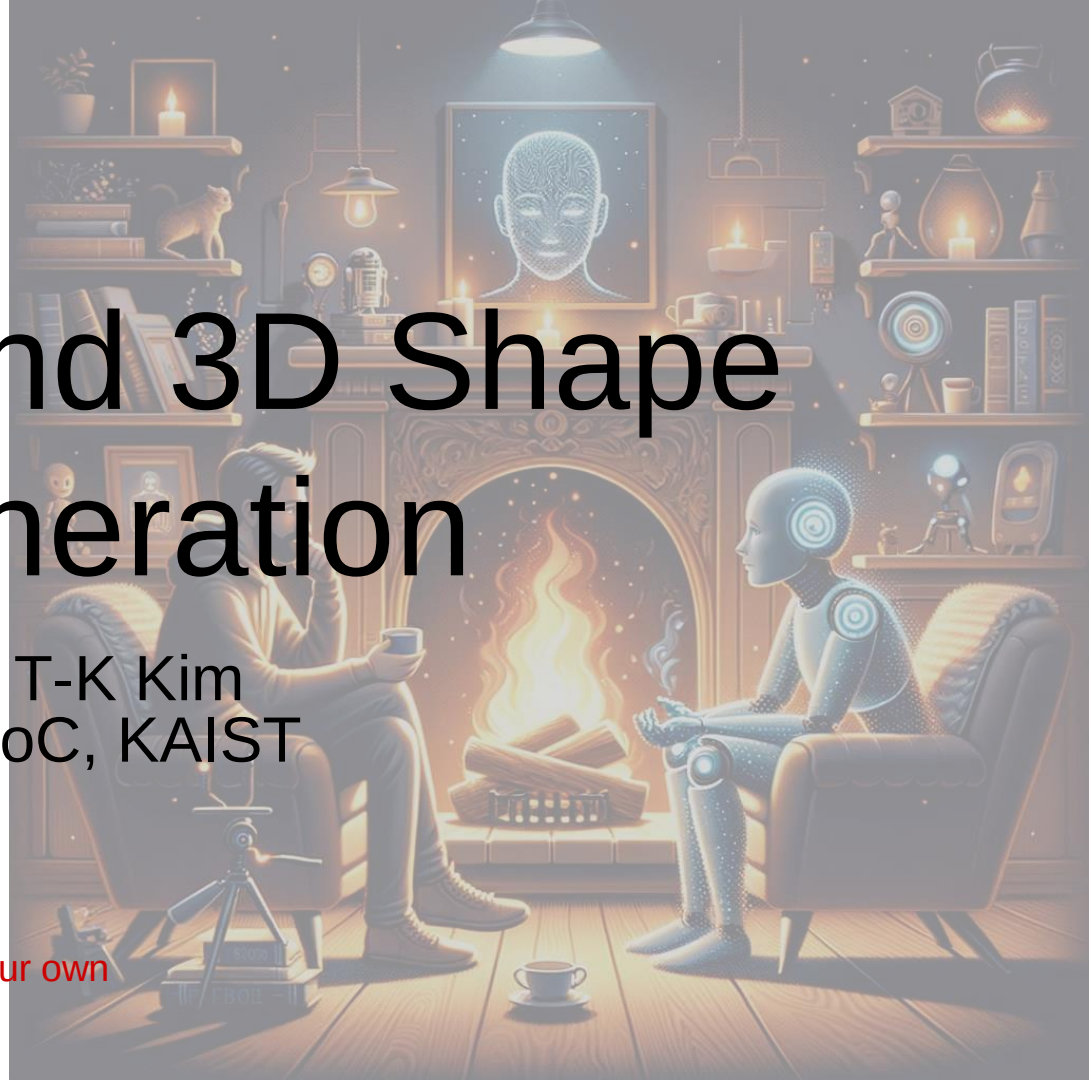
Image and 3D Shape Generation

T-K Kim
SoC, KAIST



Citation in red: our own work

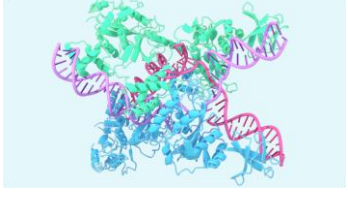
Generative (Gen) AI

LLM (chatgpt vs google bard)

Image generation (DALL-E, stable diffusion)

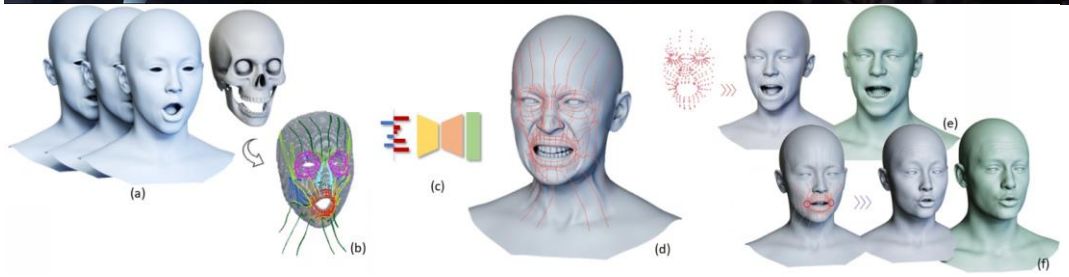
Video generation (SORA)

