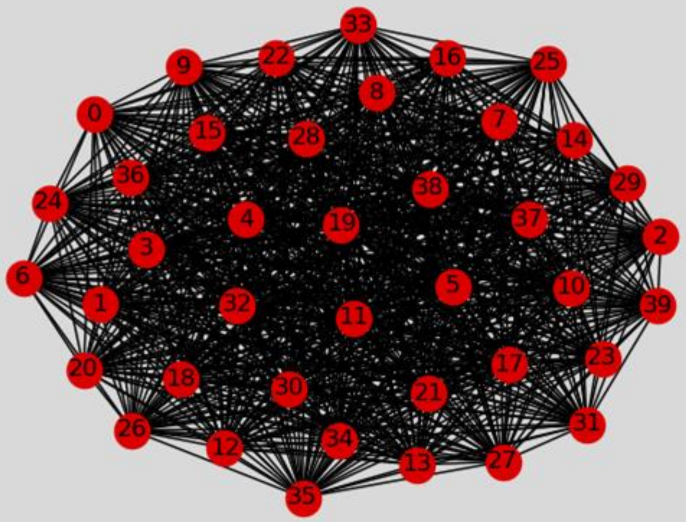
Binod Bhattarai[1], Tae-Kyun (T-K) Kim[1,2]

1. Imperial College London, UK
2. KAIST, Daegeon, South Korea

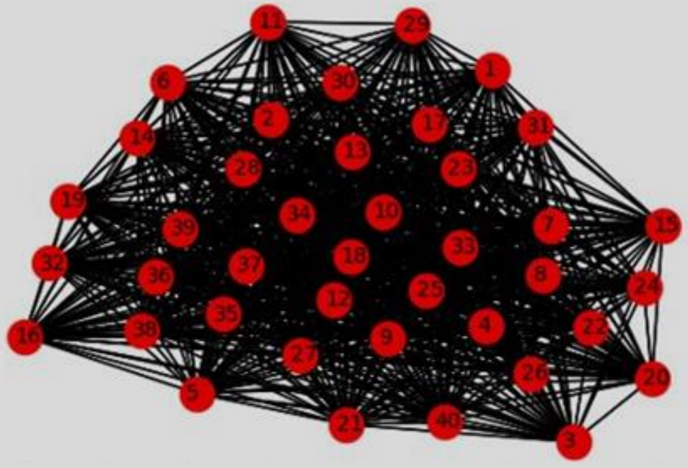
Introduction



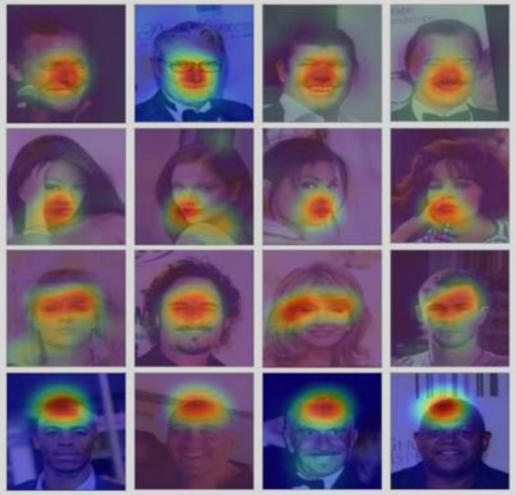
- Face attribute manipulation is an active research problem
- Labelled conditional GANs e.g. Stargan (CVPR'18,), Attgan (TIP'19), STGAN(CVPR'19) are successfully applied
- Encode target attributes in one-hot vector form
- Hand-engineered, no semantic information of attributes is embedded



One-hot Vector (ICASSP' 2020)



Word2Vec (ICASSP' 2020)



Attrbs-weights (CVPRW'18)

Semantic Representations of attributes

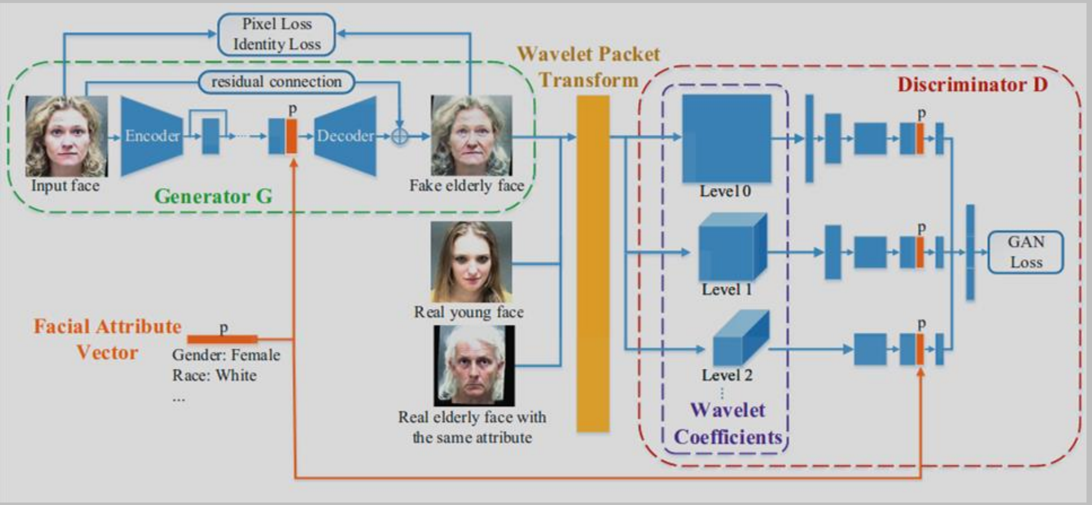
- STGAN (CVPR'19a) proposed to condition *t-s* instead *t*, its proven effective to improve attribute generation rate and other qualitative metrics
- Explored different conditioning mechanism: one-hot vector representation, semantic representations such as Word2Vec, Attrbs-weights
- These representations do not explicitly encode the co-occurences of the attributes

Conditioning both on generator and on discriminator

- Identified the issue of unnatural translation of target attributes due to lack of mechanism to retain the associated attributes of the target one

- Proposed to condition associated attributes (e.g. gender, race) in addition to main attribute (aging) both on Generator and on Discriminator to faithfully retain even after translation

- Hand-engineered, difficult to scale to arbitrary attributes manipulation



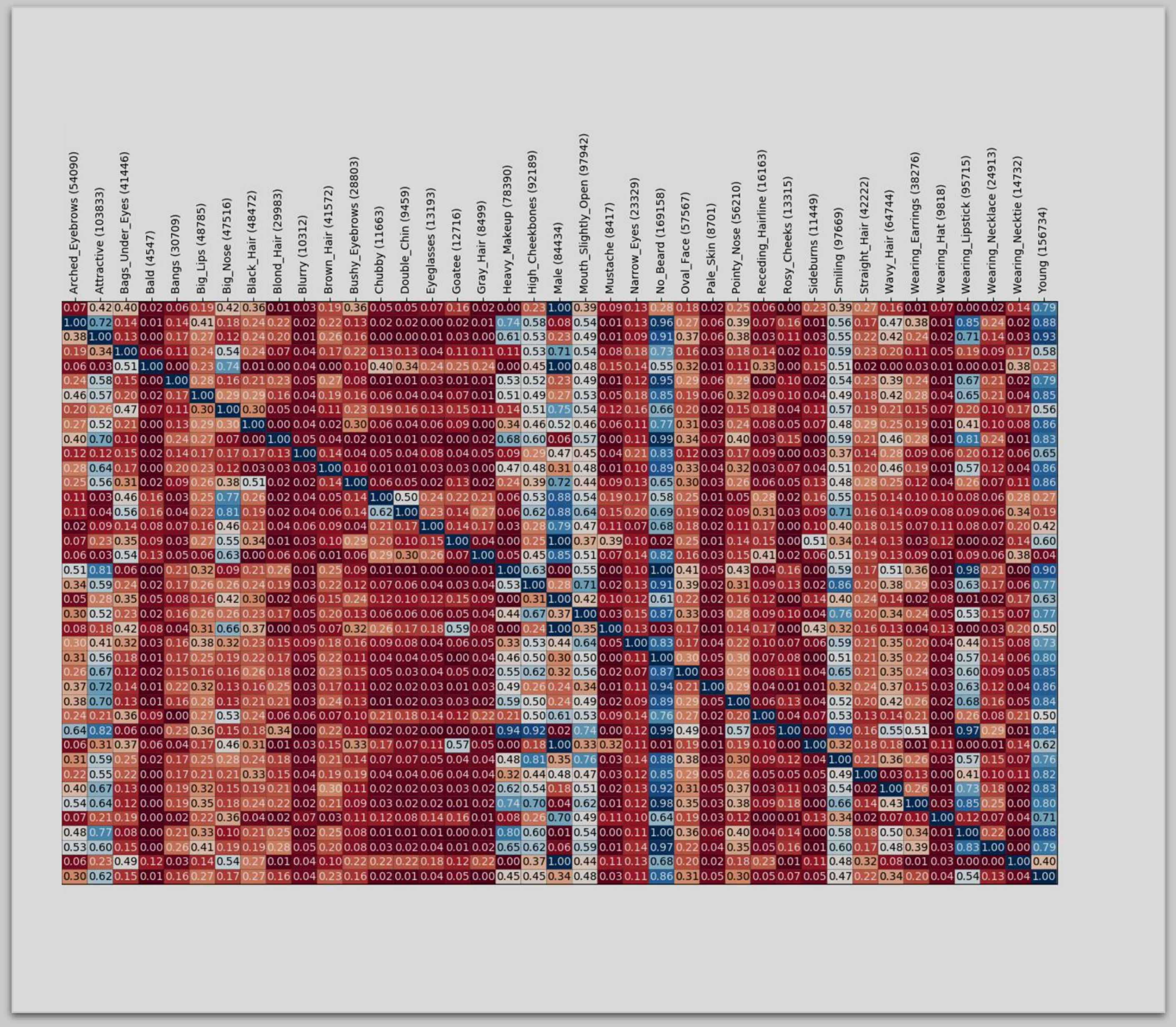
Attribute Aware Age Progression (CVPR'19b)

Key Idea

- Propose novel method to induce higher-order semantic representations of target attributes

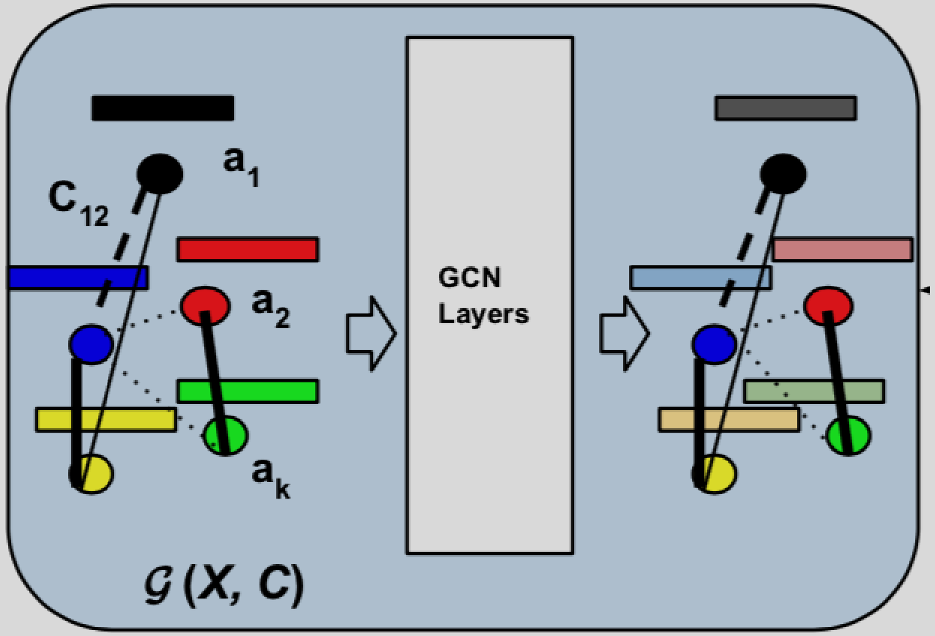
- Estimated co-occurrence probabilities from the training example and construct co-occurrence matrix

- Conditioning on both Generator and on Discriminator part



Graph Convolutional Network to induce higher order representation

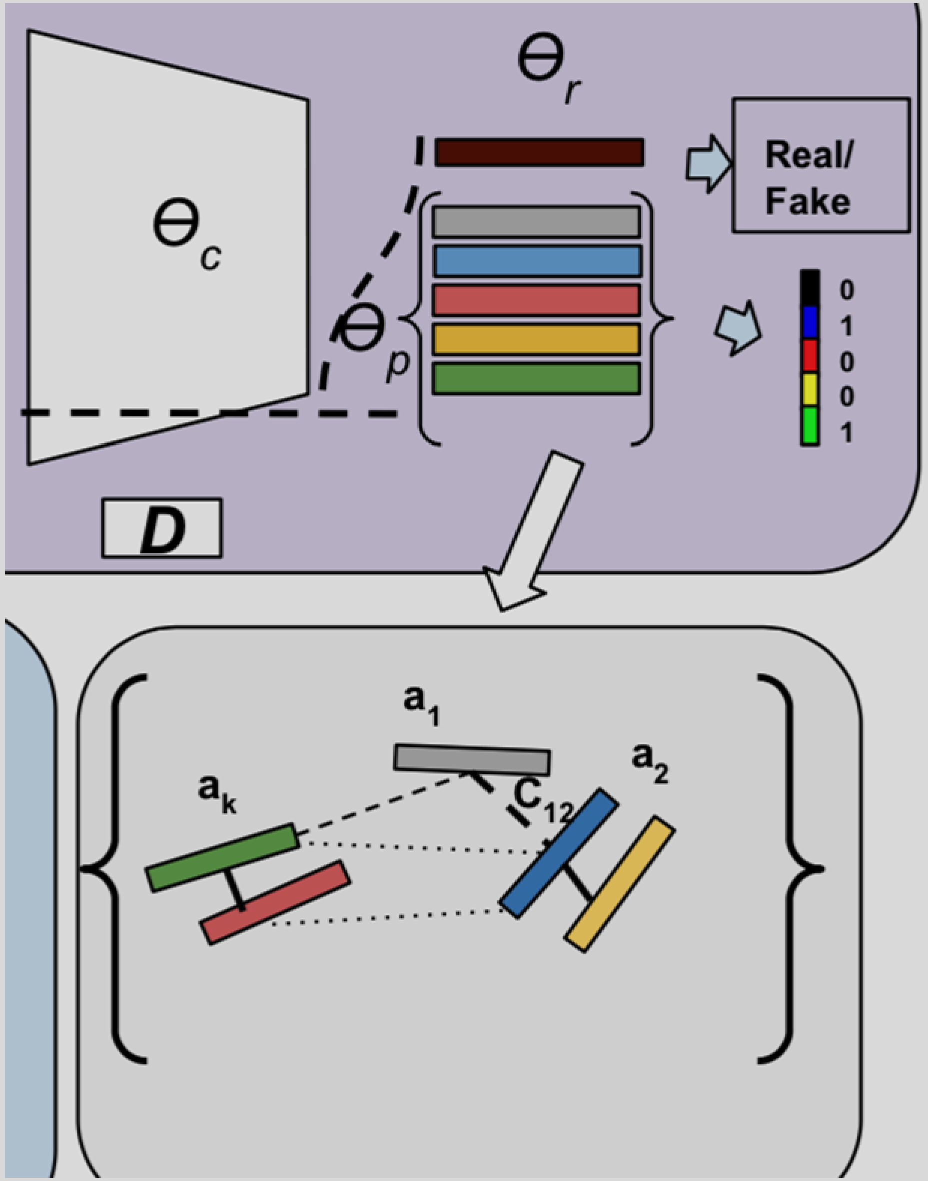
- Apply GCN framework similar to Kipf et al [ICLR'17]
- Each node of the Graph represents attribute specific information
- Edges encode relation between the attributes defined in adjacency matrix which we derive from the co-occurrence matrix
- Thickness of the edges indicate the probabilities of co-existing
- Apply Convolution operation to induce the higher order representations and feed to the generator



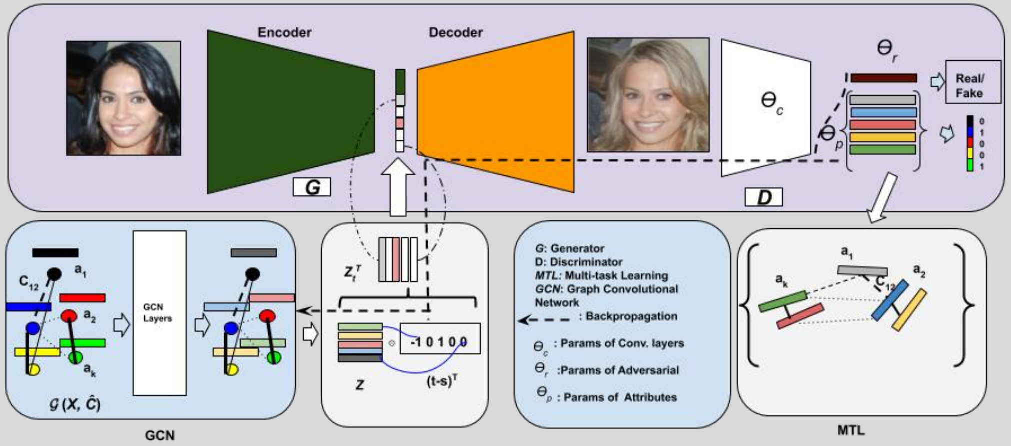
GCN

Multi-task learning on Discriminator

- Applied multi-task learning similar to Cavallanti et al [JMLR'10] on discriminator
- **Main idea:** if prediction on any attribute is wrong, update the model parameters of not only that attribute but also the related attributes
- Relations is derived from co-occurrence matrix as before
- Rate of update is determined by the magnitude of relation
- Satisfying such constraints induces similar model representations of related attributes



