De-occlusion in pose estimation

2022년도 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
윤준하

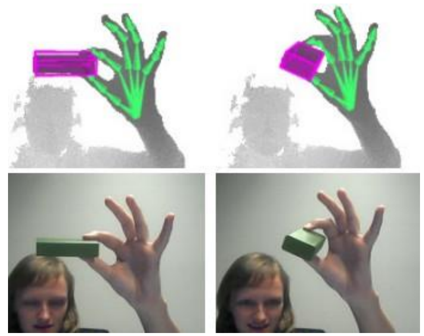
Outline

[1] “Self-supervised scene de-occlusion.” (CVPR 2020)
[2] “Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)
[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

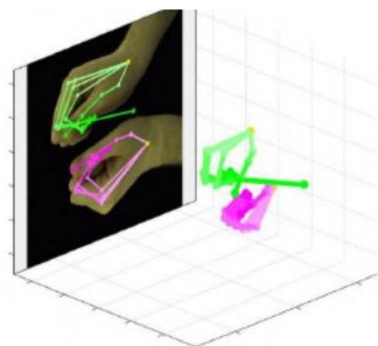
- Backgrounds
  - About topic
  - What is De-occlusion
  - Inpainting
- Human-pose de-occlusion
  - Method
  - Dataset
  - Experiment
- Hand-pose de-occlusion
  - Method
  - Dataset
  - Experiment

Backgrounds

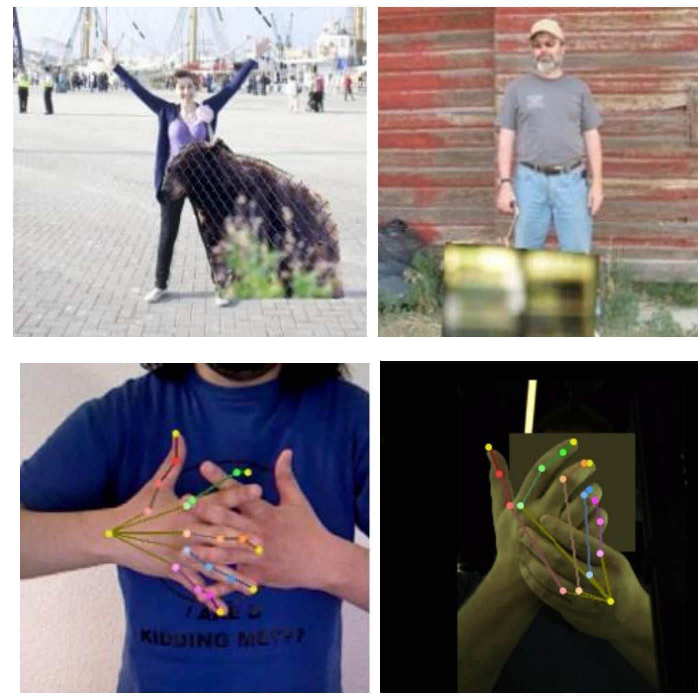
- Single pose estimation



- Multi pose estimation



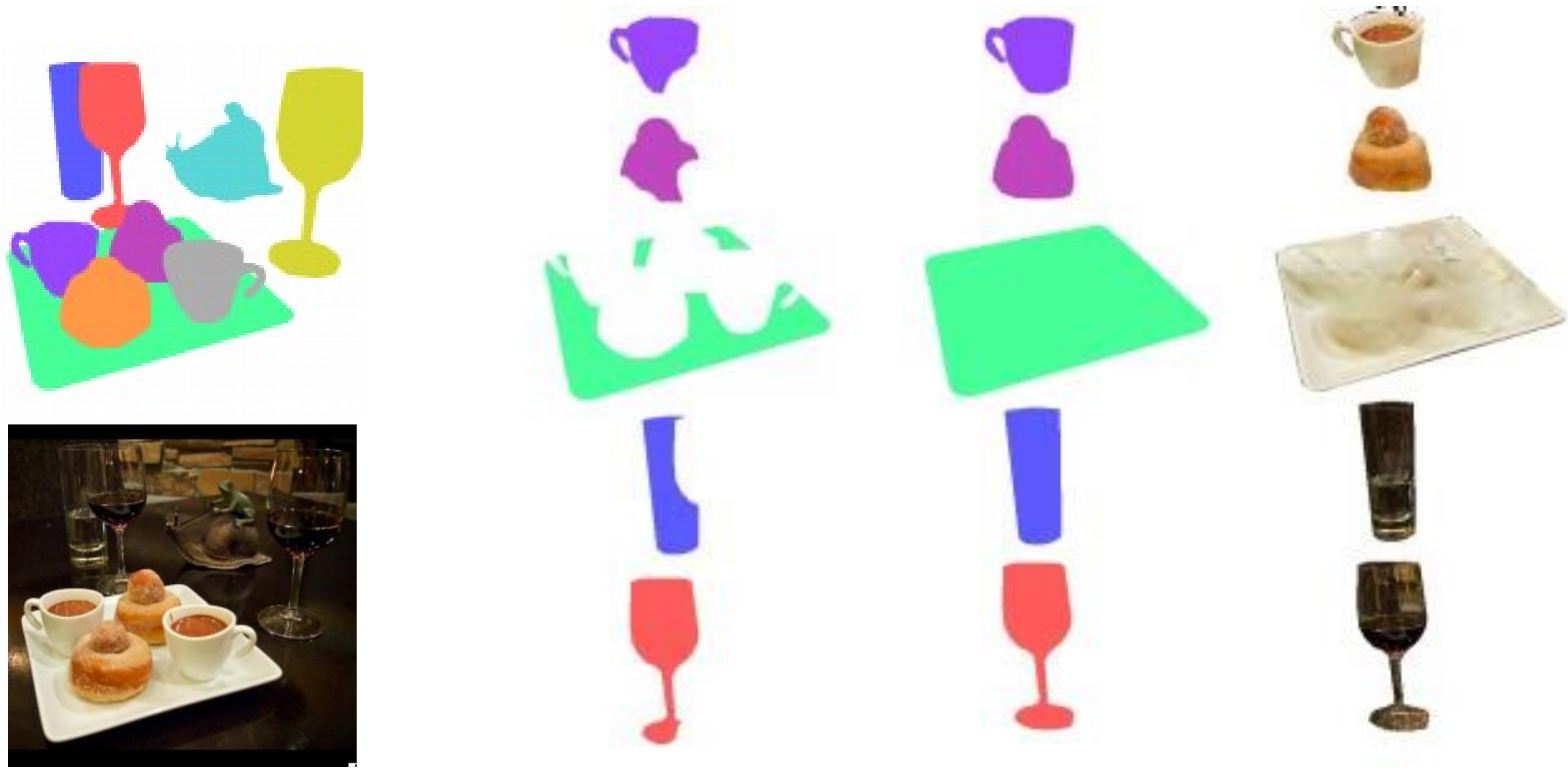
How we could approach Occlude cases ?



Backgrounds

[1] "Self-supervised scene de-occlusion." (CVPR 2020)

What is De-occlusion?