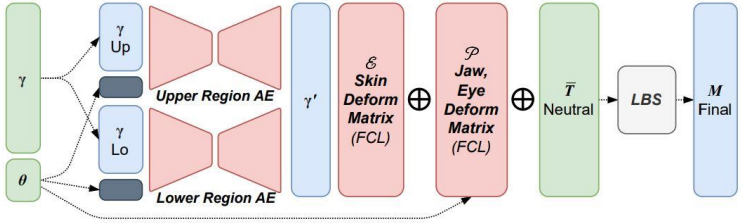
AI for medicine



Avatar 2 facial motion capture

Animatomy: an Animator-centric, Anatomically Inspired System for 3D Facial Modeling, Animation and Transfer, *SIGGRAPH Asia 2022* ([1]Wētā Digital, [2]Wētā FX, [3]Univ. of Toronto)





Li Fei-Fei Startup Officially Launched... "We Will Develop a '**World Model**' That Goes Beyond Language Models"

World Labs, a 'spatial intelligence' startup led by Stanford University professor Feifei Li, has reportedly succeeded in attracting investment of 230 million dollars (about 300 billion won). The company announced that it will develop a '**Large World Model (LWM)**' beyond the Large Language Model (LLM).

https://www.aitimes.com

Professor Lee disclosed specific information about the company on this day. He explained that the goal is to develop spatial intelligence that understands and judges the real world and to build LWM.

“The current LLM generates text and images, while the **LWM focuses on the ability to reason about how the 3D world works**,” said Professor Lee. This **spatial intelligence** “can be used in augmented reality (AR) and virtual reality (AR), as well as robotics.”

“Images and videos from generative AI models that have emerged so far **don’t adequately convey a sense of how the 3D world** is constructed,” he points out. But spatial intelligence could enable broader reasoning abilities, which could help avoid hallucinations **like counting fingers incorrectly**, he explains.

AI that understands the physical world is also a key research area for major AI companies. Overcoming the limitations of LLM, which only learns about the world through text, and acquiring knowledge of space through vision like humans is understood as **one way to achieve artificial general intelligence (AGI)**.

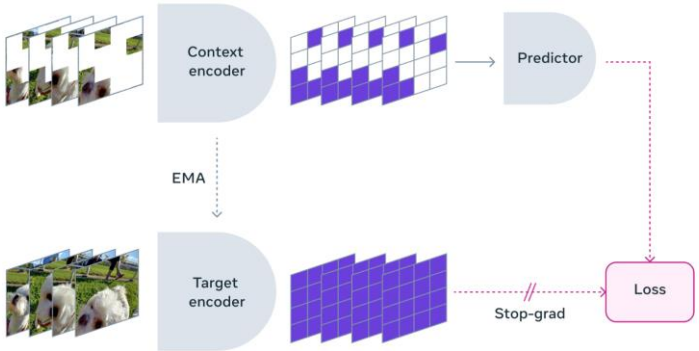
For this reason, companies aiming for AGI, including Meta, Google, OpenAI, and xAI, are all focusing on this field.

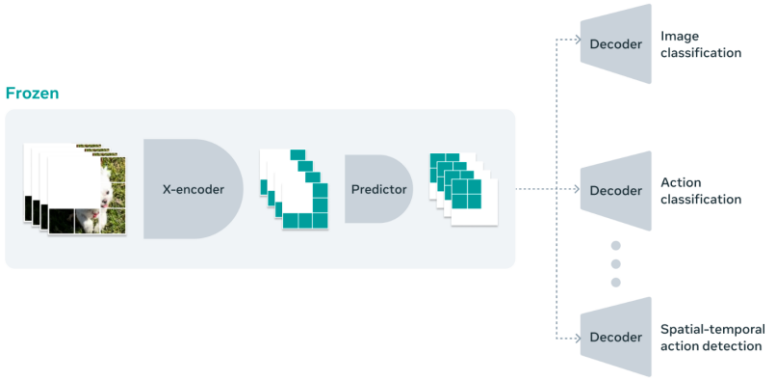
Professor Lee said that the model will be built **using images and synthetic data**, and the same **transformer-based** architecture as the LLM. However, the transformer is not everything, and other elements will be integrated, he said.

"If we want to advance AI beyond its current capabilities, we need more than AI that can see and talk. We want AI that can do things directly," he said, emphasizing that "**spatial intelligence will be the next standard that will change the direction of AI**."