Proposed pipeline



- Upper part is regular cGAN
- Apply Graph Convolutional Network (GCN) on Generator part
- Do element wise multiplication between induced representations from graph by the difference of target and source one-hot vector similar to STGAN (CVPR'19a)
- (Multi-task Learning) MTL on discriminator part

Empirical evaluations

- Baseline architecture: Stargan
- Data set : CelebA
- TARR (Target Attributes Recognition Rate): Trained a classifier on real training examples and test on synthetic examples
- Evaluated on 5 attributes: *Black, Blonde, Brown hair, Gender, Age*
- Semantic attributes (word2vec and attrbs-weights) reprs. perform better

Condition Type	**Condition Mode** Std	Diff	**Average**
one-hot vec	✓		78.6
one-hot vec		✓	80.2
co-occurrence		✓	78.6
word2vec		✓	81.3
attrbs-weights		✓	81.9
gcn-reprs		✓	**84.0**

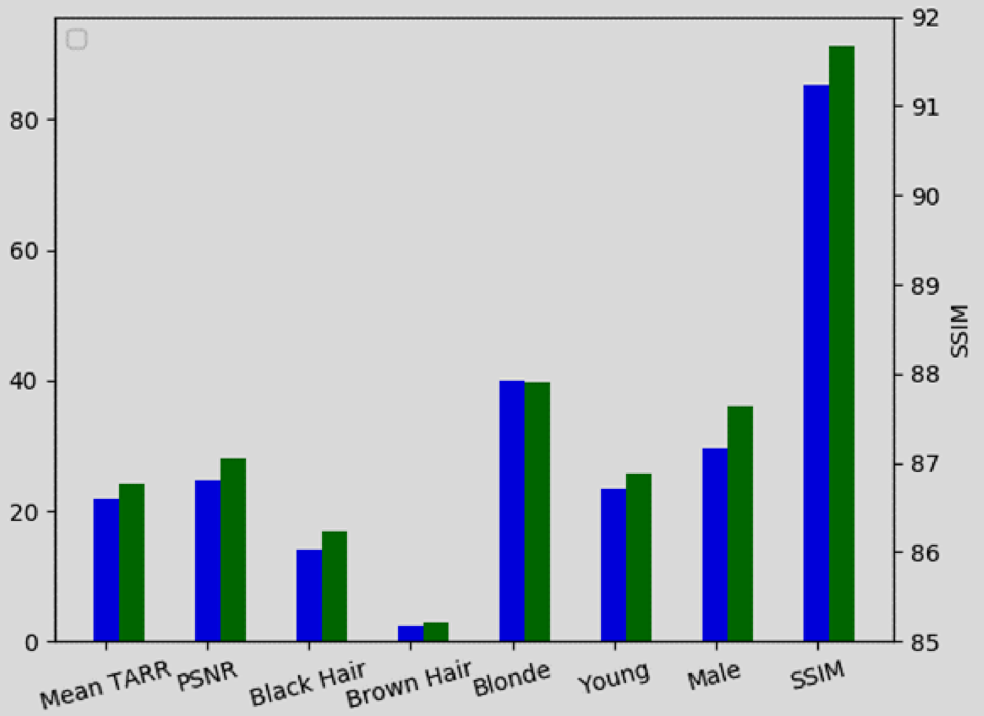
Quantitative Evaluations on CelebA

GAN Arch.	Condition Type	C. Mode: Std	C. Mode: Diff	Bald	Bangs	Black Hair	Blonde Hair	Brown Hair	B. Eyebrows	Eyeglasses	Mth. Slt. Open	Mustache	No beard	Pale Skin	Male	Young	Average
IcGAN [37]	*one-hot vec*	✓		19.4	74.2	40.6	34.6	19.7	14.7	82.4	78.8	5.5	22.6	41.8	89	37.6	43.2
FaderNet[25]	*one-hot vec*	✓		1.5	5	27	20.9	15.6	24.2	87.4	44	10	27.2	11.1	48.3	20.3	29.8
Stargan [9]	*one-hot vec*	✓		24.4	92.3	59.4	68.9	55.7	50.1	95.7	96.1	18.8	66.6	84	77.1	83.9	67.2
+ **MTL**	*one-hot vec*	✓		22.7	95.4	63	62.3	51.9	58	99.2	98.7	24	52.2	90.5	83.7	86.8	68.3
	one-hot vec		✓	41.9	93.6	74.7	75.2	67.4	65.9	99	95.3	26.8	64.3	86.2	89	89.3	74.5
	latent-reprs	✓		18.4	93.8	68.5	60.9	62.5	69.4	97.0	97.7	14.0	34.4	91.3	78.5	76.7	66.4
	latent-reprs		✓	32.5	93.2	68.9	79.5	71.5	55.3	97.2	98.4	30.0	58.5	85.1	84.0	75.1	71.4
	attrbs-weights		✓	32.7	96.0	74.4	77.5	74.1	66.7	98.8	32.2	78.2	90.9	81.6	98.5	86.7	76.0
	gcn-reprs		✓	28.2	**99.4**	**76.5**	77.1	70.9	74.2	**99.5**	99.4	37.3	89.6	92	93.4	94.9	79.4
+ **MTL**	*gcn-reprs*		✓	34.4	98.4	73.3	78.6	70.8	**85.5**	**99.5**	99.1	44.2	**90**	92.3	**95.6**	91.7	81.0

Evaluations on LFWA

- Blue Bar: Baseline, Green Bar: Our Approach

- Consistently outperforming the counter-part method



Qualitative Comparisons

Qualitative Comparisons

Female

Female

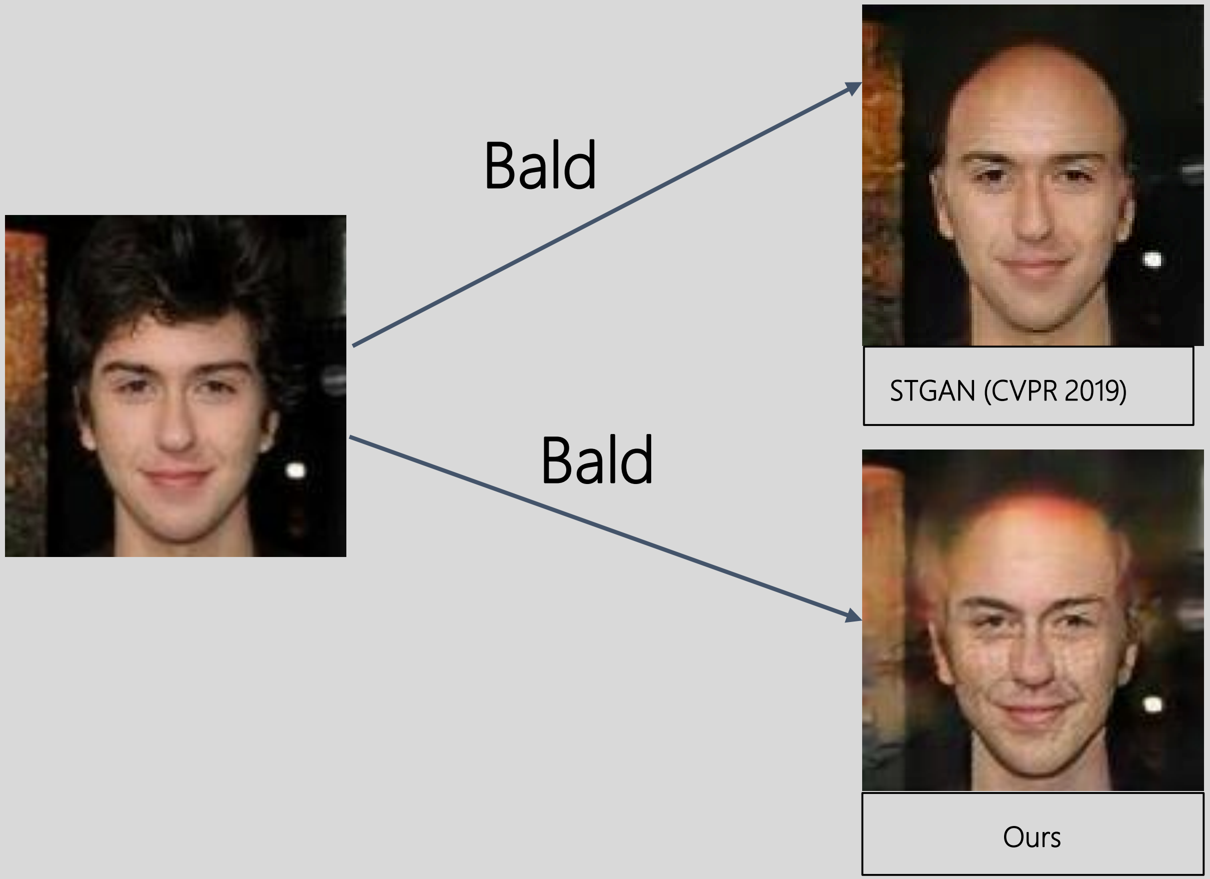
Attgan, IEEE TIP 2019

Beard not thinned yet

Ours

Beard is thinned, Arched eyebrows, lipsticks

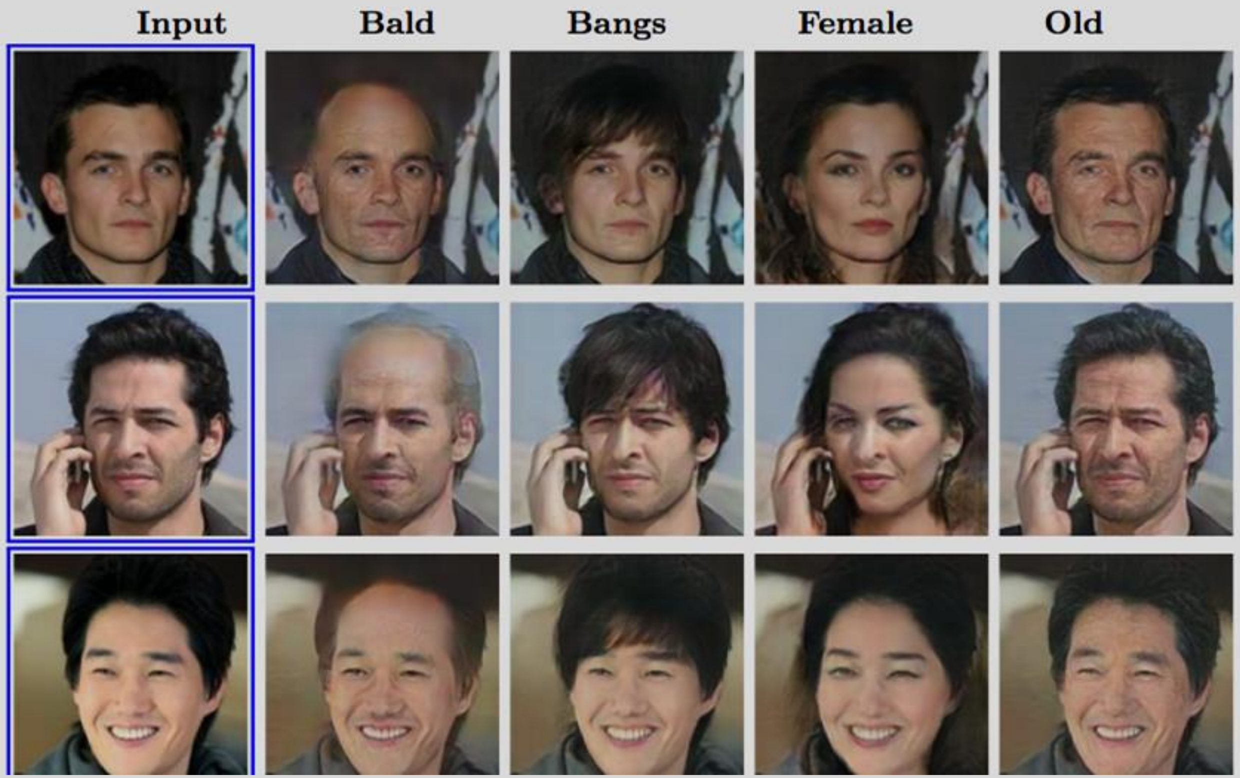
Qualitative Comparisons



A young guy turning complete bald=> unnatural and rare ?

A wrinkled face, old guy, few remaining few grey hairs and turning bald

Qualitative Results



- **Main attribute**: Bald**, Associated attributes**: wrinkles on face, few remaining grey hairs
- **Main attribute**: Old, **Associated attributes**: Wrinkles, Receding Hairlines, Grey hairs

Conclusions and Future Works

- We propose a framework to induce higher-order representations of target label for conditional GAN
- We applied Graph Convolutional Network on Generator side whereas multi-task learning on Discriminator
- Empirical evaluation demonstrates the improvement in the accuracy of the proposed method compared to the existing arts
- Qualitative evaluations shows natural translation
- In future, we explore the method to learn the edge information of the graph to synthesise naturally translated images

References

- [CVPR’18] Choi, Yunjey, et al. "*Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.*"
- [CVPR’19a] Liu, Ming, et al. "*STGAN: A unified selective transfer network for arbitrary image attribute editing.*"
- [CVPR’19b]] Liu, Yunfan, Qi Li, and Zhenan Sun. "*Attribute-aware face aging with wavelet-based generative adversarial networks.*"
- [ICASSP’20] Bhattarai, Binod, Rumeysa Bodur, and Tae-Kyun Kim. "*Auglabel: Exploiting word representations to augment labels for face attribute classification.*"
- [CVPRW’18] Taherkhani, Fariborz, Nasser M. Nasrabadi, and Jeremy Dawson. "*A deep face identification network enhanced by facial attributes prediction.*"
- [ICLR’17] Kipf, Thomas N., and Max Welling. "*Semi-supervised classification with graph convolutional networks.*"
- [JMLR’10] Cavallanti, Giovanni, Nicolo Cesa-Bianchi, and Claudio Gentile. "*Linear algorithms for online multitask classification.*"
- [TIP’19] He, Zhenliang, et al. "*Attgan: Facial attribute editing by only changing what you want.*"

Sampling Strategies for GAN Synthetic Data

Binod Bhattarai, Seungryul Baek, Rumesya Bodur, Tae-Kyun Kim
Imperial College London

Introduction

Sad Fear Surprise Happy

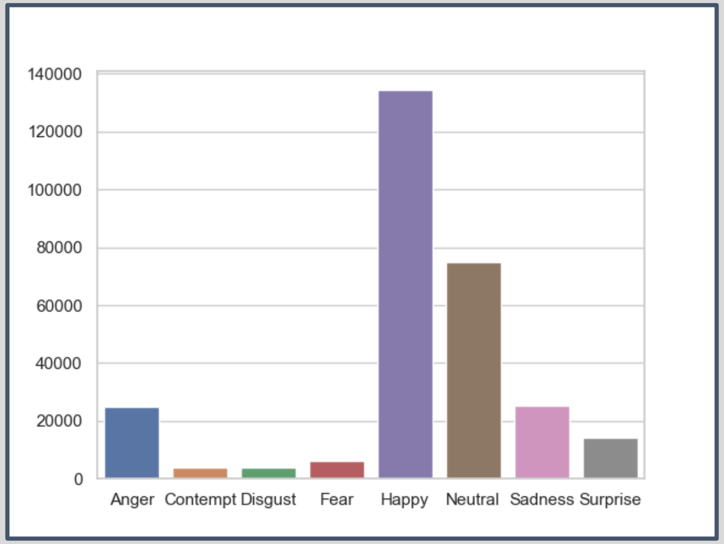


Fig. Distribution of annotated examples on AffectNet

- Uneven distribution of annotated examples.
- Some categories even lack sufficient annotated examples.
- Data augmentation has been crucial for the success of deep learning framework [a]
- Geometric transformations of an image cropping, flipping, rotation, shearing are commonly used to generate new annotated examples.
- Recently, GANs synthetic are being used to augment the real data set

[a] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks."NeurIPS. 2012.