Input image | Modal masks (occluded objects) | Amodal completion | Amodal-guided content completion

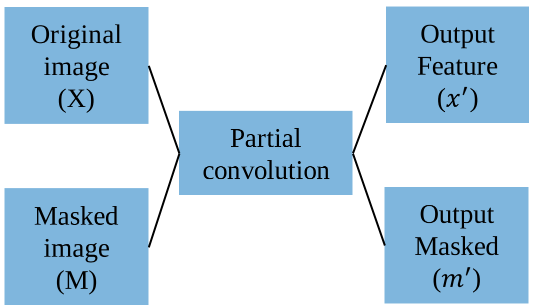
Backgrounds

Inpainting

주변 pixel값을 참고하여 hole을 채움

Backgrounds

Partial Convolution



$$x' = \begin{cases} \mathbf{W}^T(\mathbf{X} \odot \mathbf{M}) \frac{\text{sum}(\mathbf{1})}{\text{sum}(\mathbf{M})} + b, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$m' = \begin{cases} 1, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

X

50	20	30	50	50
40	30	20	10	10
30	40	50	10	30
10	10	40	30	20
20	10	30	40	50

M

1	1	1	0	0
1	0	0	0	0
1	0	0	0	0
1	1	1	1	1
1	1	1	1	1

X⊙M

50	20	30	0	0
40	0	0	0	0
30	0	0	0	0
10	10	40	30	20
20	10	30	40	50

Zero padding

0	0	0	0	0	0	0
0	50	20	30	0	0	0
0	40	0	0	0	0	0
0	30	0	0	0	0	0
0	10	10	40	30	20	0
0	20	10	30	40	50	0
0	0	0	0	0	0	0

$W^t(X \odot M) x \sum(1) / \sum(M)$

110/3	140/4	50/2	30/1	0
140/4	170/5	50/2	30/1	0
90/4	130/5	80/3	90/3	50/2
80/5	150/7	160/6	210/6	140/4
50/4	120/6	160/6	210/6	140/2

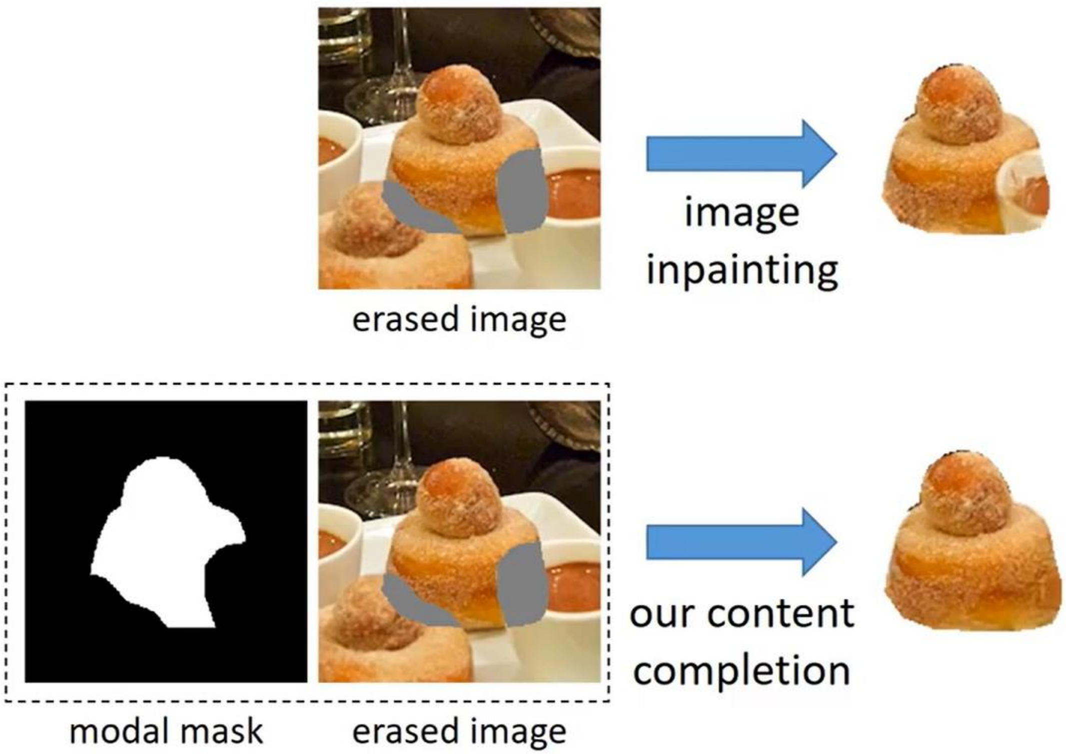
m'

1	1	1	1	0
1	1	1	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

Backgrounds

[1] “Self-supervised scene de-occlusion.” (CVPR 2020)

Compared to Image Inpainting

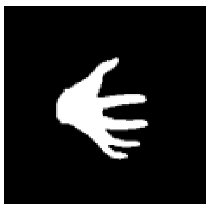


Backgrounds

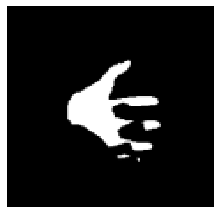
Words

- Hand Pose Estimation
  - Ground Truth

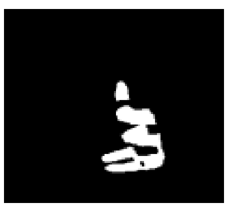
Occlusion image



Right hand Amodal Mask



Right hand Visible Mask

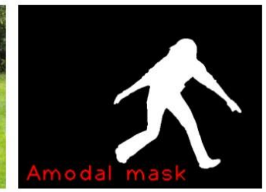


Left hand Visible Mask

De-occlusion & Removal

- Human Pose Estimation
  - Ground Truth

Occlusion image



Amodal Mask

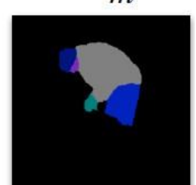
Modal Mask

Invisible Mask

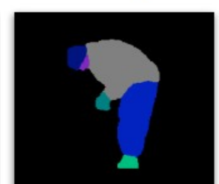
Recovered image

  - Human Parsing

$\hat{M}^p_m$

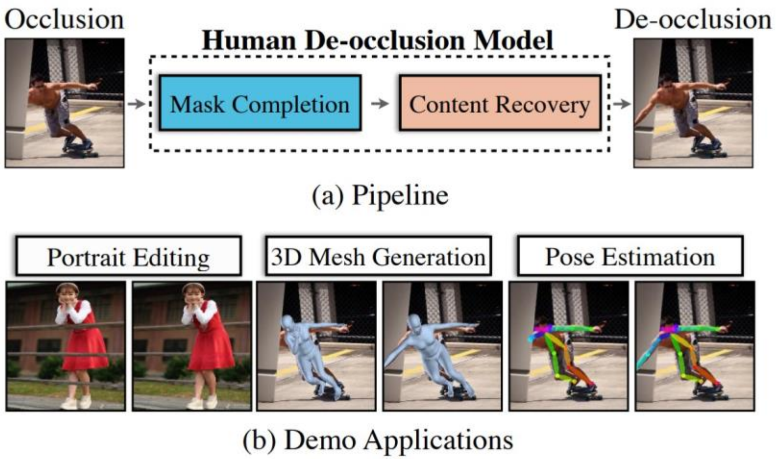


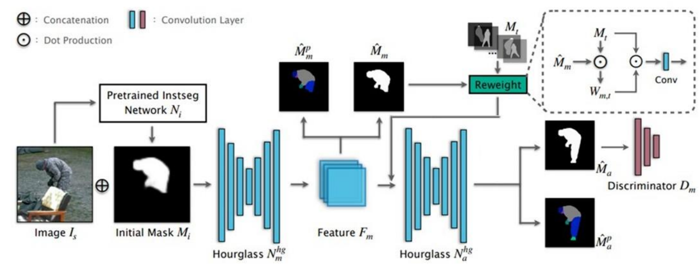
$\hat{M}^p_a$



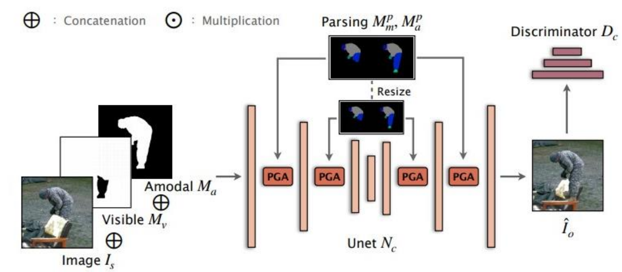
Introduction

[2] “Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)