Meta, vision-based AGI

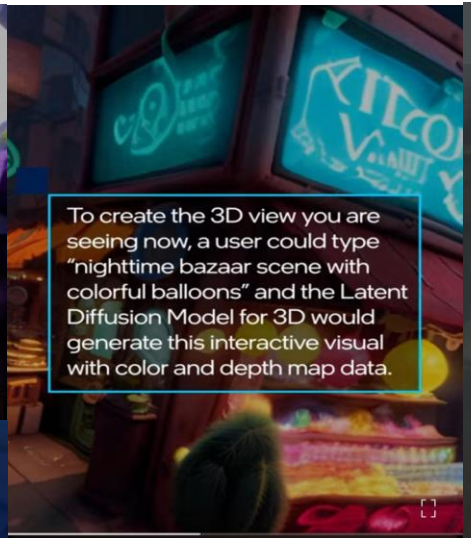
V-ZEPA: The next step toward advanced machine intelligence (AMI)
Video Joint Embedding Predictive Architecture (V-JEPA) model, a crucial step in advancing machine intelligence with a more grounded understanding of the world.
This early example of a physical world model excels at detecting and understanding highly detailed interactions between objects.
V-JEPA is a non-generative model that learns by predicting missing or masked parts of a video in an abstract representation space.
While the “V” in V-JEPA stands for “video,” it only accounts for the visual content of videos thus far. A more multimodal approach is an obvious next step, so we’re thinking carefully about incorporating audio along with the visuals.

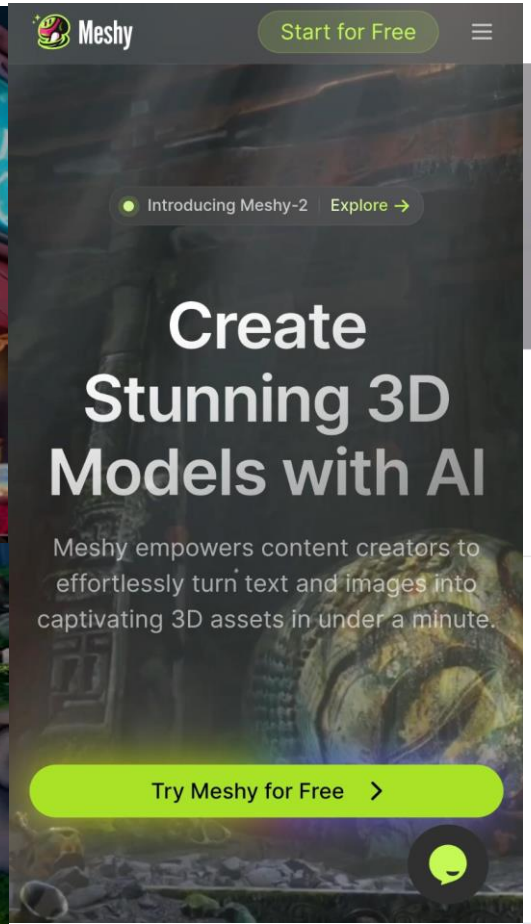


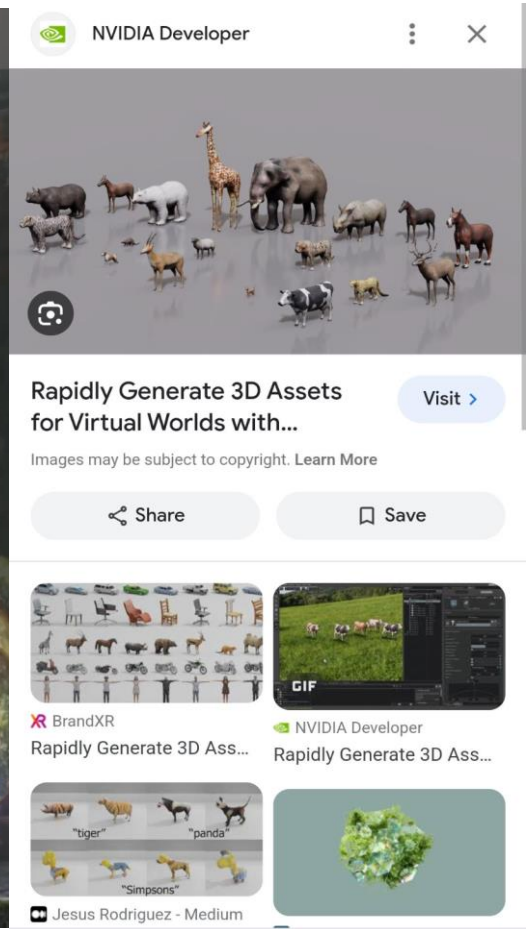


How much of this does LLM comprehend? → LWM (Large World Model)

Using powerful synthetic simulators, sim2real







Layout diffusion, 3D scene generation

Total3DUnderstanding(CVPR 2020)