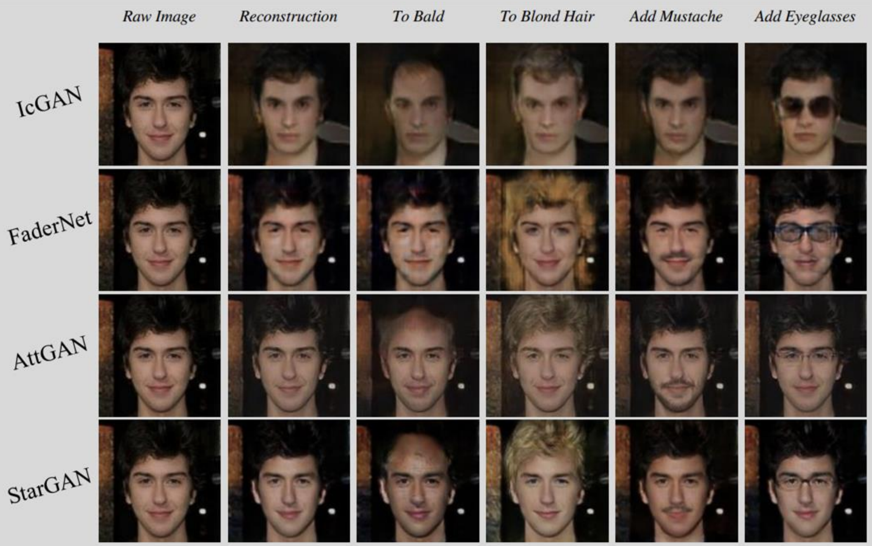
Imperial College London

Motivation: Issues with Geometric Transformation

- [a] identified the issue with applying a common set of geometric transformation to every classes
  - Rotational invariance is poorly suited while dealing with certain classes such as 6 or 9 in MNIST.
- AutoAugment [b] proposed to learn the class specific geometric transformation using RL to sub-sample from large pool of geometrically transformed synthetic image.
- Our work is focused in sub-sampling GAN synthetic data
- Advantages of GAN synthetic data: i) Images from different categories can be translated to a target category ii) Geometric Transformation can be applied

[a] Hauberg, Søren, et al. "Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation." *Artificial Intelligence and Statistics*. 2016.
[b] Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation strategies from data." *CVPR 2019*

Motivation: Issues in GAN synthetic examples



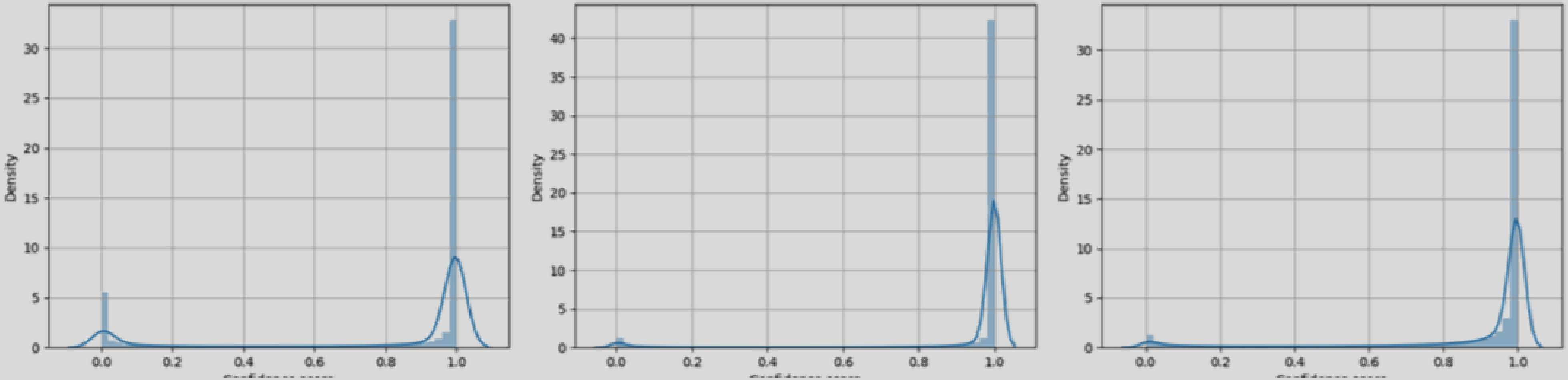
- Generates photorealistic synthetic examples.
- Randomly augmenting synthetic examples with real data [a,b] for face analysis task is getting popular.
- Not examined yet if all synthetic examples are equally important.

a. Gecer, B., Bhattarai, B., Kittler, J., & Kim, T. K.. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model. ECCV, 2018

b. Zhao, Jian, et al. "Dual-agent gans for photorealistic and identity preserving profile face synthesis." *NIPS 2017*

Imperial College London

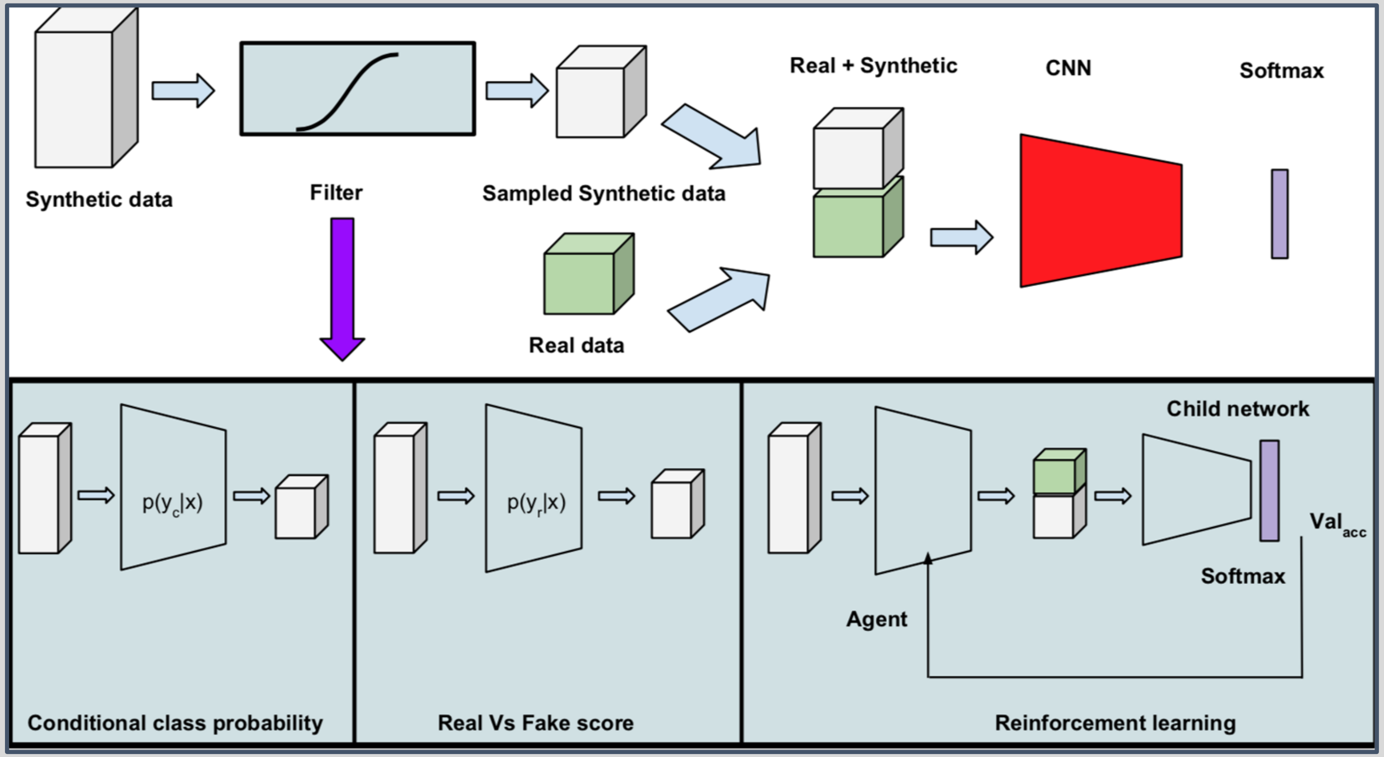
Motivation: Distribution of target label confidence score on synthetic examples of Affectnet



- Computed target label confidence score on synthetic data
- Order of target label is: Contempt, Disgust, Fear
- Large fraction of synthetic data preserve label with very low confidence

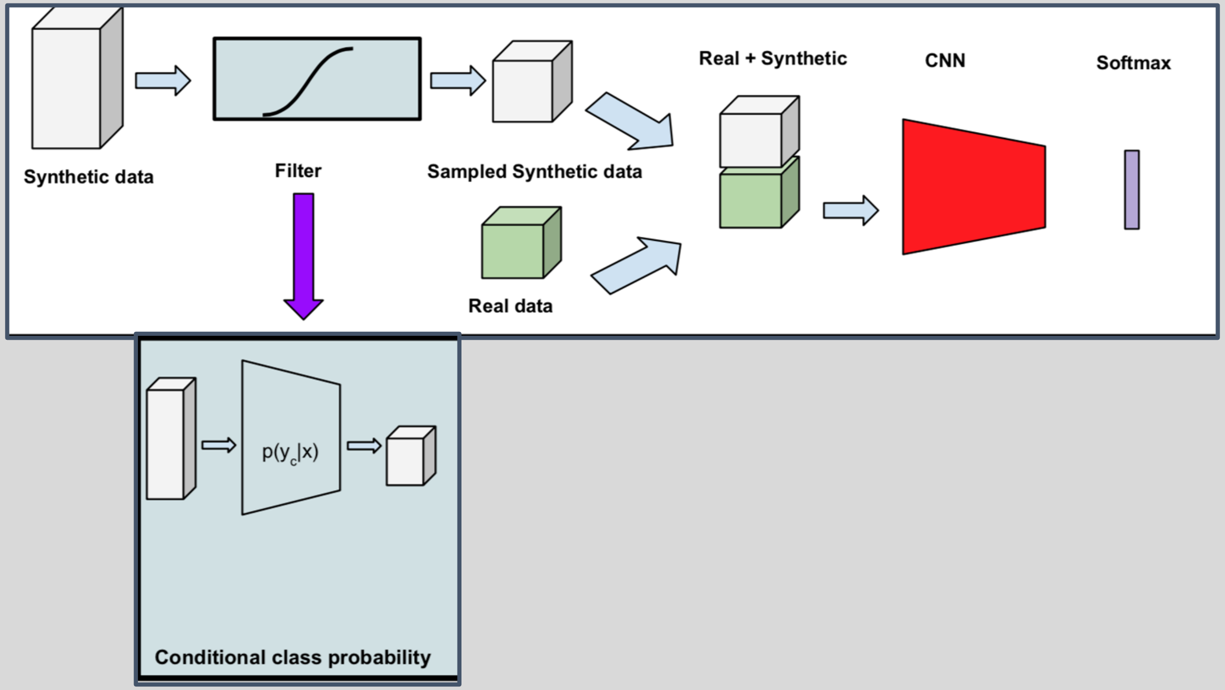
Imperial College London

Proposed method: 3 data sampling methods.



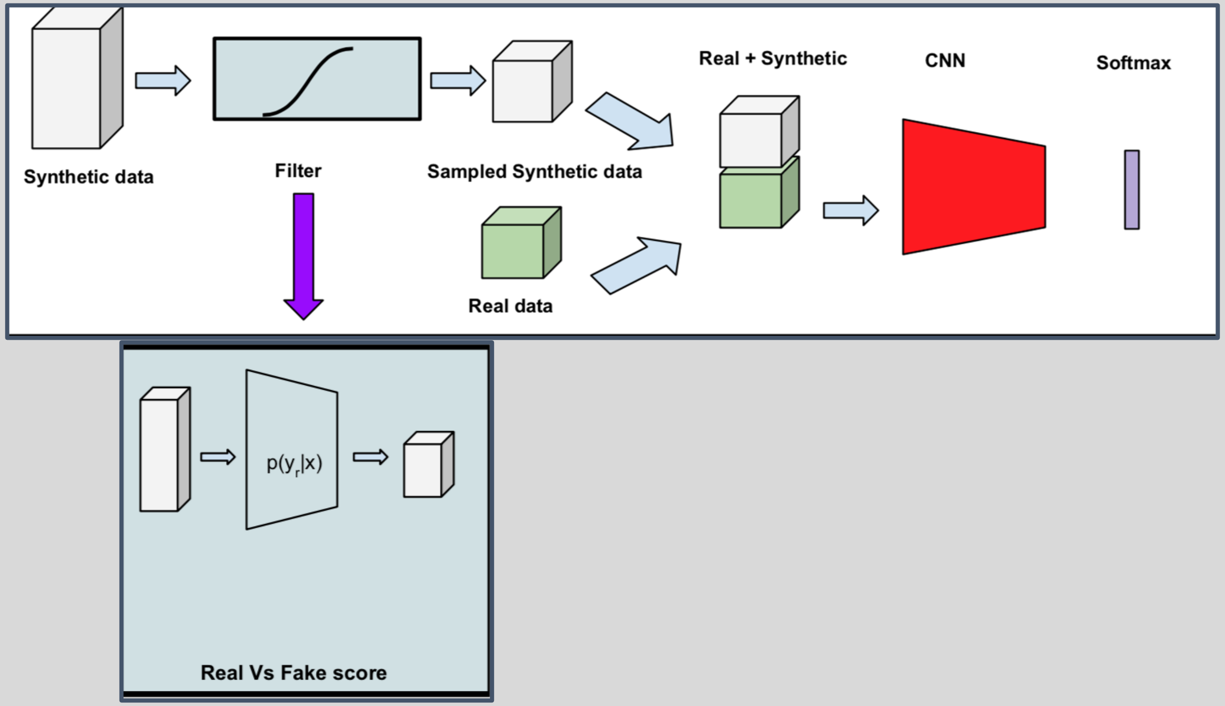
- Three different data sampling strategies based on 1) target class confidence score (cl-sam), 2) confidence on realism (cr-sam) and 3) reinforcement learning (RL)
- Evaluated independently to compare their impact

Confidence score based sampler (cl-sam)



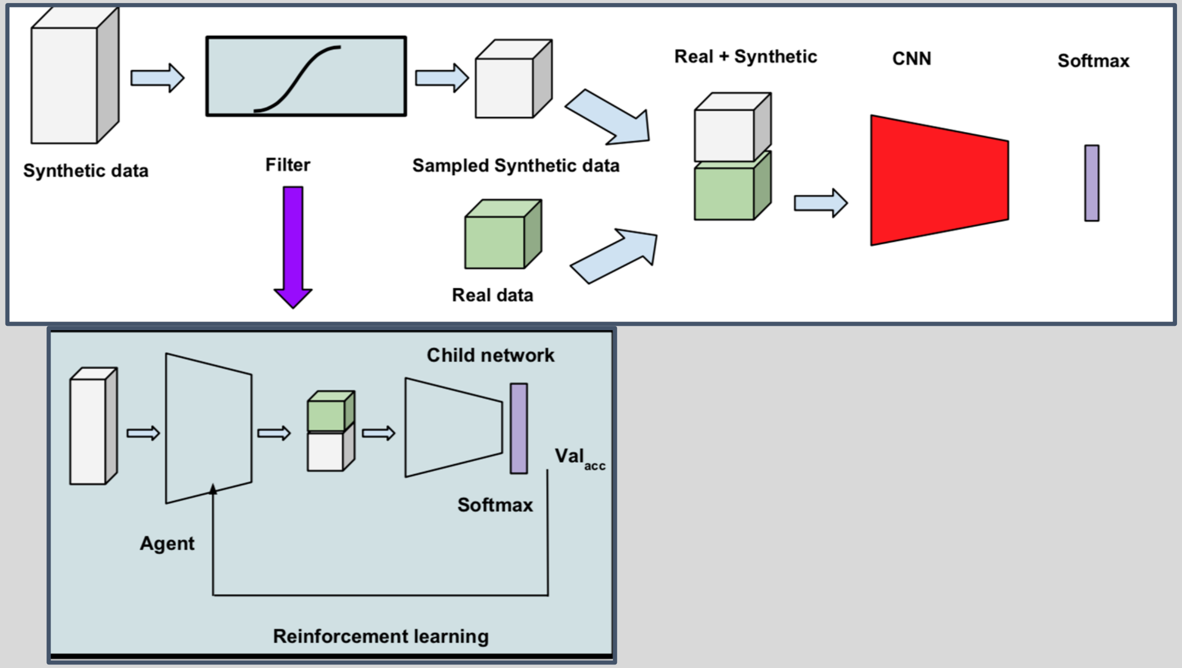
- Use Discriminator of the GAN to predict class label
- Ranked synthetic examples based on target class label confidence
- top-K ranked examples used to train the classifier

Realism score based sampler (cr-sam)



- Use Discriminator of the GAN to predict the real vs fake score
- Ranked synthetic examples based on realism confidence
- top-K ranked examples used to train the classifier.