Mask completion



Content recovery

- Mask Completion
  - First hourglass module to refine the inaccurate input modal mask
  - Second hourglass module is applied to estimate the integrated amodal mask

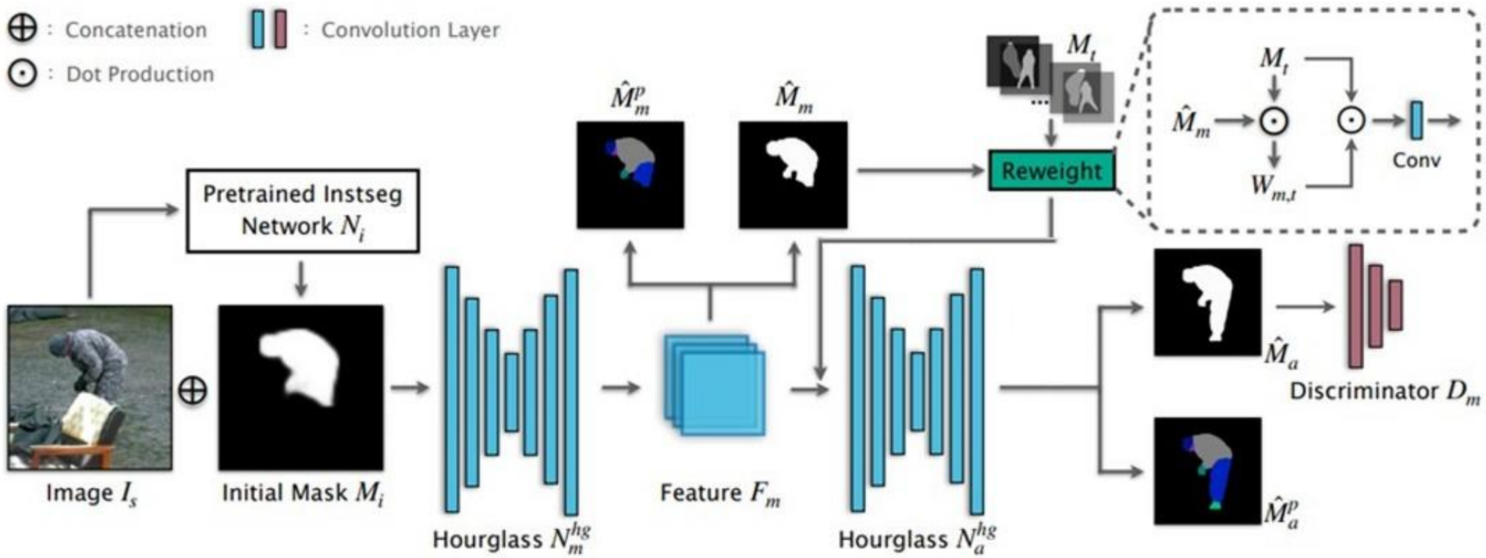
- Content recovery
  - Recovers the appearance content inside the visible portions

서강대학교 SOGANG UNIVERSITY

VDS LAB

Method - Mask Completion

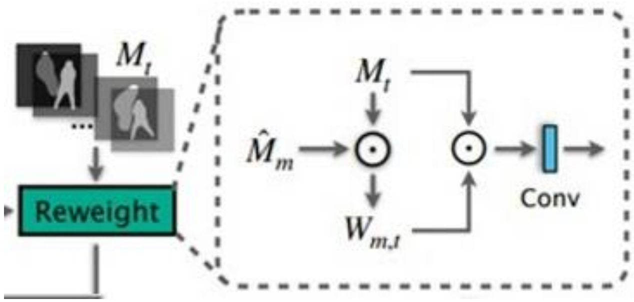
[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)

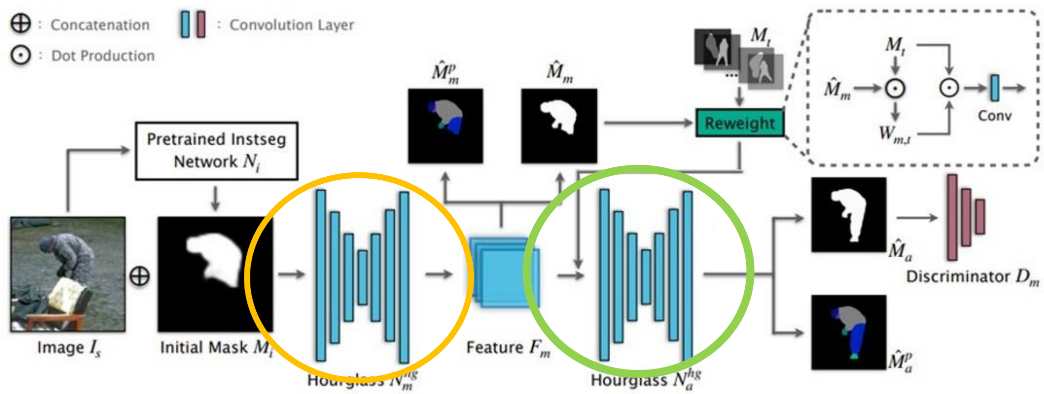


- Mask Completion
  - A pretrained instance segmentation network N_I is applied to obtain the initial modal mask M_i from input image I_s
  - One hourglass module N_m^{hg} outputs a refined modal mask $\widehat{M}_m$ and the corresponding parsing result $\widehat{M}_m^p$
  - Another hourglass module N_a^{hg} is stacked with the template masks finally outputs the amodal mask $\widehat{M}_a$ and the parsing result $\widehat{M}_a^p$
  - A discriminator D_m is applied to improve the quality of the generated amodal mask

Method - Mask Completion

[2] “Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)





- Reweight
  - M_t: k-means를 이용해 template masks를 비슷한 외형끼리 분류
  - $D_{m,t}$: l_2 distances of M_t and $\widehat{M}_m$
  - $W_{m,t} = 1/D_{m,t}$
  - $W_{m,t}$ is multiplied back with the template masks to highlight suitable candidates

- The modal recognition process (Yellow)