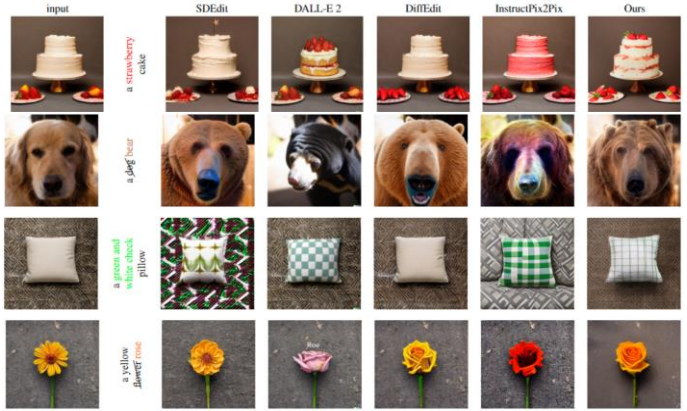
LayoutGPT(NeuralIPS 2023)

ATISS(NeurIPS, 2021)

"Text2Immersion: Generative Immersive Scene with 3D Gaussians." arXiv 2023

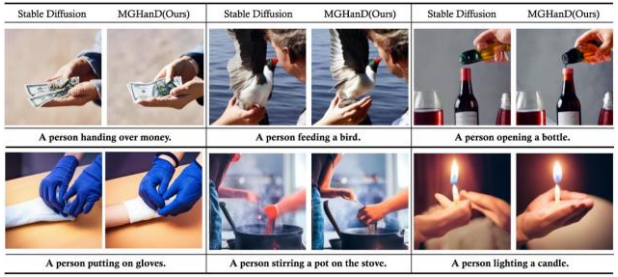
Video diffusion, consistency of continuous frame generation

VR2024



CVPR2024

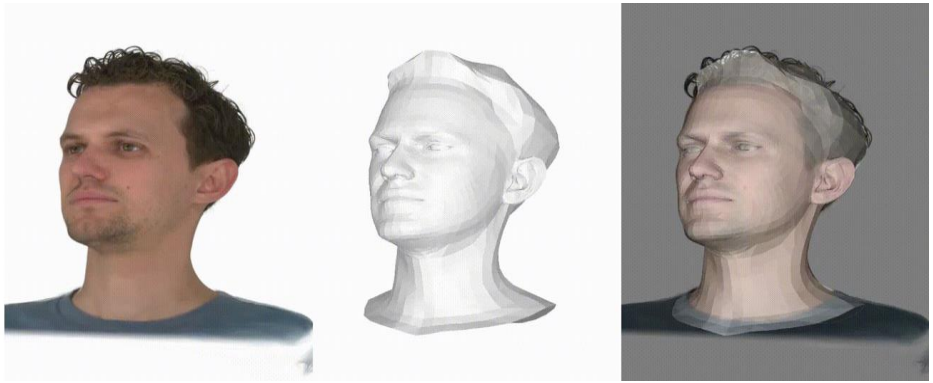
CVPR2024



CVPR 2024 papers

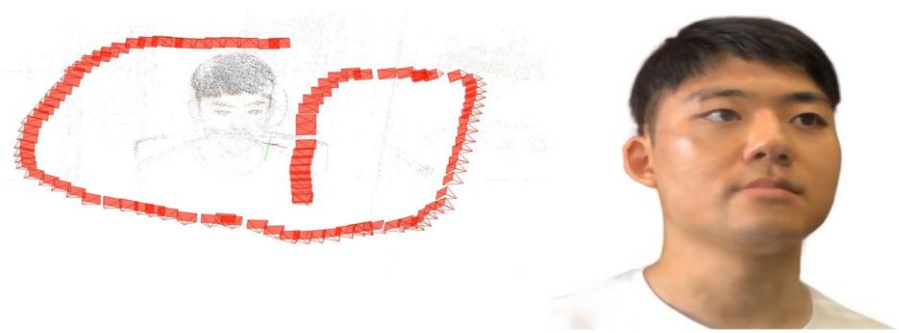
...

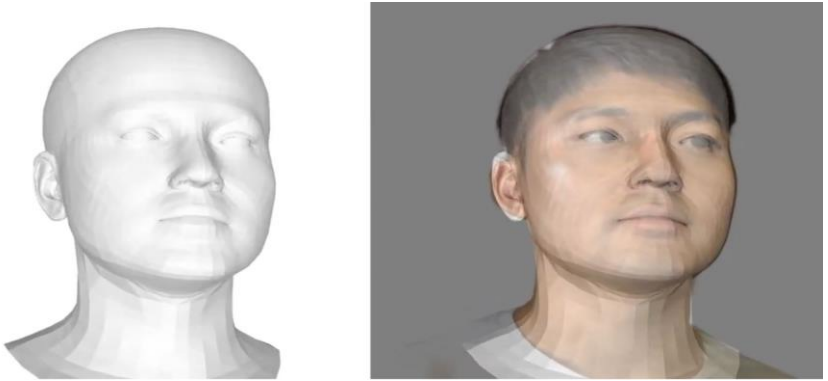
Hyper realistic 3D capture and rendering, Gaussian Splatting