RL-based Sampler



- RL : Trained an Actor-Critic RL framework to learn the policy to predict whether to augment a synthetic example or not given synthetic image
- Parameters were learned to maximise the reward
- Reward is computed from the validation score from the child network

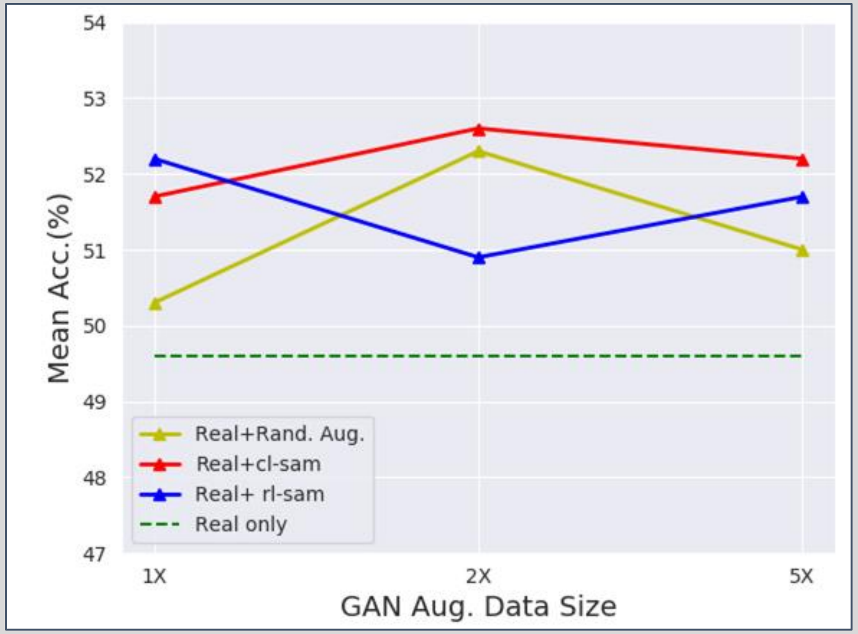
Imperial College London

Experiments

- **Data sets**: CelebA (Attributes) and Affectnet (Expression)
- **Evaluation metric**: Mean Accuracy
- **Compared methods**:
  - Random Augmentation: Most common augmentation technique
  - cr-sam : Data sampled based on confidence on realism
  - cl-sam: Data sampled based on confidence on target class label
  - RL : An agent trained to predict augment/not augment given synthetic example
- **Synthetic image generator**: StarGAN
  - Can be applied with any other GANs

Quantitative Results on Affectnet

Architecture	Resolution	Mean. Acc.	Aug.	Type
AlexNet [31]	$224 \times 224 \times 3$	**50.0**	$0\times$	No aug.
ResNet-50	$64 \times 64 \times 3$	46.1	$0\times$	No aug.
ResNet-50	$128 \times 128 \times 3$	49.6	$0\times$	No aug.
ResNet-50	$128 \times 128 \times 3$	50.3	$1\times$	Random
ResNet-50	$128 \times 128 \times 3$	51.7	$1\times$	cl-sam
ResNet-50	$128 \times 128 \times 3$	**52.2**	$1\times$	cr-sam
ResNet-50	$128 \times 128 \times 3$	52.3	$2\times$	Random
ResNet-50	$128 \times 128 \times 3$	**52.6**	$2\times$	cl-sam
ResNet-50	$128 \times 128 \times 3$	50.9	$2\times$	cr-sam
ResNet-50	$128 \times 128 \times 3$	51.0	$5\times$	Random
ResNet-50	$128 \times 128 \times 3$	**52.2**	$5\times$	cl-sam
ResNet-50	$128 \times 128 \times 3$	51.7	$5\times$	cr-sam
ResNet-50	$128 \times 128 \times 3$	51.8	$2.6\times$	RL



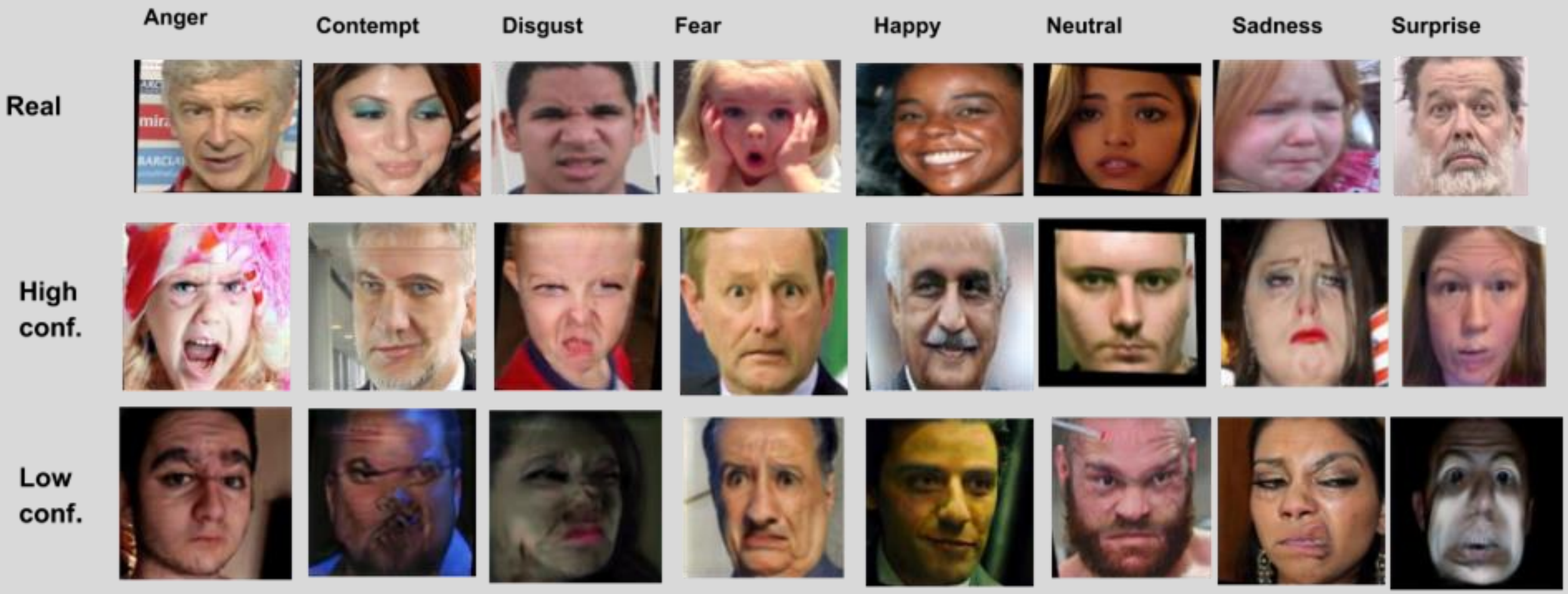
- 88K of training examples
- Augmented by different size of synthetic data
- Consistently outperforms the baseline
- RL sub-sampled synthetic data of size 2.6X of real data from the pool of 7X

Quantitative Results on CelebA

Architecture	Resolution	Mean. Acc.	Aug.	Type
ResNet-50	$64 \times 64 \times 3$	90.3	$0\times$	No aug.
ResNet-50	$64 \times 64 \times 3$	91.0	$5\times$	Random
ResNet-50	$64 \times 64 \times 3$	91.1	$5\times$	cl-sam.

- A popular benchmark with 160K training examples
- Our approach further improved the performance of the commonly used technique

Experiments (Qualitative)



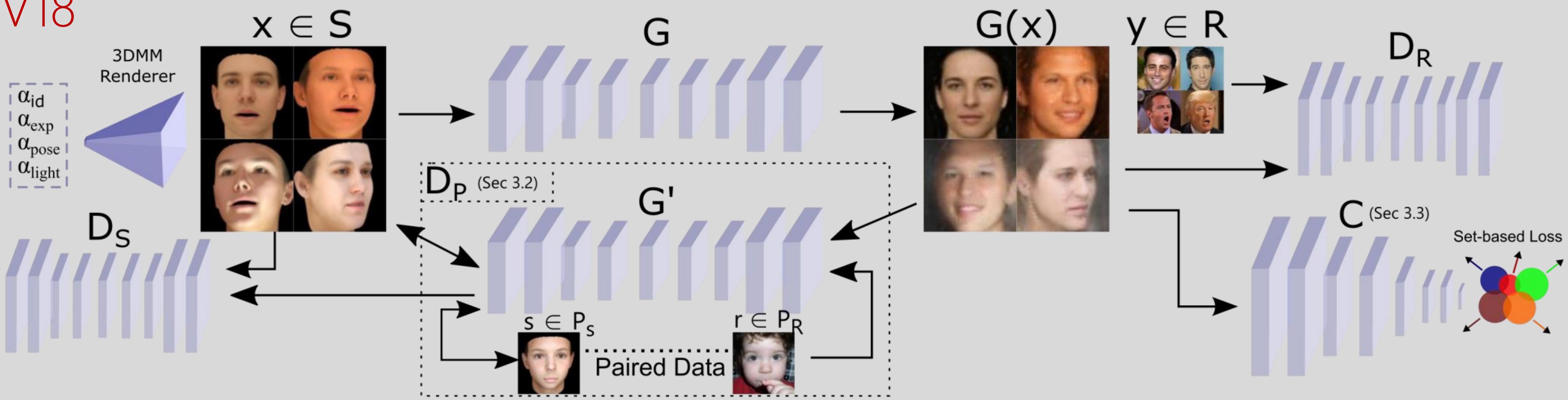
- Sampled synthetic images are closer to real images.
- Discarded images look quite different in terms of illumination and have more artifacts.

Conclusions

- We evaluated three different sampling strategies over commonly used augmentation techniques. We propose to use confidence score, realism score and RL based sampler to find a meaningful subset.
- From our extensive experiments, we observed that these three techniques outperform the commonly used random augmentation technique.
- Among these three, we observed that the class conditional and realism based methods are both efficient and accurate, RL is accurate but computationally expensive

Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3DMM

ECCV18



- Randomly generated 3DMM images with random pose, expression and lighting attributes for the new IDs.
- Unsupervised training with forward cycle consistency.
- Adversarial Pair Matching network G′ by the help of a limited number of paired data.
- ID preservation by a set-based supervision through a pretrained classification network C.

$$\mathcal{L}_{cyc} = \mathbb{E}_{x\in\mathcal{S}}\|G'(G(x)) - x\|_1 \quad (1)$$

$$\mathcal{L}_{G} = \mathbb{E}_{x\in\mathcal{S}}\|G(x) - D_R(G(x))\|_1 \quad (2)$$

$$\mathcal{L}_{G'} = \mathbb{E}_{x\in\mathcal{S}}\|G'(G(x)) - D_S(G'(G(x)))\|_1 \quad (3)$$

$$\mathcal{L}_{D_R} = \mathbb{E}_{x\in\mathcal{S},y\in\mathcal{R}}\|y - D_R(y)\|_1 - k_t^{D_R}\mathcal{L}_G \quad (4)$$

$$\mathcal{L}_{D_S} = \mathbb{E}_{x\in\mathcal{S}}\|x - D_S(x)\|_1 - k_t^{D_S}\mathcal{L}_{G'} \quad (5)$$

$$k_t^{D,G} = k_{t-1}^{D,G} + 0.001(0.5\mathcal{L}_D - \mathcal{L}_G)$$

Identity Preservation

$$\mathcal{L}_C = \mathbb{E}_{x \in \mathcal{S}, i_x \in \mathbb{N}^+} \sum_{x}^{M} -log \frac{\exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_{i_x}\|_2^2 - \eta)}{\sum_{j \neq i_x} \exp(\frac{1}{2\sigma^2}\|C(G(x)) - c_j\|_2^2)}$$

- While the quality of images is being improved during the training,
- Their projection on the embedding space is shifting.
- Centre/pushing losses is used to adapt to those changes

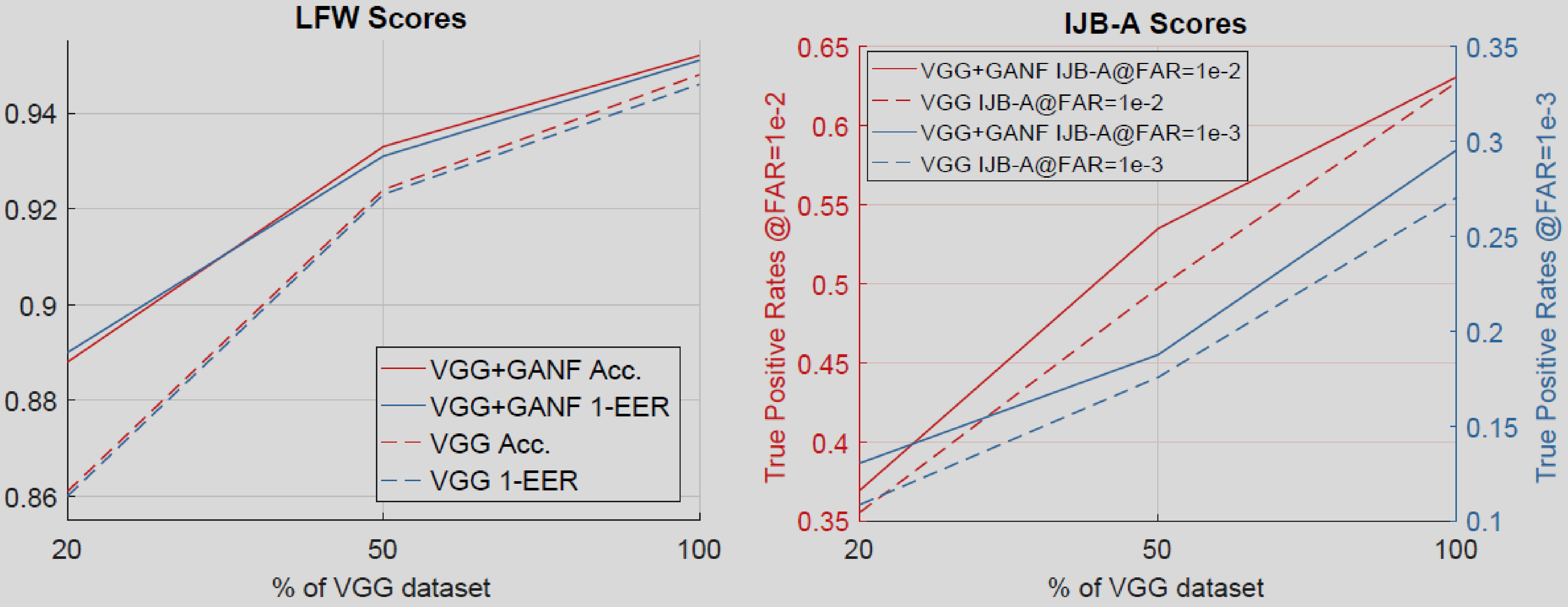
➢ Quality of 9 images of 3 identities (per row) during the training. Background plot shows the error by the proposed identity preservation layer over the iterations. Notice the changes on the level of fine-details on the faces which is the main motivation of using set-based identity preservation.

GANF aces

- ➢ Random samples from GANFaces dataset. Each row belongs to same identity. Notice the variation in pose, expression and lighting.

Quantitative Results / Less Real Data

- Contribution of GANFaces is more visible when a percentage of real data is



BMVC 2018
Newcastle upon Tyne

Multi-task Deep Network for depth-based 6D object Pose and Joint Registration in Crowd Scenarios

Juil Sock[1], Kwang-In Kim[2], Caner Sahin[1] and Tae-Kyun Kim[1]

[1]Imperial College London & [2]University of Bath

This work was possible thanks to Samsung Research

Problem Statement

Problem：

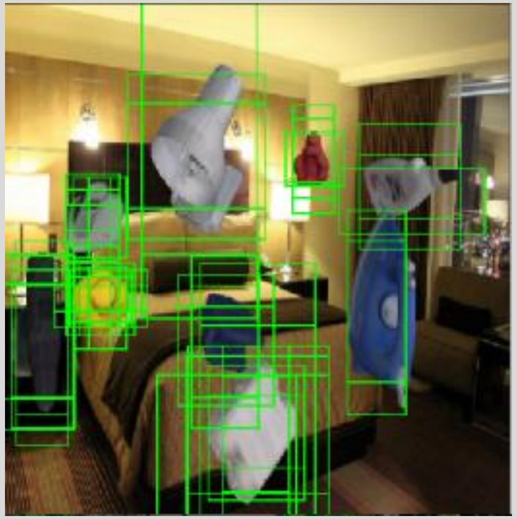
- Most of the current pose estimation system consists of a sequence of separate, independent component : object detection, pose estimation and refinement.
- The performance drops largely when the environment is crowded and cluttered as the training dataset does not contain occlusion.

Idea：

- Build a network that jointly performs object detection, 6D object pose estimation, and joint registration.
- Synthetic images are generated with physics simulation to capture the realistic occlusion pattern for training.

Related work

- Most previous works considers relation between object candidate hypothesis at testing stage only(e.g. NMS for detection network), and the training of their pose estimators is agnostic to such occlusion behaviors: They are trained on isolated object instances.
- Previous works either performs different components separately or use approximation to infer 6D parameters based on assumption that there is no occlusion.
  - SSD6D estimate 3d translation indirectly from bounding box. Only works when tight bounding box is recovered from isolated unoccluded objects.
  - BB8[2] performs detection and pose estimation separately.