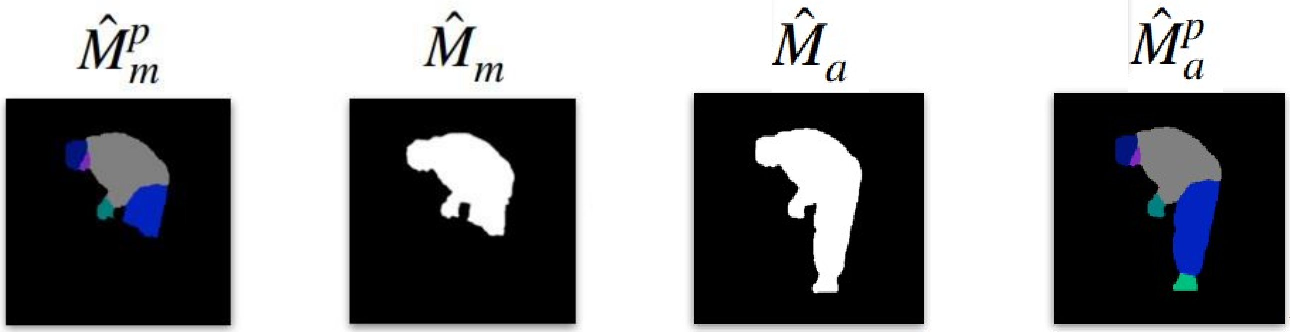
$$\hat{M}_m,\ \hat{M}_m^p = N_m^{hg}\left(I_s,\ N_i\left(I_s\right)\right)$$

- The amodal completion process (Green)

$$\hat{M}_a,\ \hat{M}_a^p = N_a^{hg}\left(F_m \oplus \hat{M}_m,\ \mathrm{Conv.}(M_t \odot W_{m,t})\right)$$

Method - Mask Completion

[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)



- Modal, amodal and the human parsing loss

$$\mathcal{L}_{seg} = \mathcal{L}_{CE}(\hat{M}_m, M_m) + \mathcal{L}_{CE}(\hat{M}_a, M_a) + \mathcal{L}_{CE}(\hat{M}_m^p, M_m^p) + \mathcal{L}_{CE}(\hat{M}_a^p, M_a^p)$$

- Discriminator loss

$$\mathcal{L}_{adv} = \underbrace{\mathbb{E}_{\hat{M}_a}[\log(1 - D_m(\hat{M}_a))]}_{\text{Minimal}} + \underbrace{\mathbb{E}_{M_a}[\log D_m(M_a)]}_{\text{Maximal}}$$

- Reconstruct loss

$$\mathcal{L}_{gen} = \mathcal{L}_{\ell 1}(\hat{M}_a, M_a) + \mathcal{L}_{prec}(\hat{M}_a, M_a)$$

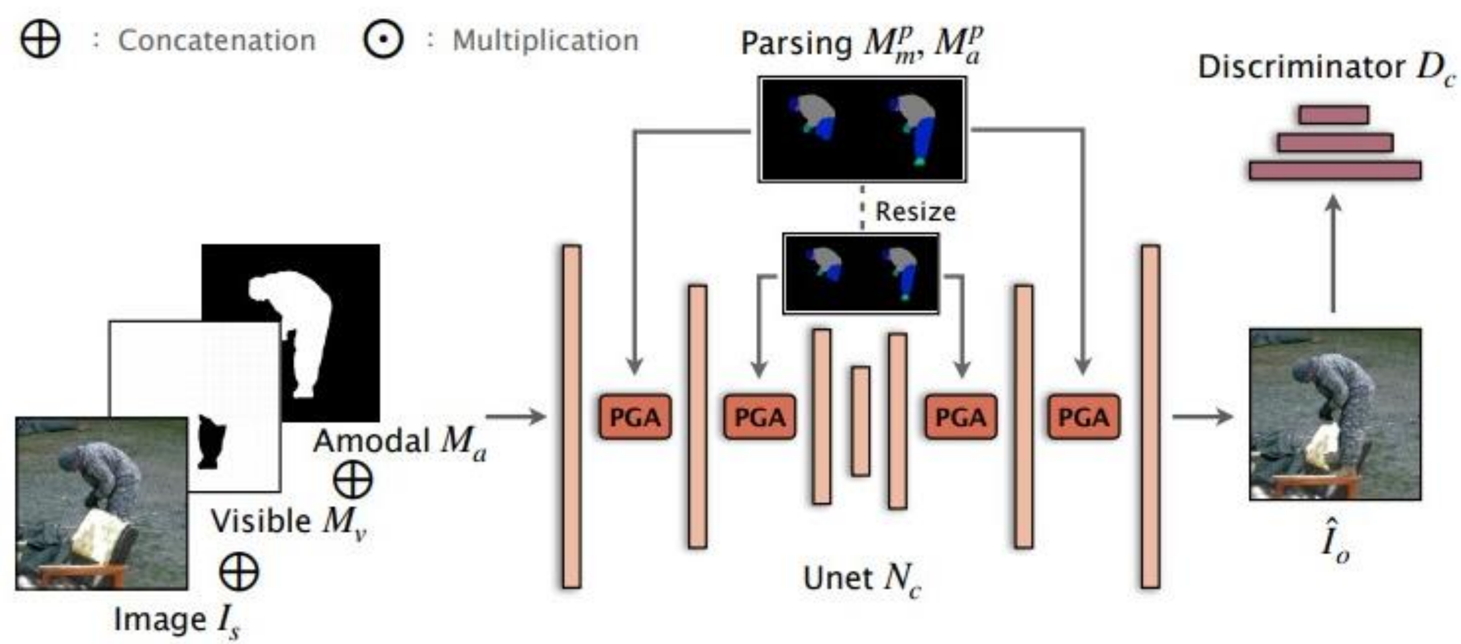
perceptual loss

- Final Loss

$$\therefore\ \mathcal{L}_m = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{gen}$$

Method - Content Recovery

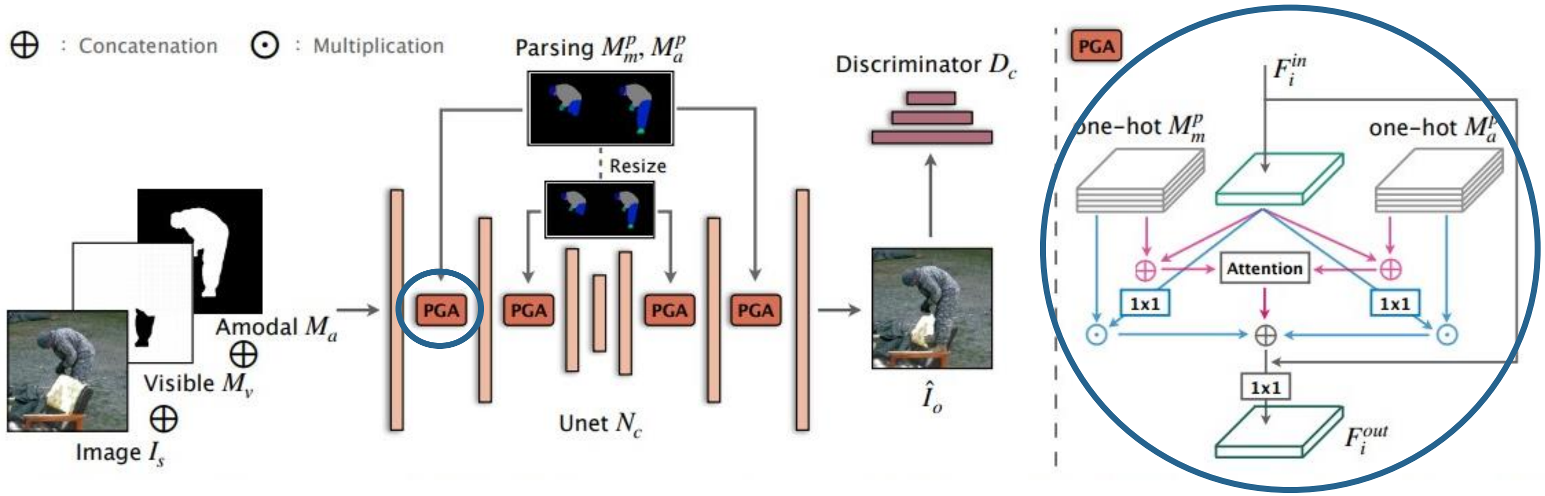
[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)