Gaussian Avatars CVPR24,

Quality degrades with less number of images: 15->8->4 views

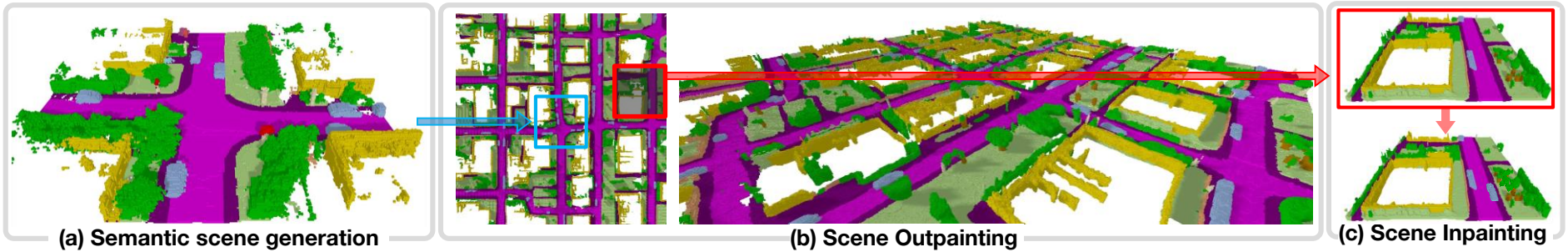




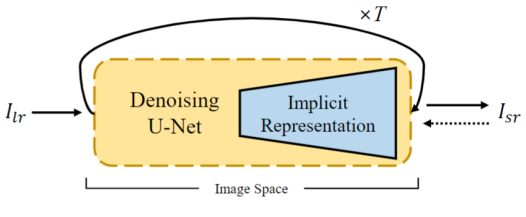
Using a free motion camera, head avatar generation.
Regularization loss to improve FLAME/GS.