Sample image of training data from SSD6D[1]

Sample image of training data from BB8[2]

[1] SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again, Kehl et al., ICCV'17
[2] BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth

Training data generation

Environment model

Object model

Object pose generation (Mujoco)

Object pose

Mesh generation (Blender)

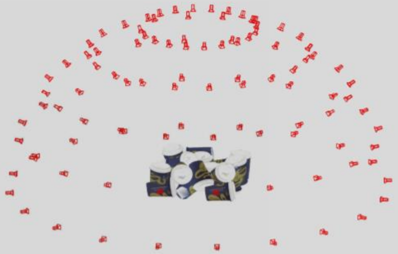
Mesh model

Ambient condition

Camera model

RGBD Rendering (OpenGL)

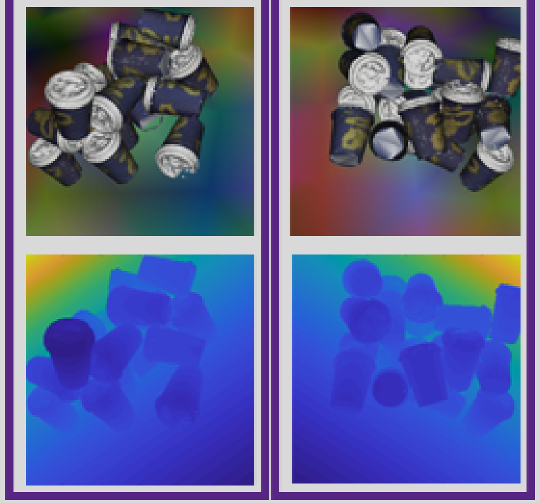
Visualization of simulation(video)



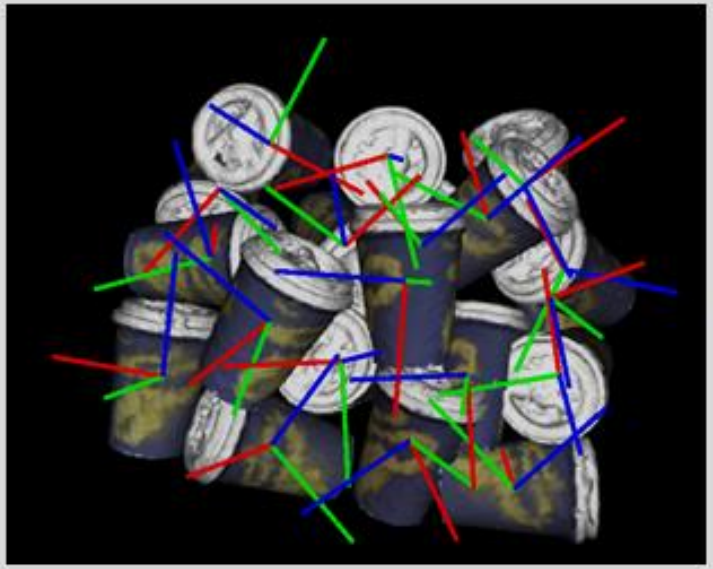
View-point sampling

Rendered RGB

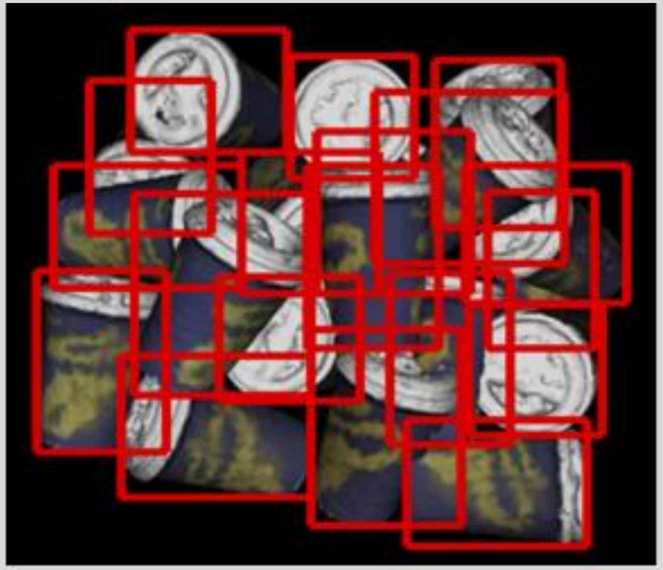
Rendered Depth



Training data generation


Pose annotations for objects


Visibility mask for objects

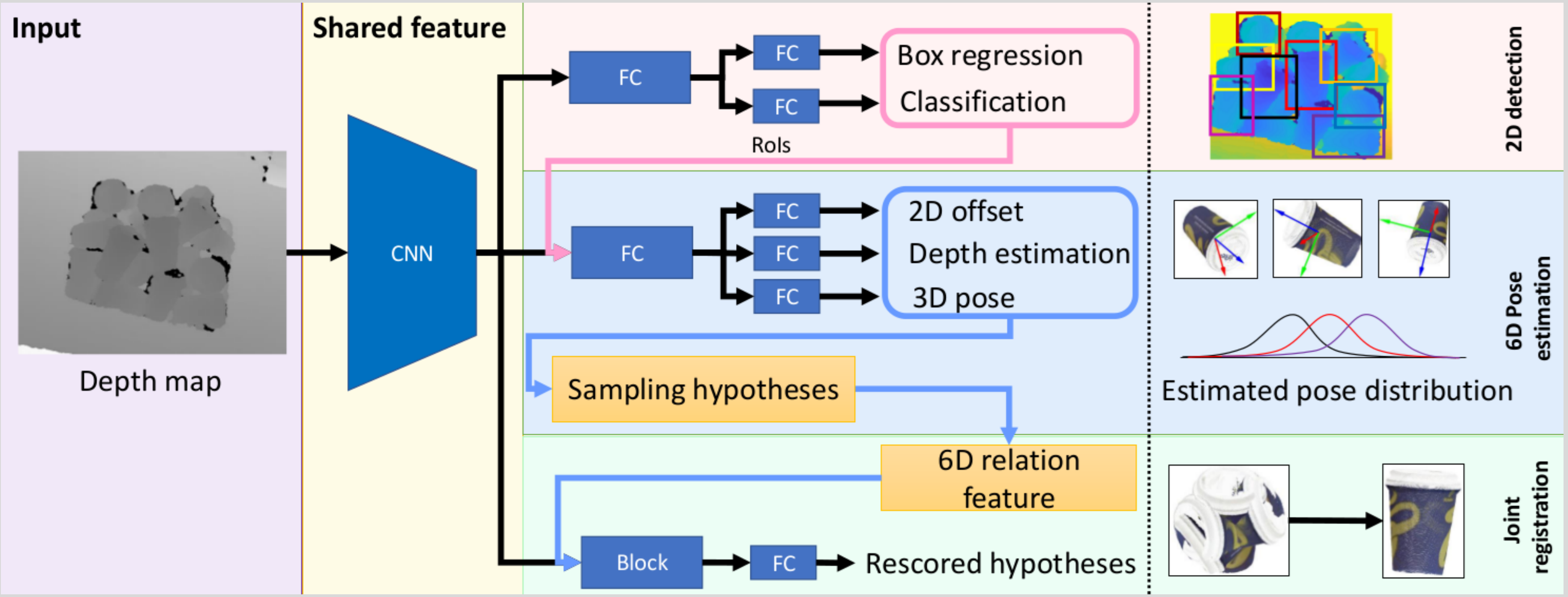

Bounding boxes for objects

Training data generation

Framework

- Modular design
  - There are 3 modules responsible for 3 different tasks required for estimating pose of multiple objects:
    - Object bounding box detection
    - Object pose estimation
    - Joint registration
  - Proposed system optimizes the following loss function:

$$L = L_{Det} + L_O + L_D + L_P + L_J$$



Framework

- L_{Det} : Region Proposal Network(RPN) detects objects and regresses bounding boxes around detected object.
- L_O : 2D object center estimation(x,y) regresses the offset between the object center and the bounding box corner estimated from RPN.
- $L_{P(D)}$: Under severe occlusion, regressing a single reliable pose estimate is challenging. Both 3D pose(roll, pitch, yaw) and depth(z) are formulated as classification which allows us to sample multiple hypotheses.
- L_J : 6D pose estimation module generates a pool of hypotheses which contains multiple false positives. Joint registration classifies each hypothesis into false positive or true positive.

Detection

Correct detection

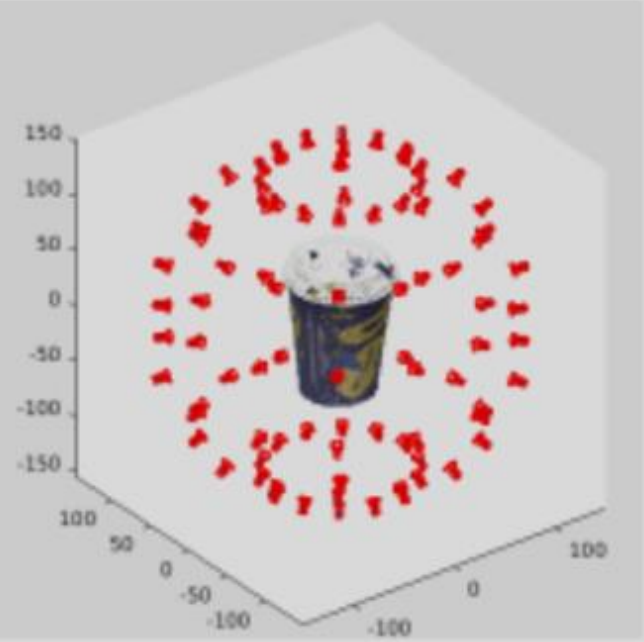
False positives

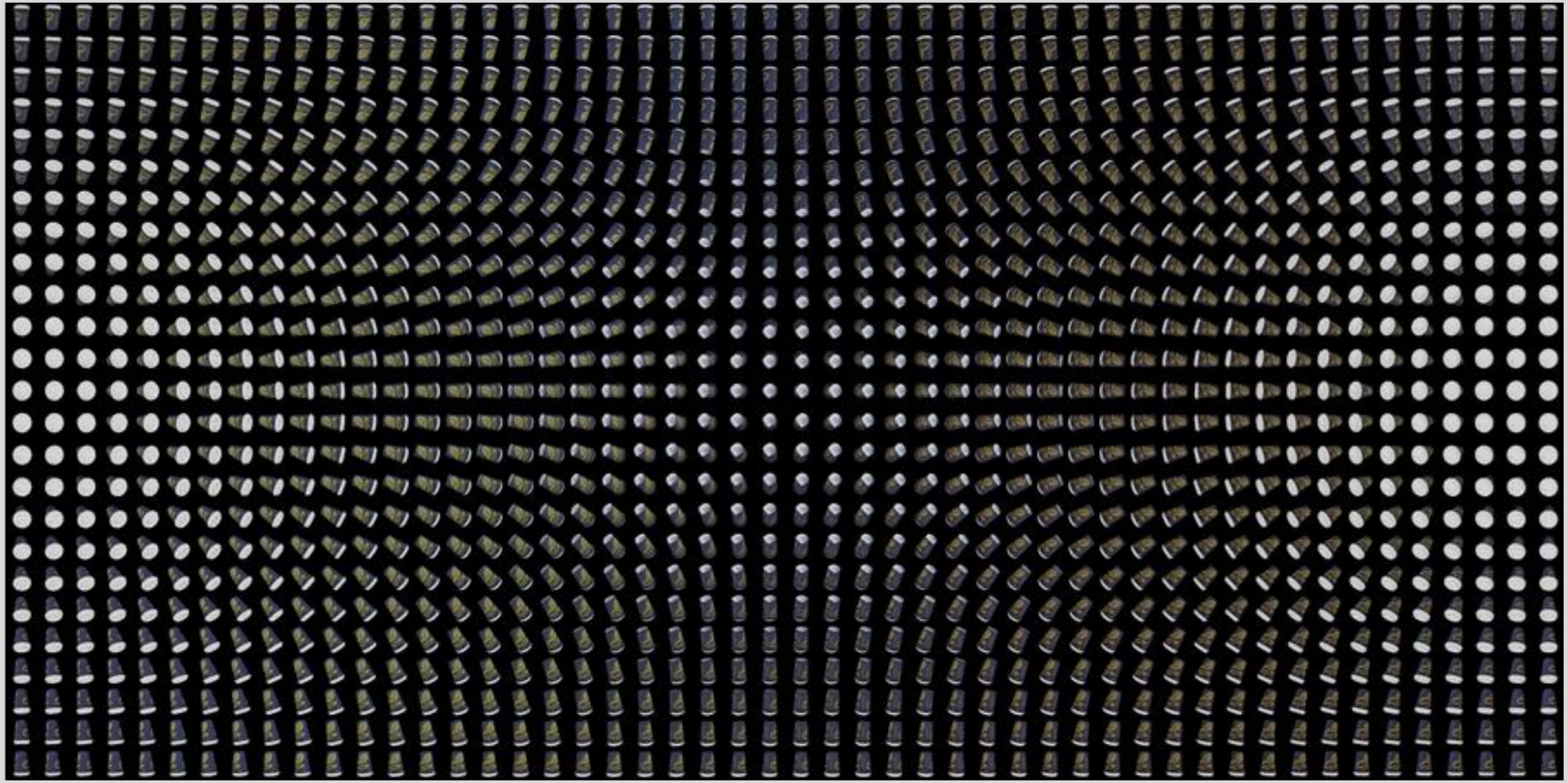
False negative

Framework

- L_{Det} : Region Proposal Network(RPN) detects objects and regresses bounding boxes around detected object.
- L_O : 2D object center estimation(x,y) regresses the offset between the object center and the bounding box corner estimated from RPN.
- $L_{P(D)}$: Under severe occlusion, regressing a single reliable pose estimate is challenging. Both 3D pose(roll, pitch, yaw) and depth(z) are formulated as classification which allows us to sample multiple hypotheses.
- L_J : 6D pose estimation module generates a pool of hypotheses which contains multiple false positives. Joint registration classifies each hypothesis into false positive or true positive.

Pose estimation





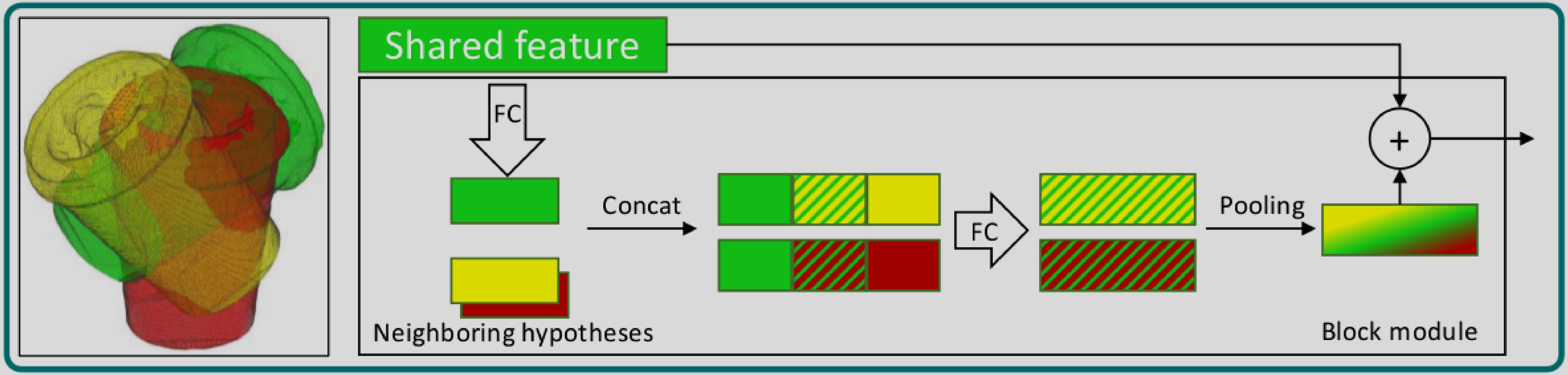
- Figure above visualises pose classes.
- Unlike regression, classification can provide multiple pose hypothesis.

Framework

- L_{Det} : Region Proposal Network(RPN) detects objects and regresses bounding boxes around detected object.
- L_{O} : 2D object center estimation(x,y) regresses the offset between the object center and the bounding box corner estimated from RPN.
- $L_{P(D)}$: Under severe occlusion, regressing a single reliable pose estimate is challenging. Both 3D pose(roll, pitch, yaw) and depth(z) are formulated as classification which allows us to sample multiple hypotheses.
- L_{J} : 6D pose estimation module generates a pool of hypotheses which contains multiple false positives. Joint registration classifies each hypothesis into false positive or true positive.

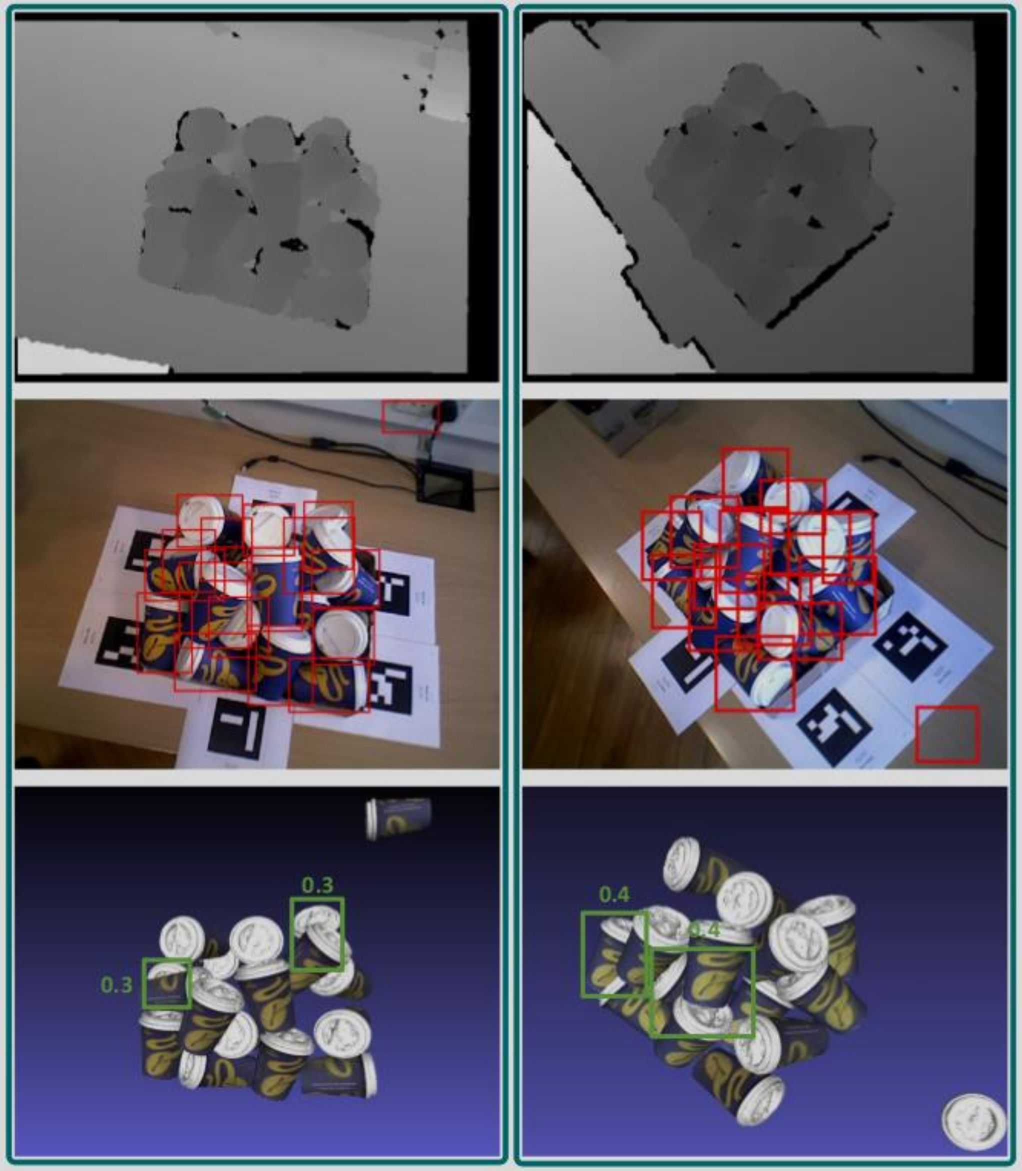
Joint registration

- Inspired by GossipNet[1] architecture which models and learn relational feature between hypotheses.
- Cast selection problem as classification problem as in [2].