- Content recovery
  - The image I_s concatenated with the visible mask M_v and the amodal mask M_a is passed into the network N_c
  - Unet with partial convolution as the basic architecture
  - Parsing Guided Attention (PGA) module is used
  - A discriminator D_C is applied to identify the quality of the output image I_o

Method - Content Recovery

[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)

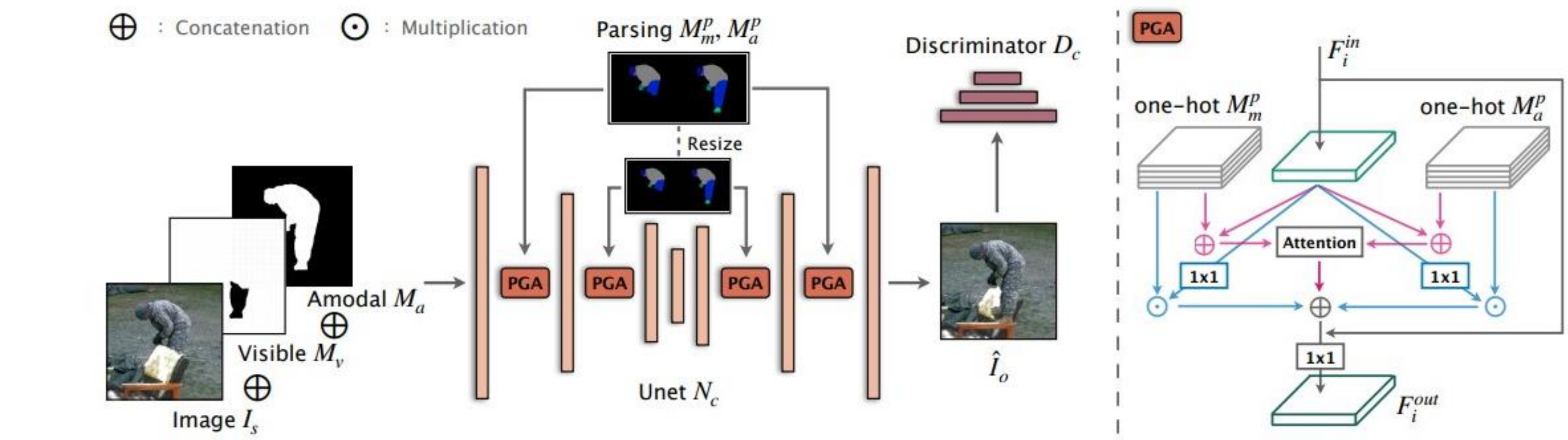


- The first stream (cyan)
  - Decomposes the feature into different body parts and compare them.
  - Feature F_i^{in} is reduced to the same channel number with the parsing logits (i.e. 19)
  - Multiplied with the two logits to distribute the feature in different body parts.
  - Two distributed features are concatenated and a 1 × 1 convolution layer is applied

- The second stream(magenta)
  - Establish the pixel-level relationship between the visible context and the invisible regions.

Method - Content Recovery

[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)



- Attention

$$\boxed{K_{amo} = \psi(F_i^{in} \oplus M_a^p)}$$ Amodal

$$\boxed{K_{vis} = \phi(F_i^{in} \oplus M_m^p)}$$ Modal

Visible, Modal   Invisible, Amodal

$$\tilde{R} = \underline{(M_v \odot K_{vis})^T}\ (\underline{(1 - M_v) \odot K_{amo}});$$

$$R = \mathrm{Softmax}(\tilde{R},\ \mathrm{dim} = 0) \in \mathbb{R}^{HW \times HW}$$

- Network process

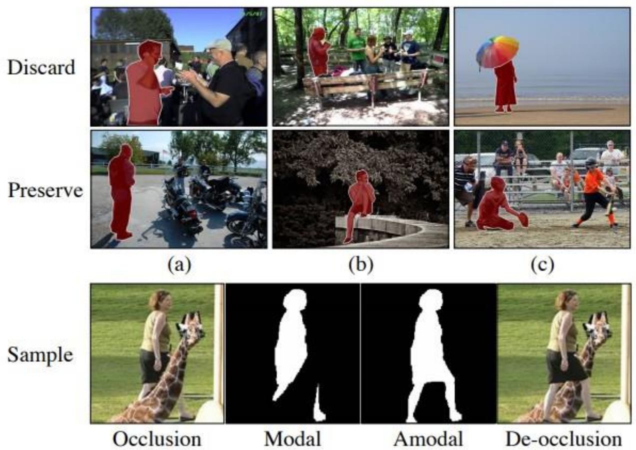
$$\hat{I}_o \ = \ N_c\,(I_s \oplus M_v \oplus M_a,\ M_m^p,\ M_a^p)$$

- Final Loss

$$\mathcal{L}_c = \beta_1 \left(\mathbb{E}_{\hat{I}_o}[\log(1 - D_c(\hat{I}_o))] + \mathbb{E}_{I_o}[\log D_c(I_o)]\right) + \beta_2\, \mathcal{L}_{\ell 1}(\hat{I}_o,\ I_o) + \beta_3\, \mathcal{L}_{prec}(\hat{I}_o,\ I_o) + \beta_4\, \mathcal{L}_{style}(\hat{I}_o,\ I_o)$$

AHP Dataset

[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)



- AHP(The Amodal Human Perception Dataset)
- Image Acquisition
  - We collect human images from several large-scale instance segmentation and detection datasets
  - Ex) COCO, VOC(with SBD), LIP, Objects365, and OpenImages
- Filtering Scheme
  - Discard : the human is occluded by other instances (e.g. desk, car or human or parts of him/her out of view
  - Preserve : the human is not occluded and the segmentation is fine
  - Refine : the human is not occluded but the segmentation result is not satisfied
- Ground Truth
  - AHP contains occlusion image, amodal mask, modal mask, invisible mask, and Recovered image

Experiment

[2] "Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)

- Quantitative comparison
  - The comparison results of mask completion task on our AHP dataset

Method	Syn.		Real	
	$\ell_1 \downarrow$	IoU $\uparrow$	$\ell_1 \downarrow$	IoU $\uparrow$
Mask-RCNN [13]	0.2402	78.4/26.9	0.2511	75.6/23.8
Deeplab [5]	0.2087	70.7/20.9	0.2179	75.7/23.5
Pix2Pix [18]	0.2329	69.6/19.2	0.2376	68.0/16.0
SeGAN [6]	0.2545	76.7/23.6	0.2544	77.7/19.0
OVSR [50]	0.1830	80.2/28.1	0.1809	82.9/25.6
PCNets [54]	0.1959	83.1/29.1	0.2218	81.3/31.2
Ours	**0.1500**	**84.6/43.7**	**0.1635**	**86.1/40.3**