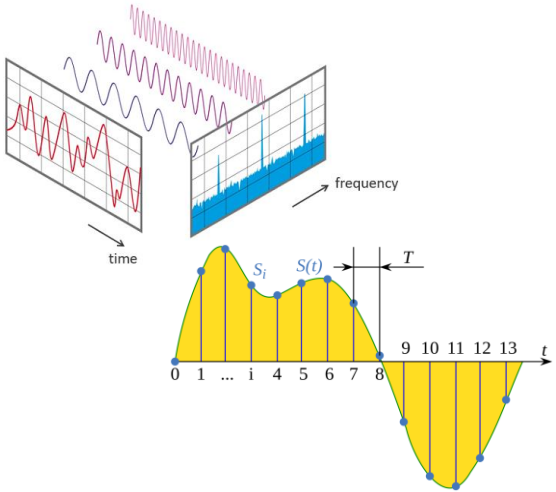
Upscaling to bring back contents

Arbitrary-scale upscaling



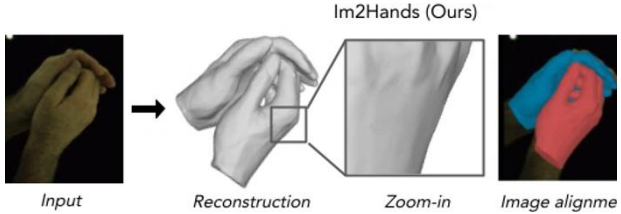
S. Yoon, SoC KAIST





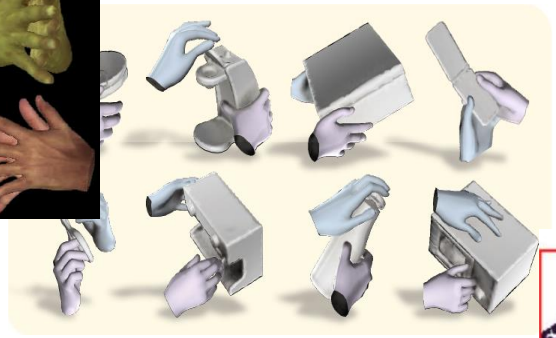
CVPR2024

Compositional Generalisation

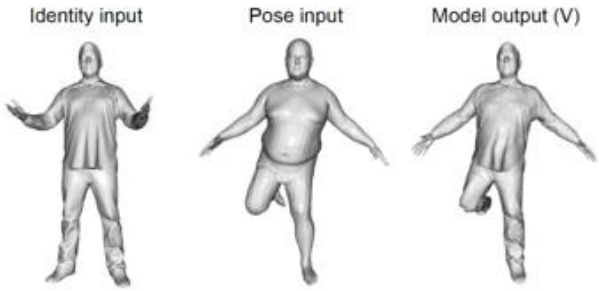


CVPR2023

CVPR2024

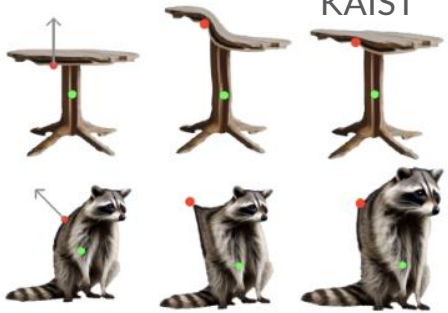


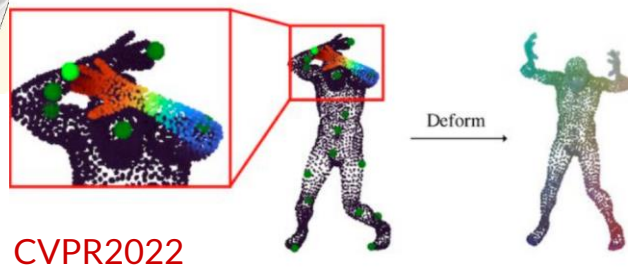
CVPR2024



ICCV2023

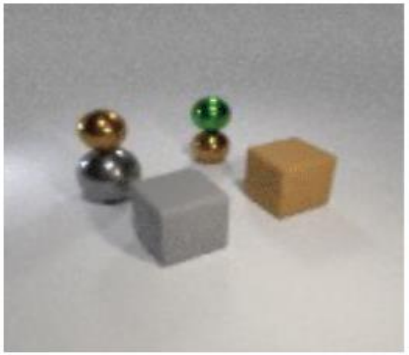
M. Sung, SoC
KAIST