[1] Learning non-maximum suppression, Hosang et al., CVPR'17
[2] A Global Hypotheses Verification Method for 3D Object Recognition, Aldoma et al., ECCV'12

Experiments



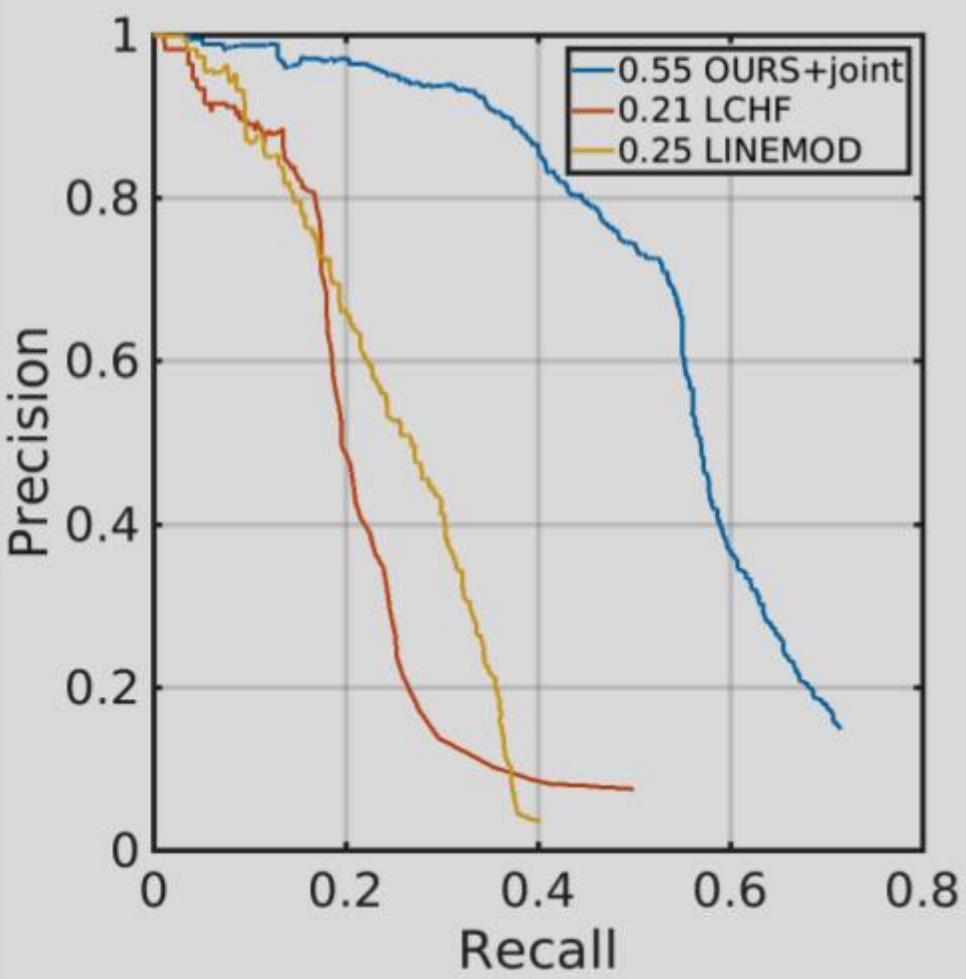
Experiments

Input

LINMO
D

LCHF

Ours



Summary

- A model can jointly learn multiple tasks without harming the performance: detection, 3D localisation, orientation estimation and joint registration.
- A pipeline to generate synthetic dataset with varying level of occlusion is proposed.

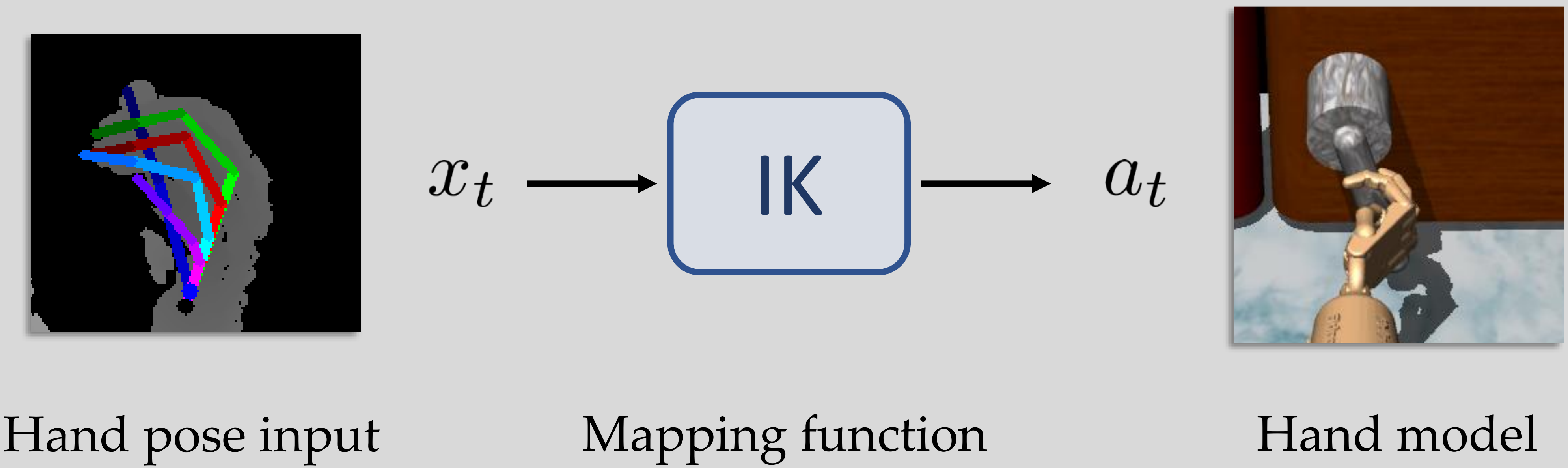
IROS 2020

Imperial College London
NIANTIC
KAIST

Physics-Based Dexterous Manipulations with Estimated Hand Poses and Residual Reinforcement Learning

Guillermo Garcia-Hernando
Edward Johns
Tae-Kyun Kim

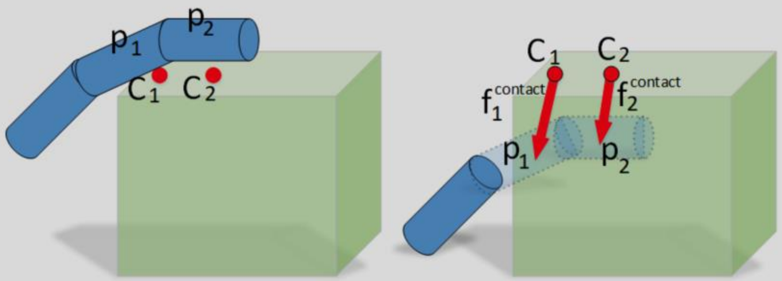
This work was done at Imperial and it was possible thanks to **Samsung Research**


x_t
IK
a_t
Hand pose input
Mapping function
Hand model

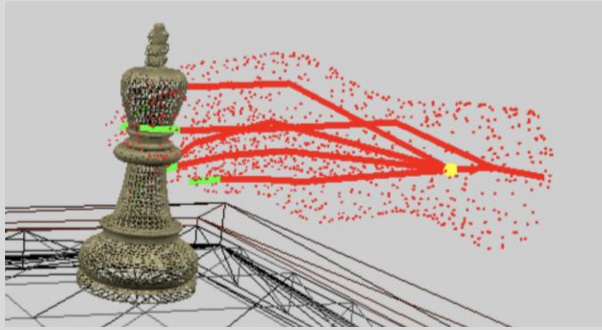
This is often not enough to interact with the world:

- Rich contact physics.
- High jitter noise from hand pose estimator.
- Kinematics and domain gap.

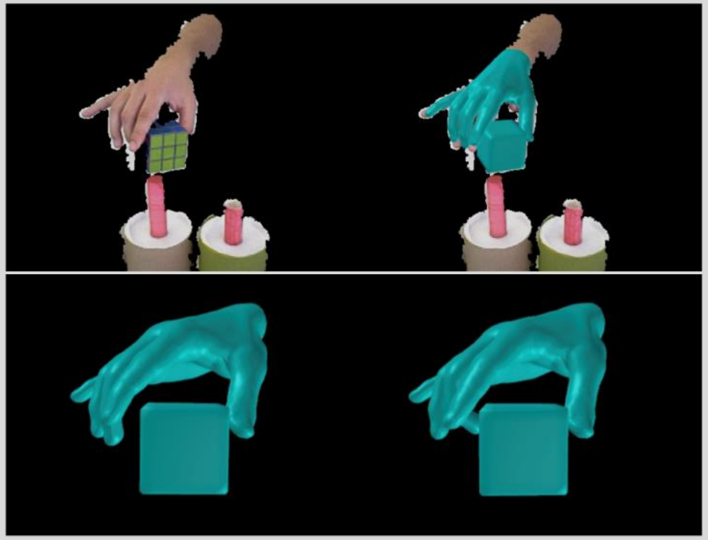
Related work: hand poses and manipulations in VR



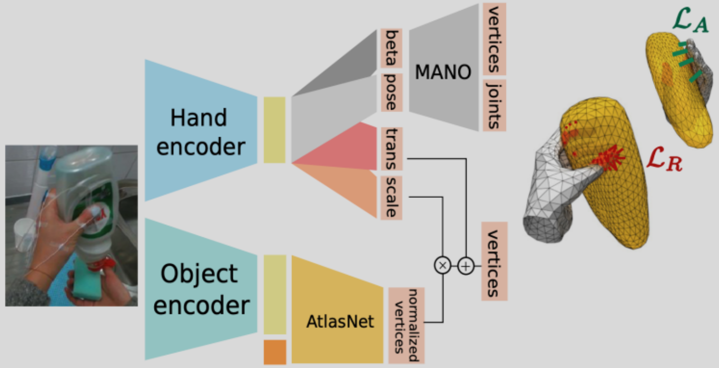
Höll et al., VR 2018.



Kim and Park, ICRA 2015.



Tzionas et al., IJCV 2016.



Hasson et al., CVPR 2019.

Related work: hand poses and manipulations in VR

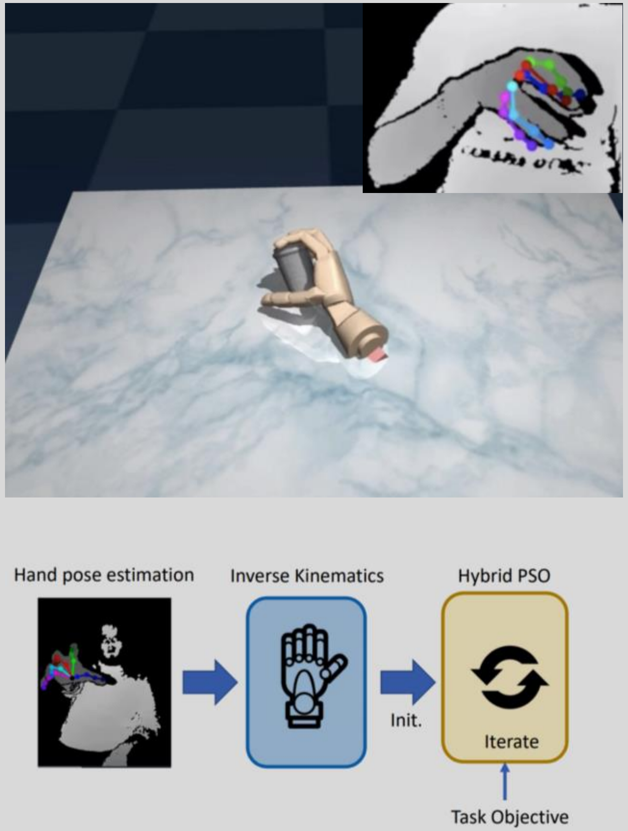
Leap Motion Interaction Engine

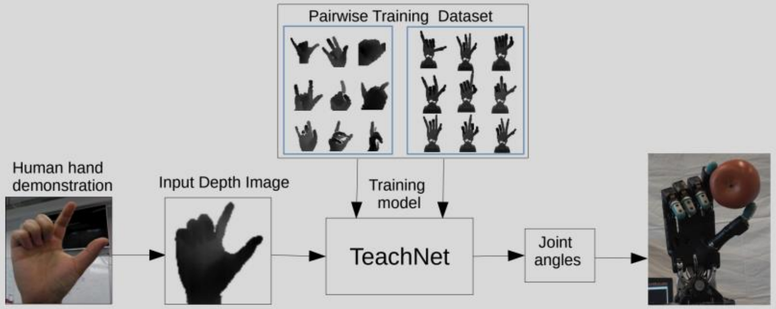
Oculus Quest (Hand Physics Lab)

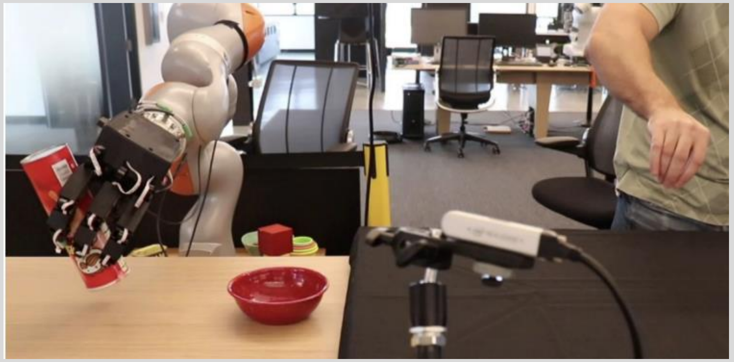
Source: https://www.youtube.com/watch?v=6MivRLfbKYs

Source: https://www.youtube.com/watch?v=J0KhC1GpLSQ

Related work: vision-based teleoperation





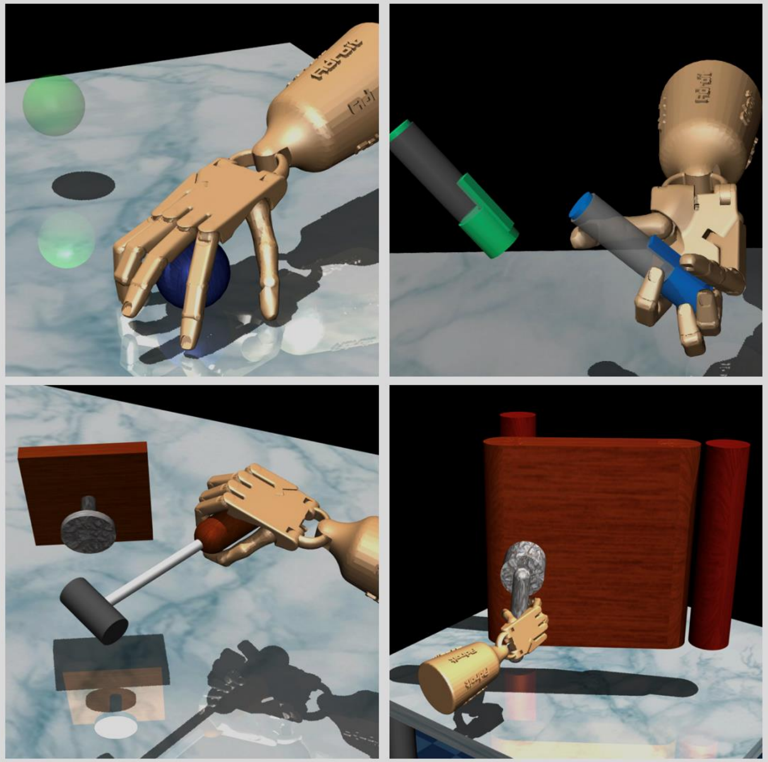


Antotsiou et al., ECCV W 2018.

Li et al., ICRA 2019.