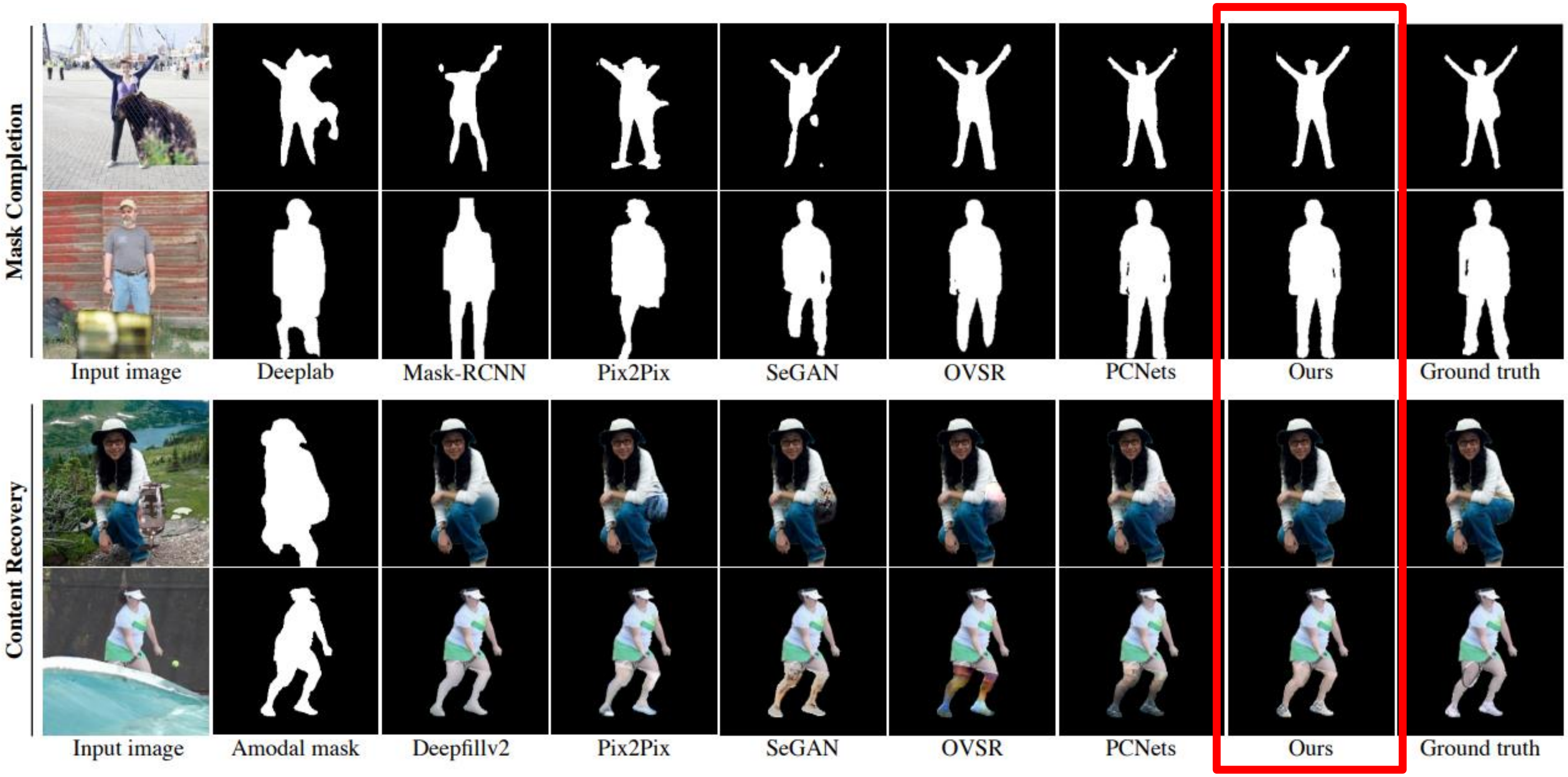
  - The comparison results of content recovery task on our AHP dataset

Method	Syn.		Real	
	$\ell_1 \downarrow$	FID $\downarrow$	$\ell_1 \downarrow$	FID $\downarrow$
Pix2Pix [18]	0.1126	19.66	0.1031	29.63
Deepfillv2 [52]	0.1127	21.61	0.1026	32.48
SeGAN [6]	0.1122	23.01	0.1027	35.21
OVSR [50]	0.0940	27.15	0.0917	36.23
PCNets [54]	0.0936	18.50	0.0911	28.30
Ours	**0.0519**	**13.85**	**0.0617**	**19.49**

Experiment

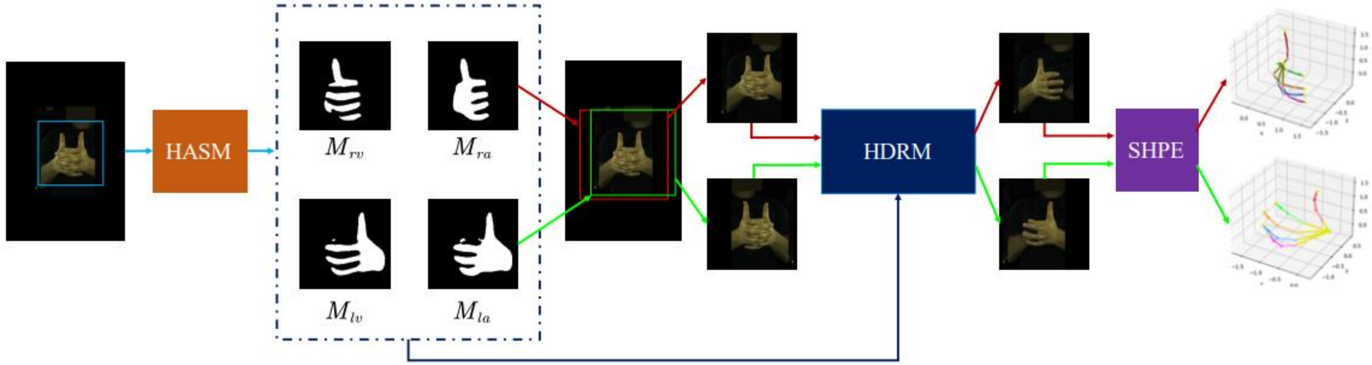
[2] “Human De-occlusion: Invisible Perception and Recovery for Humans." (CVPR 2021)

- Qualitative comparison



Introduction

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

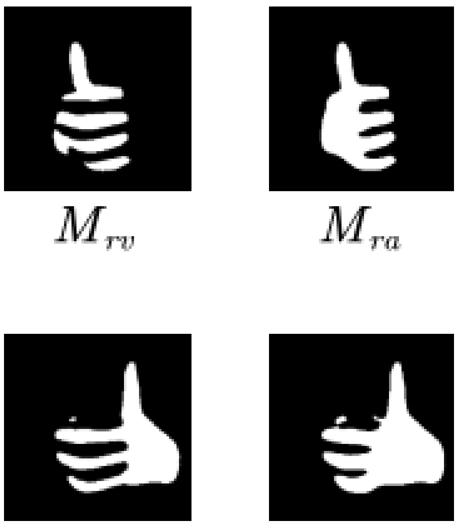


HDR(Hand De-occlusion and Removal) Framework

- HASM **(**Hand Amodal Segmentation Module**)**
  - Segment the amodal and modal masks of the left and the right hand in the image
- HDRM **(**Hand De-occlusion and Removal Module**)**
  - locate and crop the image patch centered at each hand
  - recovers the appearance content of the occluded part of one hand and removes the other distracting hand simultaneously
- SHPE (Single Hand Pose Estimator**)**
  - Get the final 3D hand poses