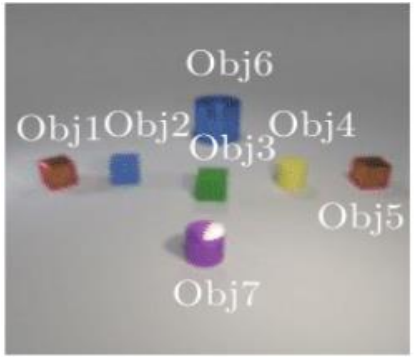


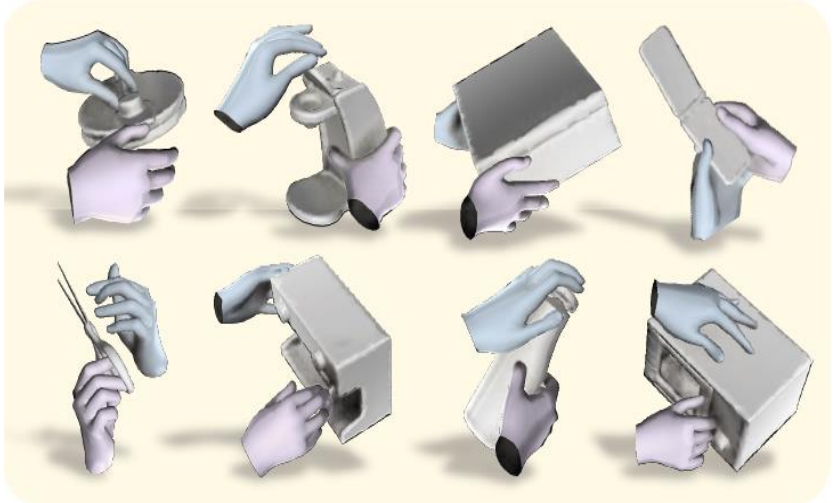


CVPR2022

Compositional Generalisation

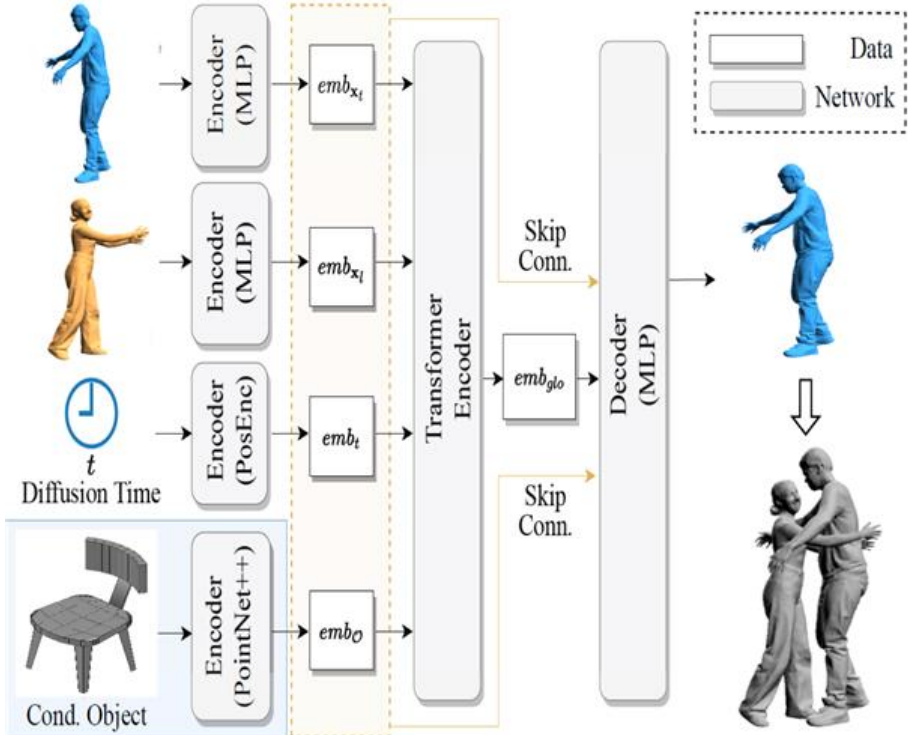




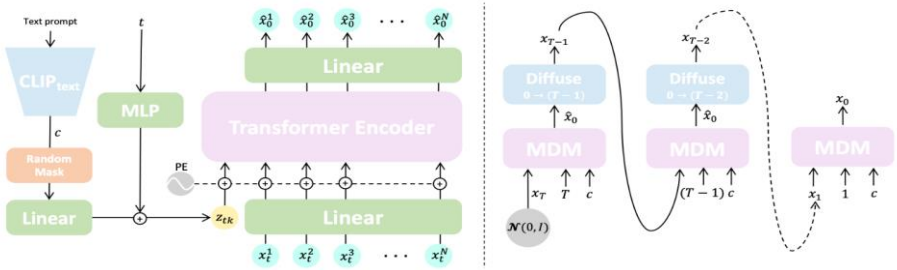


CVPR2024

S. Ahn, SoC
KAIST

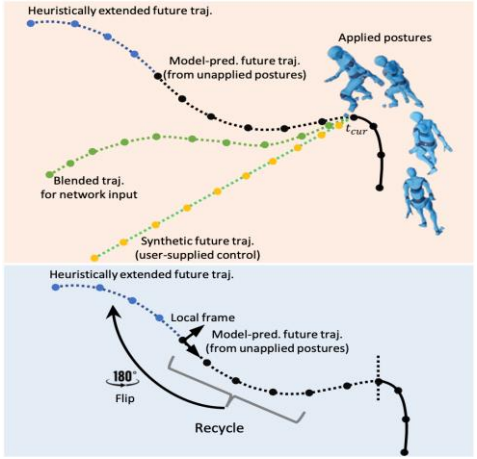


Motion Diffusion

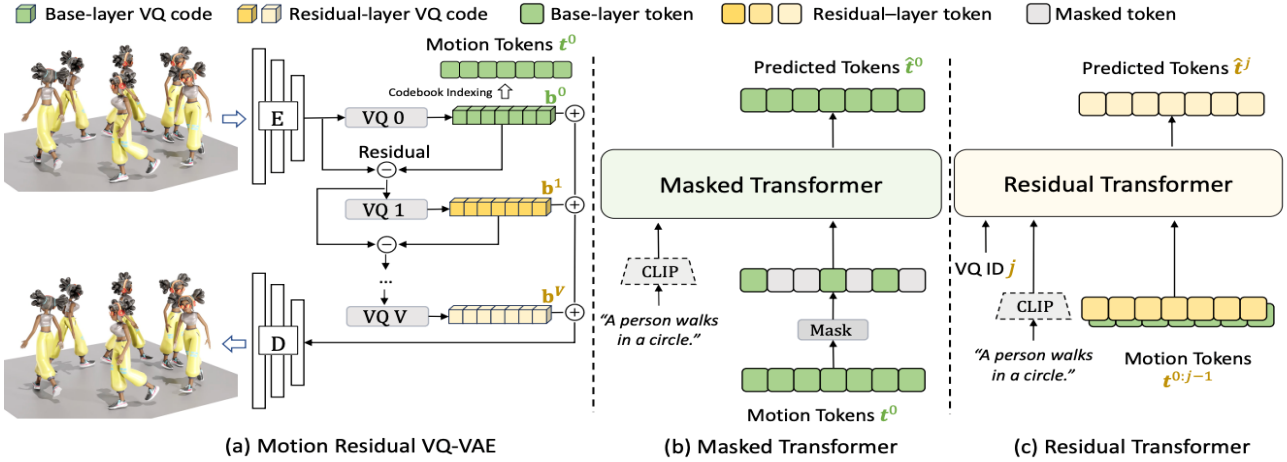


Motion Diffusion Model, ICLR 2023

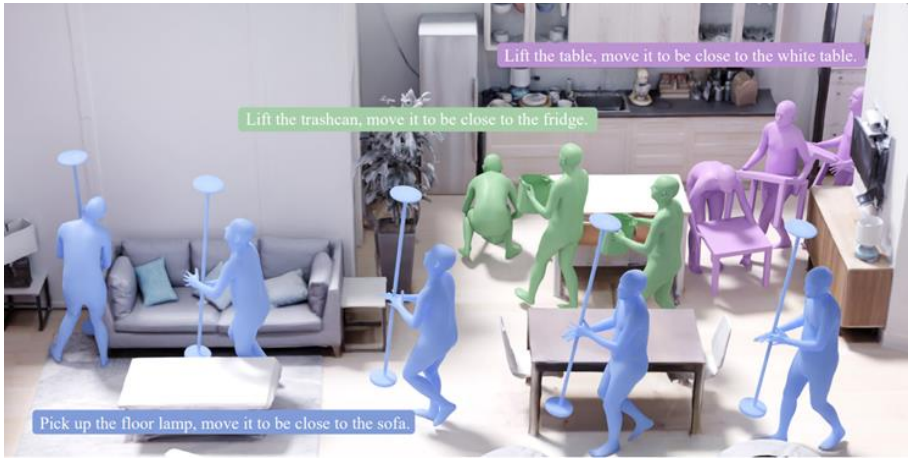
Taming Diffusion Probabilistic Models for Character Control, SIGGRAPH 2024,