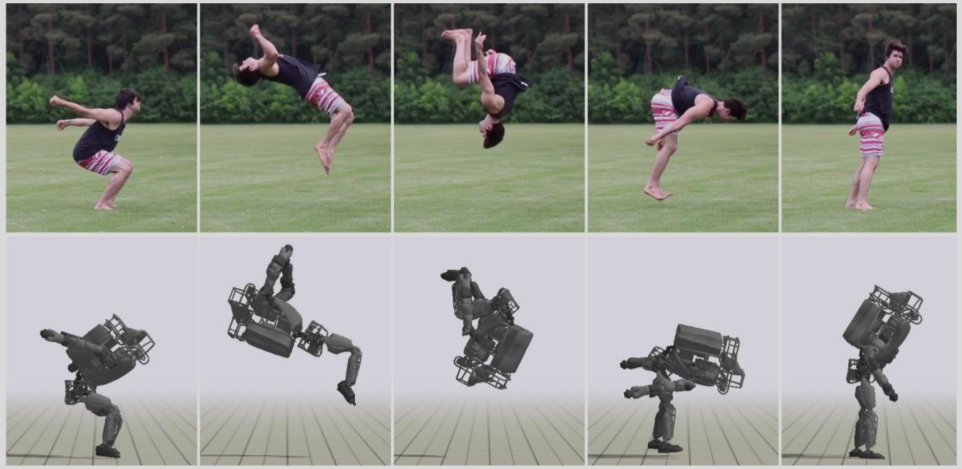
Handa et al., ICRA 2020.

Related work: dexterous manipulations and RL/IL



Rajeswaran et al., RSS 2018.

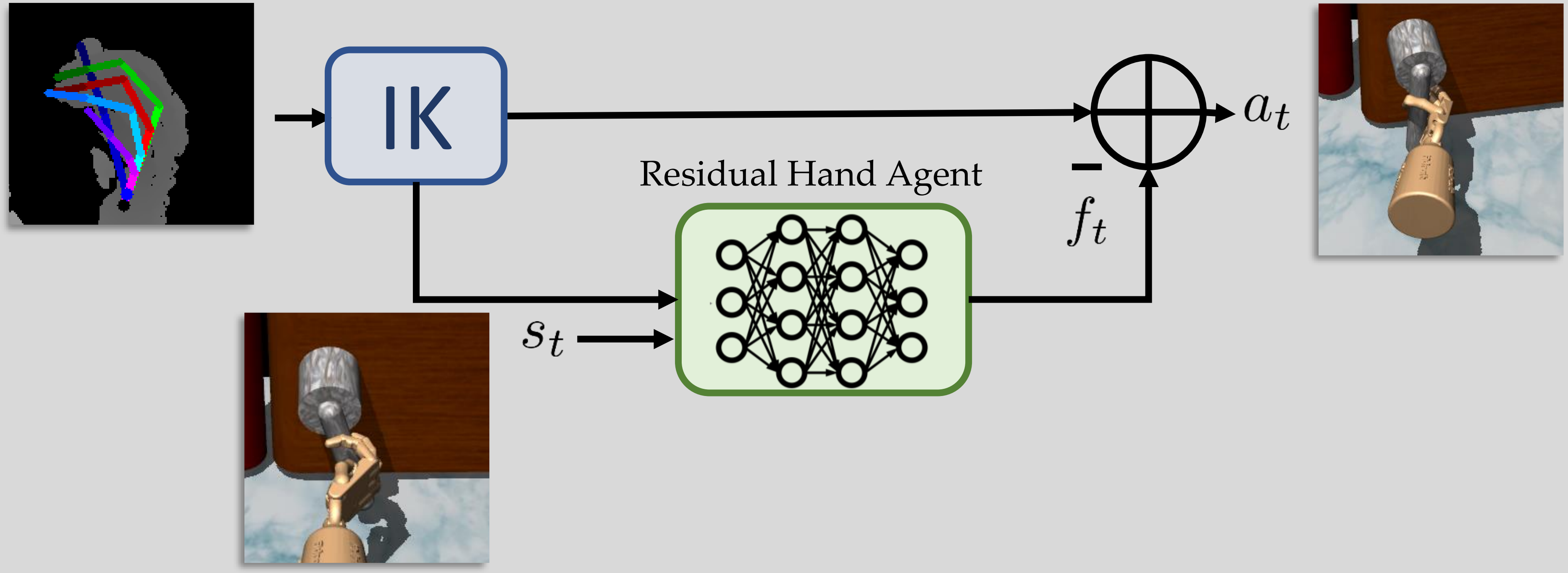
OpenAI, IJRR 2020.



Peng et al., SIGGRAPH Asia 2018.

Residual Hand Agent: Overview

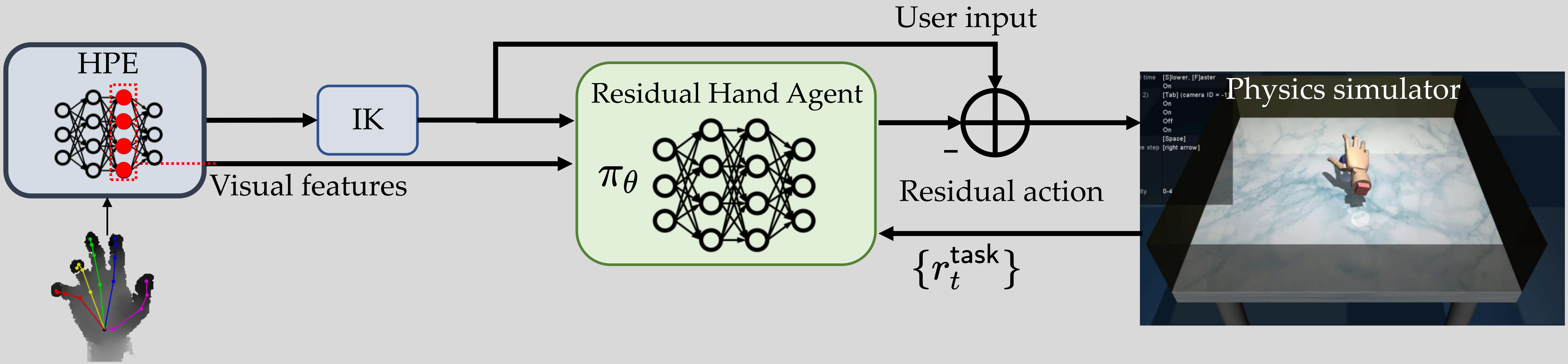
We propose a Residual Hand Agent to correct this imperfect user input:



Noisy user input

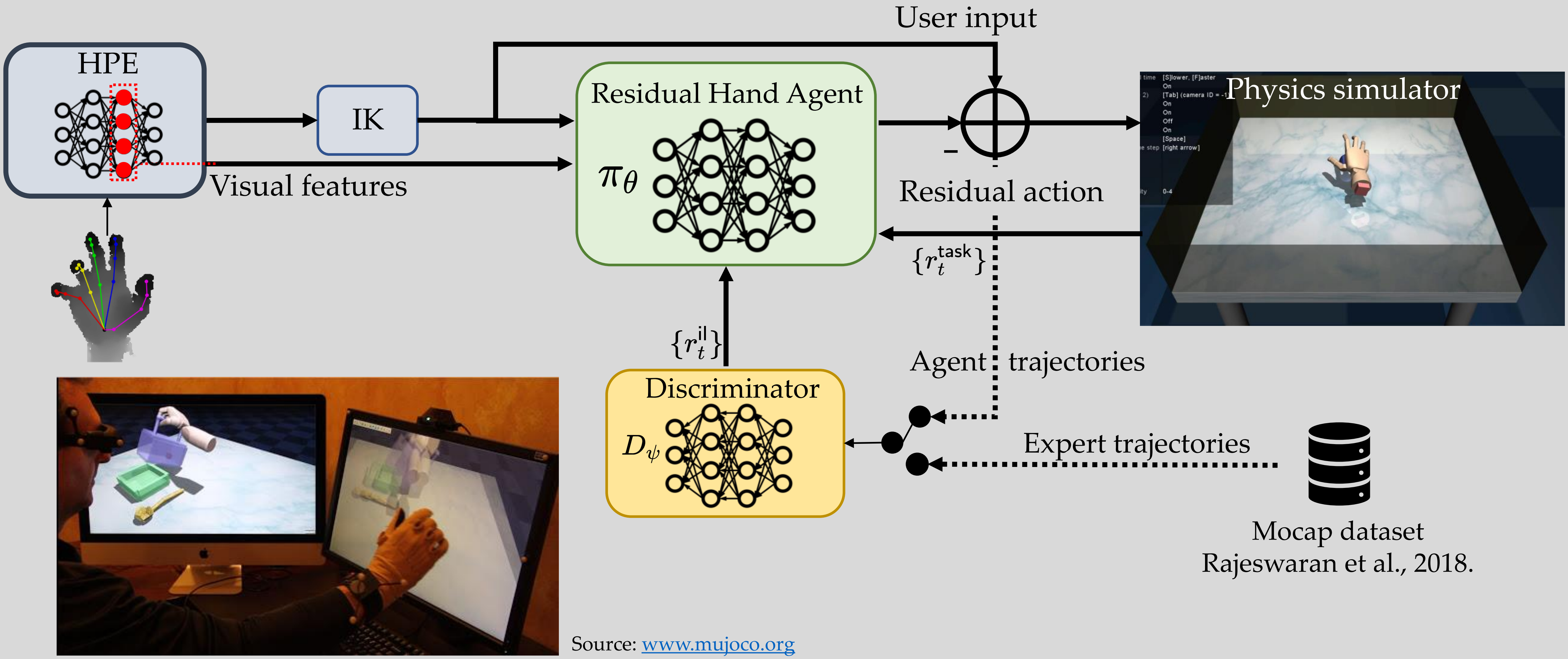
Residual Hand Agent

Residual Hand Agent: Task Reward



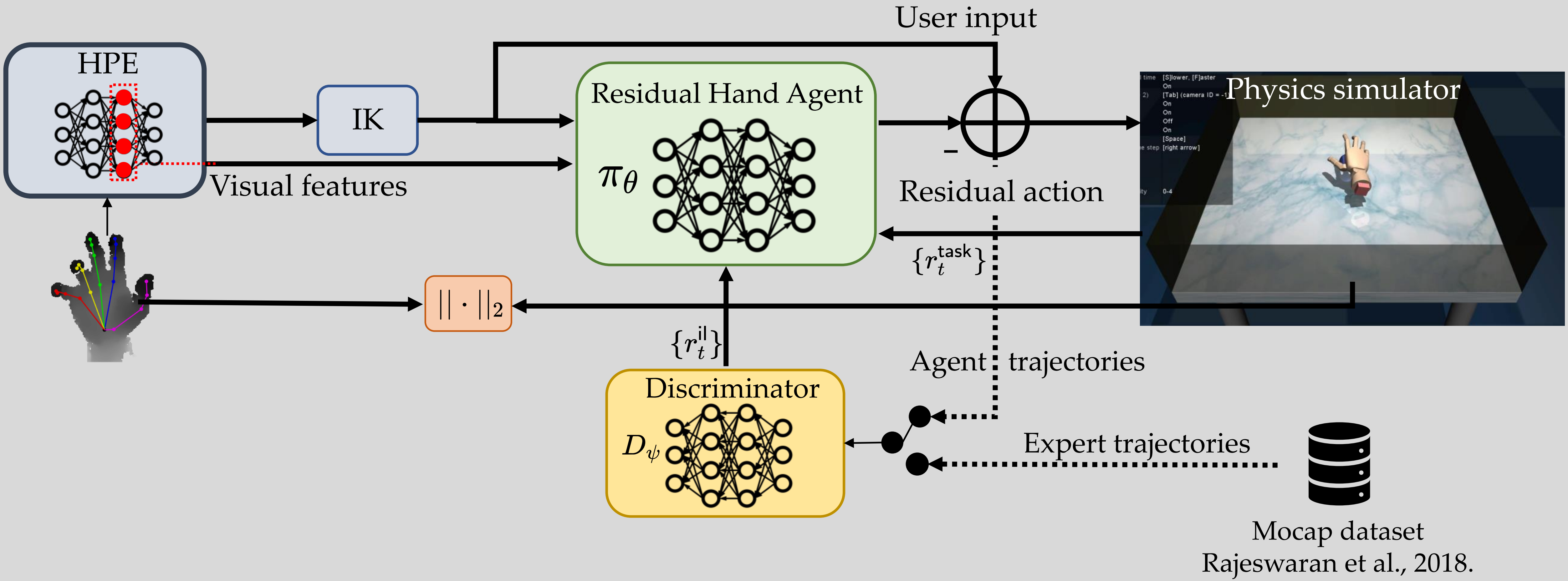
Source: Rajeswaran et al., 2018.

Residual Hand Agent: IL Reward

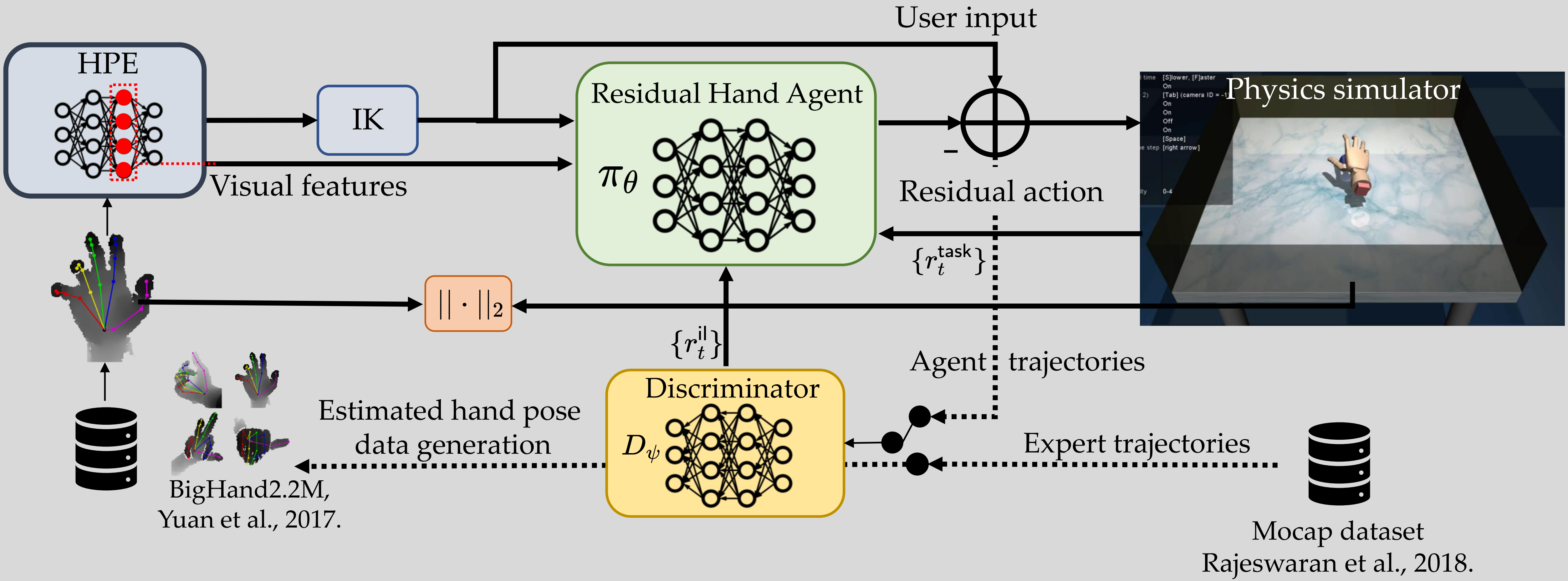


Source: www.mujoco.org

Residual Hand Agent: Hand Pose Reward



Residual Hand Agent: Data Generation



Experiments

A. Performing dexterous manipulations in the virtual space with estimated hand poses in mid-air.

B. Physics-based hand-object sequence reconstruction.

Experiment A: Overcoming random noise on demonstrations

Noisy user input

Residual Hand Agent

Residual Hand Agent
Low level input noise

Residual Hand Agent
High level input noise

Experiment A: Comparison with baselines

RL - no user reward

Residual Hand Agent

RL – no user reward

Hybrid (RL+IL)
+ user reward
– no residual

Residual Hand Agent

Experiment A: Overcoming structured hand pose estimation error

Baseline (IK)

Residual Hand Agent

(Sequence generated with our data generation scheme)

Baseline (IK)

Residual Hand Agent

(Sequence generated with our data generation scheme)

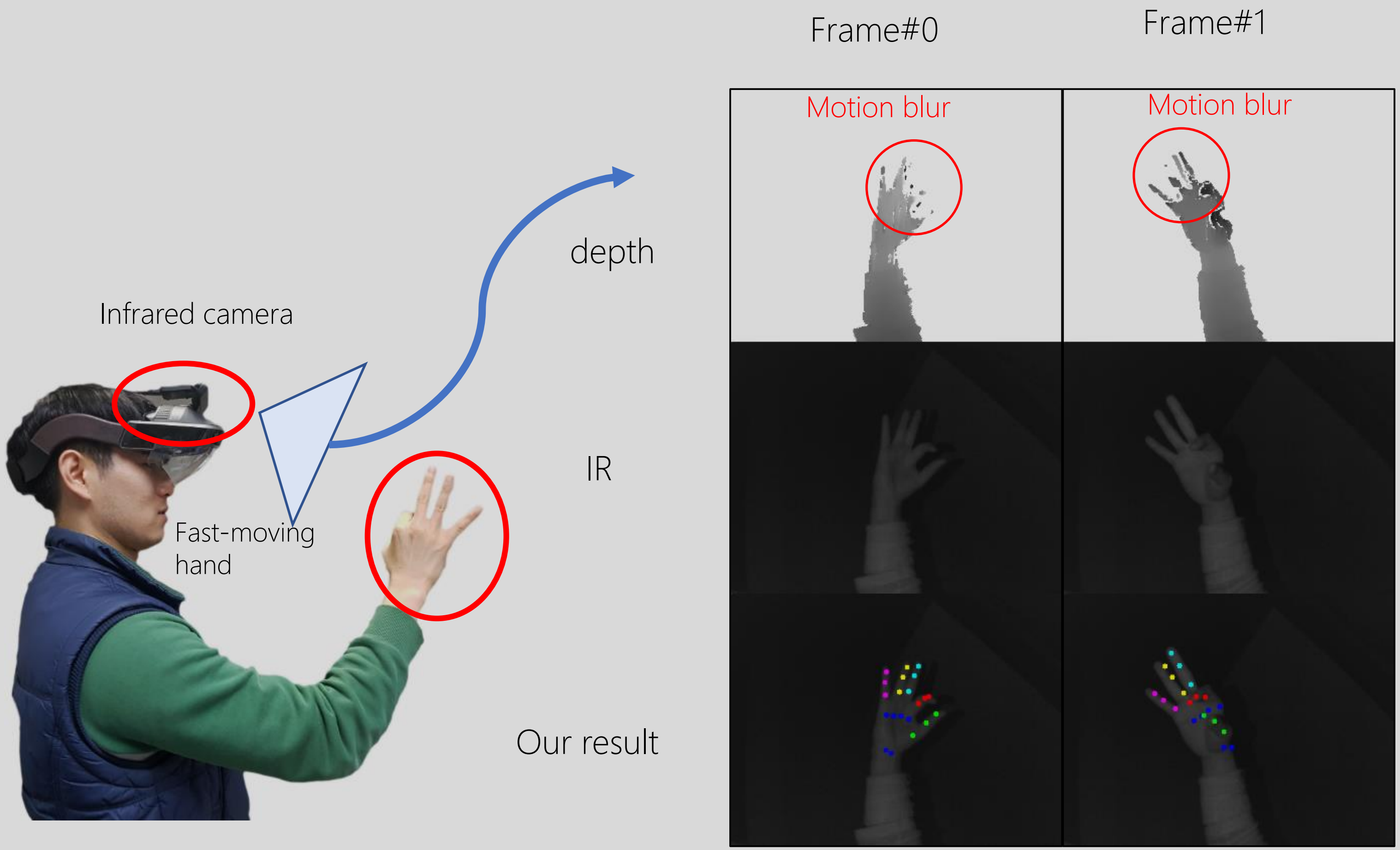
Experiment B: Physics-based hand-object sequence reconstruction.