Method - HASM

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

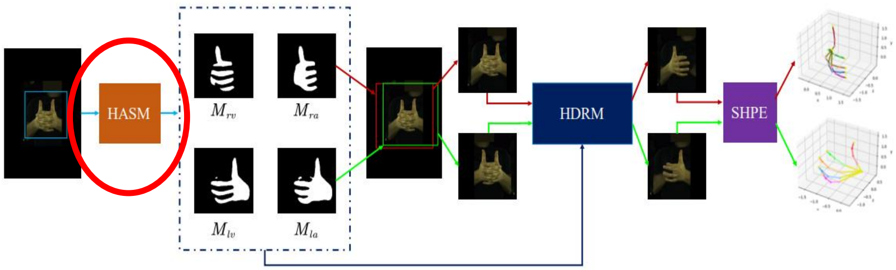


M_{rv} M_{ra}

M_{lv} M_{la}

Backbone : SegFormer

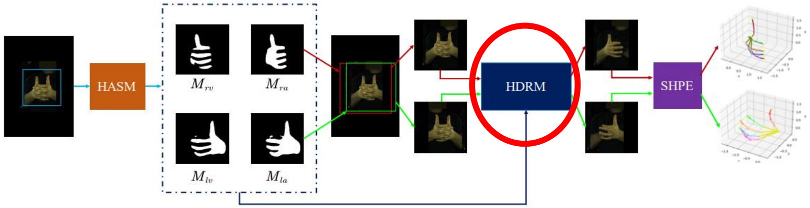
- Obtain the amodal and visible masks of both hands using the Hand Amodal Segmentation Module (HASM)

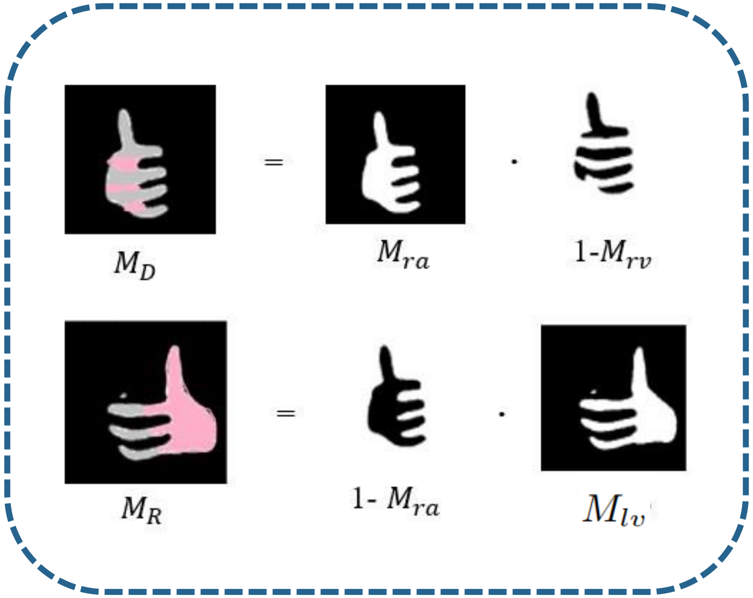


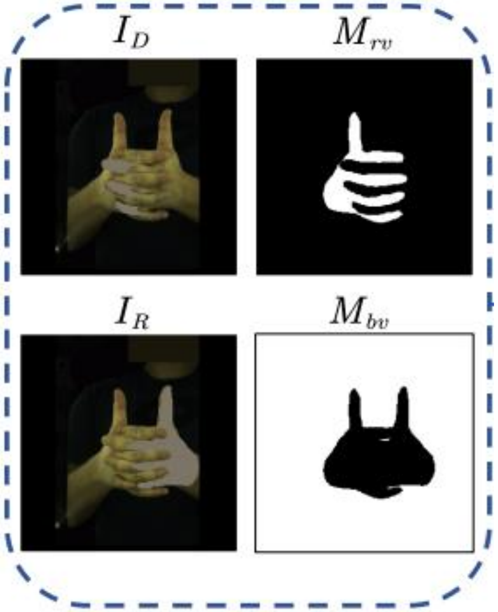
$$\mathcal{L}_{HAS} = \mathcal{L}_{BCE}\left(M_{ra}, M_{ra}^{*}\right) + \mathcal{L}_{BCE}\left(M_{lv}, M_{lv}^{*}\right) + \mathcal{L}_{BCE}\left(M_{la}, M_{la}^{*}\right) + \mathcal{L}_{BCE}\left(M_{lv}, M_{lv}^{*}\right)$$

Method - HDRM

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)







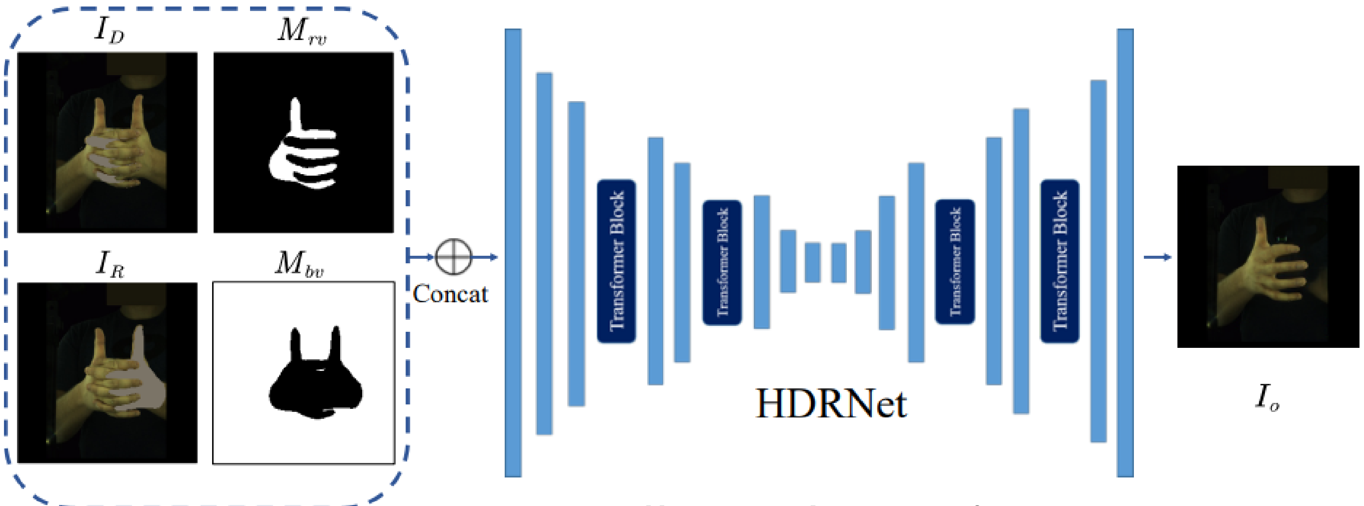
- MD denote the region where the target hand is occluded by the other hand
- MR denote the region where the distracting hand occupies

$$M_D = M_{ra} \cdot (1 - M_{rv})$$
$$M_R = (1 - M_{ra}) \cdot M_{lv}.$$

$$I_D = I_s \cdot (1 - M_D),$$
$$I_R = I_s \cdot (1 - M_R),$$
$$M_{bv} = (1 - M_{ra}) \cdot (1 - M_{la})$$