Pour juice action: qualitative examples

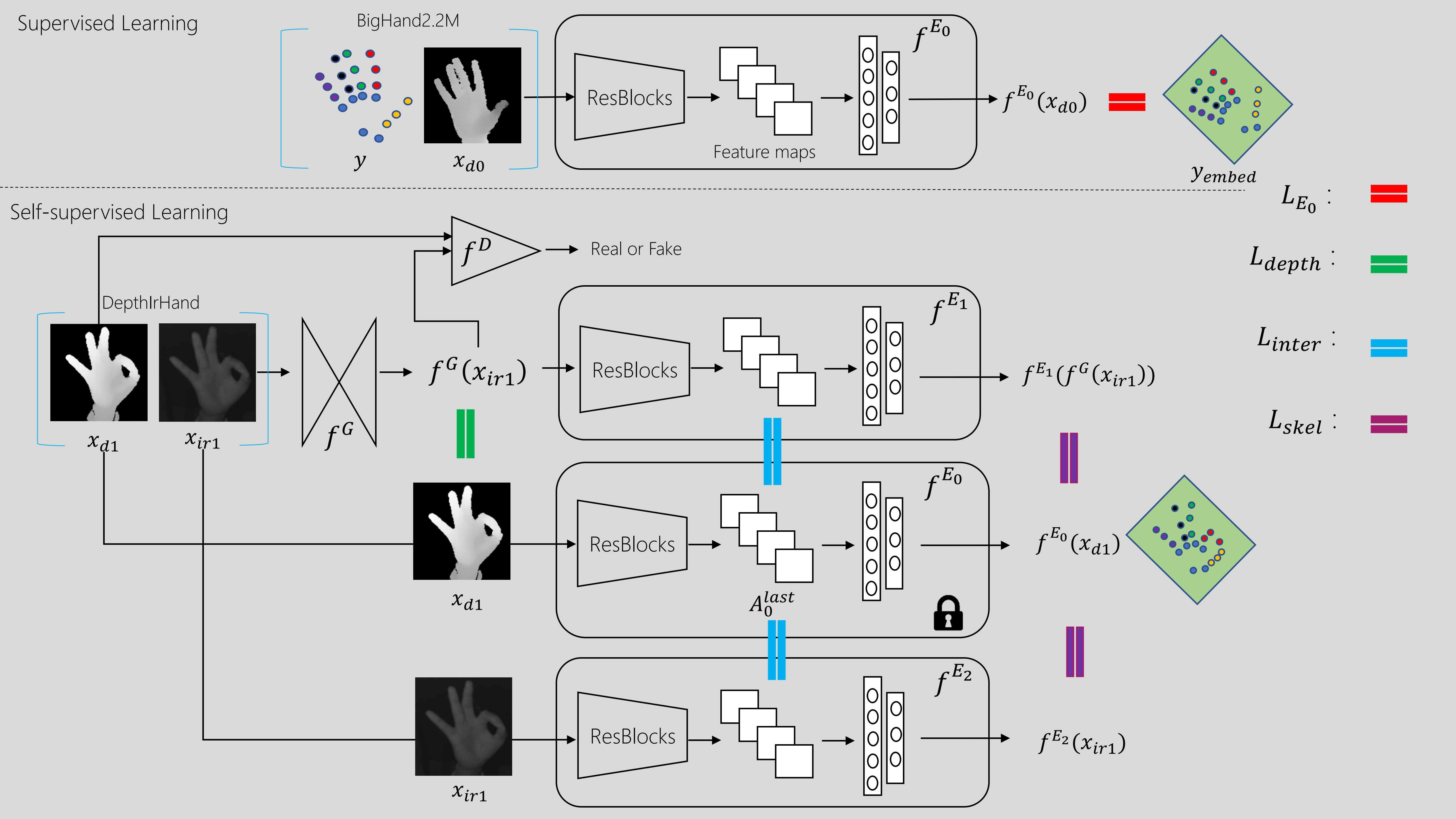
Summary and future work

- Residual framework that can perform manipulation skills by simply using a hand pose estimator and a camera.
- We showed two different applications of our approach.
- Future work: end-to-end approach with 6D object pose estimation in the loop. The use of synthetic hand data generation can help.
- Future work: study of generalization to different tasks and environments.

Domain Transfer



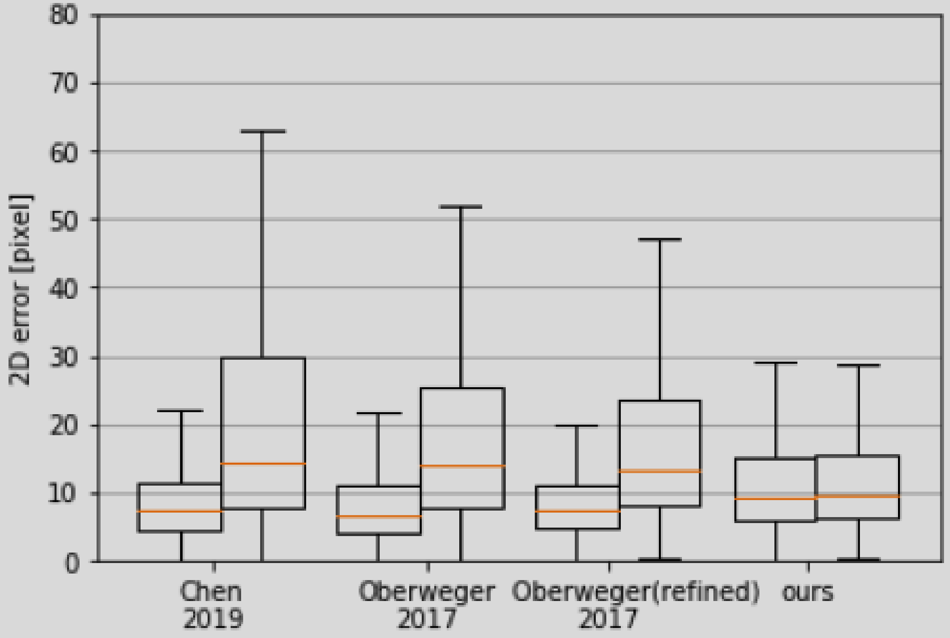
G. Park, T-K Kim, and W. Woo, “3D Hand Pose Estimation with a Single Infrared Camera via Domain Transfer Learning”, ISMAR (2020)



Supervised Learning
BigHand2.2M
y
x_{d0}
ResBlocks
Feature maps
f^{E_0}
$f^{E_0}(x_{d0})$
y_{embed}
Self-supervised Learning
f^D
Real or Fake
DepthIrHand
x_{d1}
x_{ir1}
f^G
$f^G(x_{ir1})$
ResBlocks
f^{E_1}
$f^{E_1}(f^G(x_{ir1}))$
x_{d1}
ResBlocks
A_0^{last}
f^{E_0}
$f^{E_0}(x_{d1})$
x_{ir1}
ResBlocks
f^{E_2}
$f^{E_2}(x_{ir1})$
L_{E_0} :
L_{depth} :
L_{inter} :
L_{skel} :

Comparison

<Quantitative comparison>



Left bar: Slow motion
Right bar: Fast motion

<Qualitative comparison>

[1] Oberweger et al., “DeepPrior++: Improving Fast and Accurate 3D Hand Pose Estimation” (Hands17 Workshop)
[2] Chen et al., “Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation” (Neurocomputing 2019)

Video

References

- S. Baek, K. I. Kim, T-K. Kim, Weakly-supervised Domain Adaptation for 3D Hand Pose Estimation under Hand-Object Interaction via GAN and 3D Mesh Model, Prof. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Seattle, Washington, USA, 2020 (oral, best paper nominee).
- L. Wang, T-K. Kim, K.-J. Yoon, EventSR: From Asynchronous Events to Image Reconstruction, Restoration, and Super-Resolution via End-to-End Adversarial Learning, Prof. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Seattle, Washington, USA, 2020.
- J. Sock, G. Garcia-Hernando, T-K. Kim, Active 6D Multi-Object Pose Estimation inA. Armagan, G. Garcia-Hernando, S. Baek, S. Hampali, M. Rad, Z. Zhang, S. Xie, N. Chen, B. Zhang, F. Xiong, Y. Xiao, Z. Cao, J. Yuan, P. Ren, W. Huang, H. Sun, M. Hrúz, J. Kanis, Z. Krňoul, Q. Wan, S. Li, D. Lee, L. Yang, A. Yao, Y-H. Liu, A. Spurr, P. Molchanov, U. Iqbal, P. Weinzaepfel, R. Brégier, G. Rogez, V. Lepetit, T-K. Kim, Measuring Generalisation to Unseen Viewpoints, Articulations, Shapes and Objects for 3D Hand Pose Estimation under Hand-Object Interaction, Proc. of European Conference on Computer Vision (ECCV), Edinburgh, UK, 2020.
- W. Im, T-K. Kim, S. Yoon, Unsupervised Learning of Optical Flow with Deep Feature Similarity, Proc. of European Conference on Computer Vision (ECCV), Edinburgh, UK, 2020.
- B. Bhattarai, T-K. Kim, Inducing Optimal Attributes Representations for Conditional GANs, Proc. of European Conference on Computer Vision (ECCV), Edinburgh, UK, 2020.
- X. Shi, Z. Chen, T-K. Kim, Distance-Normalized Unified Representation for Monocular 3D Object Detection, Proc. of European Conference on Computer Vision (ECCV), Edinburgh, UK, 2020.
- J. Sock, G. Garcia-Hernando, T-K. Kim, Active 6D Multi-Object Pose Estimation in Cluttered Scenarios with Deep Reinforcement Learning, Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA, 2020.
- G. Garcia-Hernando, E. Johns, T-K. Kim, Physics-Based Dexterous Manipulations with Estimated Hand Poses and Residual Reinforcement Learning, Prof. of IROS, 2020.
- K. Park, S. Kim, Y. Yoon, T-K. Kim, G. Lee, DeepFisheye: Near-Surface Multi-Finger Tracking Technology Using Fisheye Camera, Proc. of ACM Symposium on User Interface Software and Technology (UIST), Minneapolis, MN, USA, 2020.
- G. Park, T-K. Kim, W. Woo, 3D Hand Pose Estimation with a Single Infrared Camera via Domain Transfer Learning, IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR), 2020.
- B. Bhattarai, S. Baek, R. Bodur, T-K. Kim, Sampling Strategies for GAN Synthetic Data, Prof. ICASSP, Barcelona, Spain, 2020.
- J. Wang, Y. Zhang, T-K. Kim, Y. Gu, Shapley Q-value: A Local Reward Approach to Solve Global Reward Games, Proc. of AAAI Conf. on Artificial Intelligence (AAAI), New York, USA, 2020 (oral).

References

- S. Baek, K.I. Kim, T-K. Kim, Pushing the Envelope for RGB-based Dense 3D Hand Pose Estimation via Neural Rendering, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, California, USA, 2019.
- C. Loy, X. Liu, T-K. Kim, F. Torre, R. Chellappa, Editorial: Special Issue on Deep Learning for Face Analysis, *Int. Journal of Computer Vision (IJCV)*, 127(6-7):533-536, 2019.
- T-K. Kim, S. Zafeiriou, B. Glocker, S. Leutenegger, Editorial: Special Issue on Machine Vision, *Int. Journal of Computer Vision (IJCV)*, 2019. https://doi.org/10.1007/s11263-019-01201-4
- W. Luo, B. Stenger, X. Zhao, T-K. Kim, Trajectories as Topics: Multi-Object Tracking by Topic Discovery, *IEEE Trans on Image Processing (TIP)*, 28 (1), 240-252, 2019.
- Q. Ye, T-K. Kim, Occlusion-aware Hand Pose Estimation Using Hierarchical Mixture Density Network, Proc. of European Conf. on Computer Vision (ECCV), Munich, Germany, 2018 (oral).
- B. Gecer, B. Bhattarai, J. Kittler, T-K. Kim, Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model, Proc. of European Conf. on Computer Vision (ECCV), Munich, Germany, 2018.
- T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T-K. Kim, J. Matas, C. Rother, BOP: Benchmark for 6D Object Pose Estimation, Proc. of European Conf. on Computer Vision (ECCV), Munich, Germany, 2018.
- J. Sock, K.I. Kim, C. Sahin, T-K. Kim, Multi-Task Deep Networks for Depth-Based 6D Object Pose and Joint Registration in Crowd Scenarios, Proc. of British Machine Vision Conference (BMVC), Newcastle upon Tyne, UK, 2018.
- D. Tang, Q. Ye, S. Yuan, J. Taylor, P. Kohli, C. Keskin, T-K. Kim, J. Shotton, Opening the Black Box: Hierarchical Sampling Optimization for Estimating Human Hand Pose, *IEEE Trans on PAMI (TPAMI)*, accepted to appear, 2018.
- S. Baek, K.I. Kim, T-K. Kim, Augmented skeleton space transfer for depth-based hand pose estimation, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 2018 (oral, accept rate=2.1%).
- S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, J. Yuan, X. Chen, G. Wang, F. Yang, K. Akiyama, Y. Wu, Q. Wan, M. Madadi, S. Escalera, S. Li, D. Lee, I. Oikonomidis, A. Argyros, T-K. Kim, 3D Hand Pose Estimation: From Current Achievements to Future Goals, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 2018 (spotlight).
- G. Garcia-Hernando, S. Yuan, S. Baek, T-K. Kim, First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 2018.
- S. Kasaei, J. Sock, L. Lopes, A. Tome, T-K. Kim, Perceiving, Learning, and Recognizing 3D Objects: An Approach to Cognitive Service Robots, Proc. of the Association for the Advancement of Artificial Intelligence (AAAI), New Orleans, USA, 2018 (oral).
- Y. Pei, Y. Yi, G. Ma, T-K. Kim, T. Xu, and H. Zha, Finding Spatially-Consistent Supervoxel Correspondence of Cone-Beam Computed Tomography Images, *IEEE Trans. on Medical Imaging*, accepted to appear, 2018.
- V. Balntas, A. Doumanoglou, C. Sahin, J. Sock, R. Kouskouridas, T-K. Kim, Pose guided RGBD embedding learning for object pose estimation, Proc. of IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, 2017.
- S. Baek, Z. Shi, M. Kawade, T-K. Kim, Kinematic-layout-aware random forests for depth-based action recognition, Proc. of British Machine Vision Conference (BMVC), London, UK, 2017 (oral, accept rate = 5.6%).
- G. Garcia-Hernando, T-K. Kim, Transition Forests: Learning Discriminative Temporal Transitions for Action Recognition, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017.
- Z. Shi, T-K. Kim, Learning and Refining of Privileged Information-based RNNs for Action Recognition from Depth Sequences, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017.
- S. Yuan, Q. Ye, B. Stenger, S. Jain, T-K. Kim, Big Hand 2.2M Benchmark: Hand Pose Data Set and State of the Art Analysis, Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017.


https://labicvl.github.io/
https://sites.google.com/view/tkkim/