Method - HDRM

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

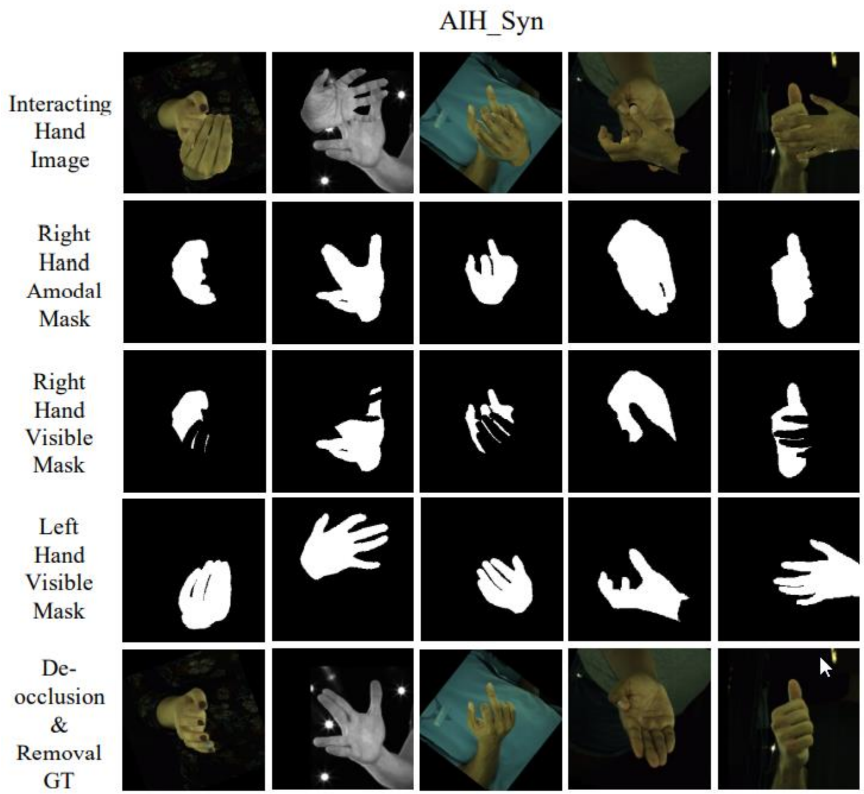


Backbone : Partial conv + transformer

$$\mathcal{L}_{HDR} = \lambda_1(\mathbb{E}_{I_o}[\log(1 - D(I_o))] + \mathbb{E}_{I_o^*}[\log(D(I_o^*))]) + \lambda_2\mathcal{L}_{\ell 1}(I_o, I_o^*) + \lambda_3\mathcal{L}_{prec}(I_o, I_o^*) + \lambda_4\mathcal{L}_{style}(I_o, I_o^*)$$

AIH Dataset

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)



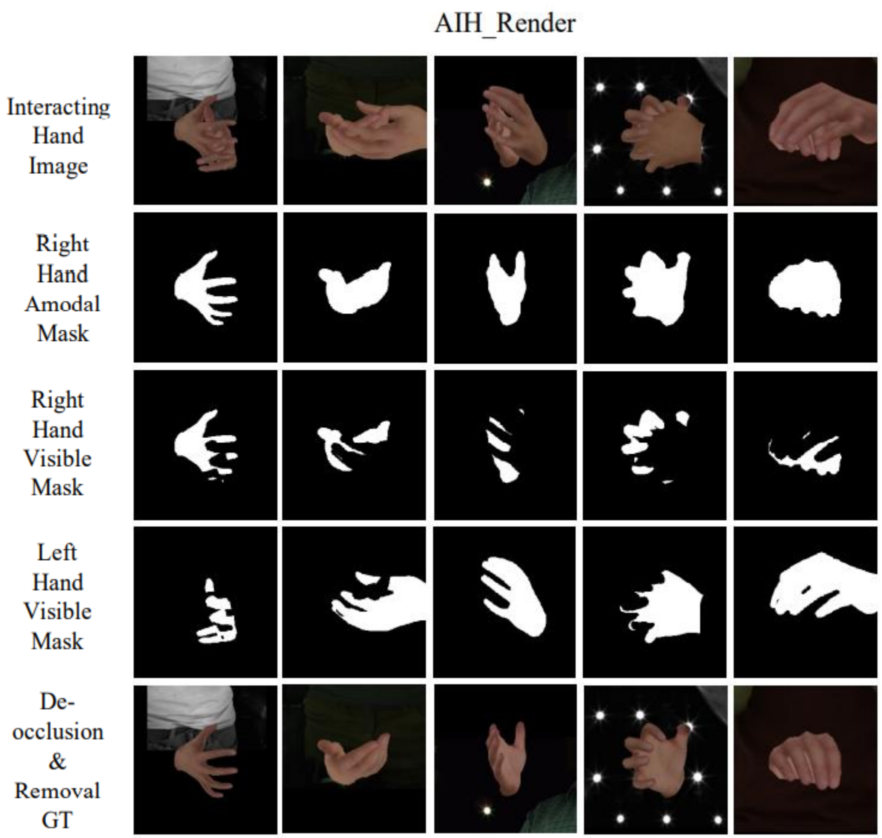
- AIH Syn
  - Single hand
    - AIH Syn contains 2.2M samples from the InterHand2.6M V1.0 dataset
    - 250K cropped single-hand images with masks
    - AIH Syn is generated by simple 2D image-level copy and paste
    - Copy the left single-hand image and paste it on the right single-hand image
  - Interacting hand
    - Two hands with similar texture from both sides
    - Then we crop the left hand region given its amodal mask and paste it on the right hand

서강대학교 SOGANG UNIVERSITY

VDS LAB

AIH Dataset

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)



- AIH Render
  - AIH Render is generated by rendering the textured 3D interacting hand mesh to the image plane.
  - Suffer from the appearance gap because the rendered texture is synthetic.
  - AIH Render contains over 0.7M samples.

Experiment

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

- Comparisons with the state-of-the-art methods
  - 'ALL' branch and the 'machine - annotator (M)' branch of InterHand2.6M V1.0 Dataset
  - MPJPE (mm) is adopted to evaluate the 3D joint estimation accuracy.

Methods	InterHand2.6M - ALL branch				InterHand2.6M - M branch		
	IH26M-SH	IH26M-IH	IH26M-ALL	IH26M-Inter	IH26M-SH	IH26M-IH	IH26M-ALL
*Boukhayma *et al.* [4]	-	-	27.14	31.46	-	-	-
*Pose2Mesh [5]	-	-	27.10	32.11	-	-	-
*BiHand [35]	-	-	25.10	28.23	-	-	-
*Rong *et al.* [27]	-	-	17.12	20.66	-	-	-
DIGIT [7]	-	14.27	-	-	-	-	-
InterNet [21]	12.16	16.02	14.21	18.04	12.52	18.04	15.28
HDR (Ours)	**8.51**	**13.12**	**10.97**	**14.74**	**8.52**	**14.98**	**11.74**

Methods	Train (M, IH26M-SH)		Train (M, IH26M-SH +AIH)	
	IH26M-IH	IH26M-ALL	IH26M-IH	IH26M-ALL
SHPE [39]	40.98	25.78	32.27	21.66
+HDR (Ours)	25.45	17.98	24.59	17.80

서강대학교 SOGANG UNIVERSITY

VDS LAB

Experiment

[3] "3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal ." (ECCV 2022)

- Qualitative Results

(a) InterHand2.6M (b) Tzionas

Q&A