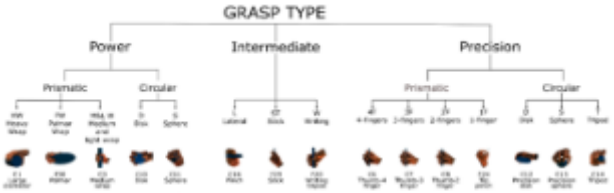
MoMask: Generative Masked Modeling of 3D Human Motions, CVPR 2024



Motion Generation with Physics



arXiv:2312.03913



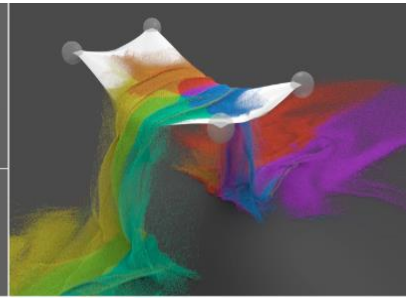
GRASPNET, ECCV 2024

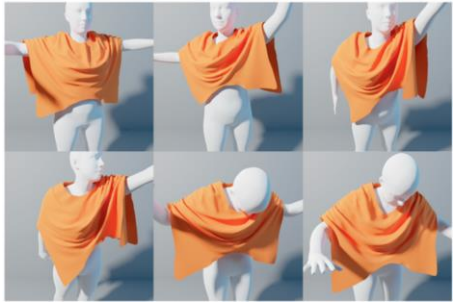
https://youtu.be/fiu5canEgOA

Bi-dexhands (Yuanpei et al., 2023)

IROS2020

PhysGaussian
CVPR 2024





KAIST COMPUTER VISION AND LEARNING LAB (KCVL)

TAE-KYUN (T-K) KIM, PHD

Professor

School of Computing, KAIST

연구책임자

KAIST 전산학부 김태균

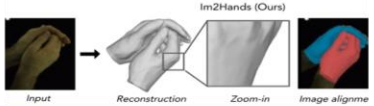
- 연구실명 : KAIST Computer Vision and Learning Lab (https://sites.google.com/view/tkkim/home)
- 현 연구실 구성원 : 석사과정 14명, 박사과정 6명
- 대표 약력
  - KAIST 전산학부, 교수 (2020-)
  - Imperial College London 겸직 교수
  - KAIST 공대 Impact Research Award 수상 (2022)
  - Imperial College London, EEE, 조교수, 정년 부교수 (2010-2020)
  - BMVC program chair (2023), BMVC general chair (2017)
  - CVPR best paper finalist (2020)
  - ASCE Journal of Computing best paper award (2016), ICRA best paper award (2014)

대표성과 (5개 이내)

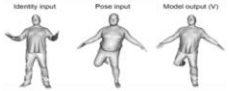
- (Google Scholar) h-index = 62, i10-index = 154, total Citation = 14,215
- Multiple object tracking: A literature review, W Luo, J Xing, A Milan, X Zhang, W Liu, X Zhao, TK Kim, Journal of Artificial Intelligence (+1200 citations)
- Discriminative learning and recognition of image set classes using canonical correlations, TK Kim, J Kittler, R Cipolla, IEEE Trans. on PAMI 29 (6), 1005-1018 (Image set을 이용한 새로운 인식 방법을 제안, +800 citations)
- First-person hand action benchmark with RGB-D videos and 3D hand pose annotations, G Garcia-Hernando, S Yuan, S Baek, TK Kim, Proc. of IEEE Conf. on CVPR 2018 (관련 가장 널리 사용되는 benchmark중 하나로, +120개 이상의 해외 기관들이 사용중)

Expert group in Generative AI based on 3D Vision

CVPR2022



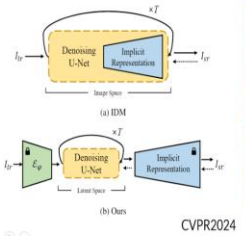
CVPR2023



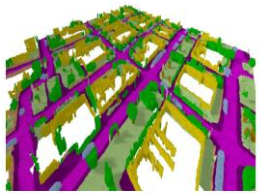
ICCV2023

CVPR2024

CVPR2024



CVPR2024



Scene Outpainting

VR2024

InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Diffusion

CVPR 2024


Jihyun Lee [1]


Shunsuke Saito [2]

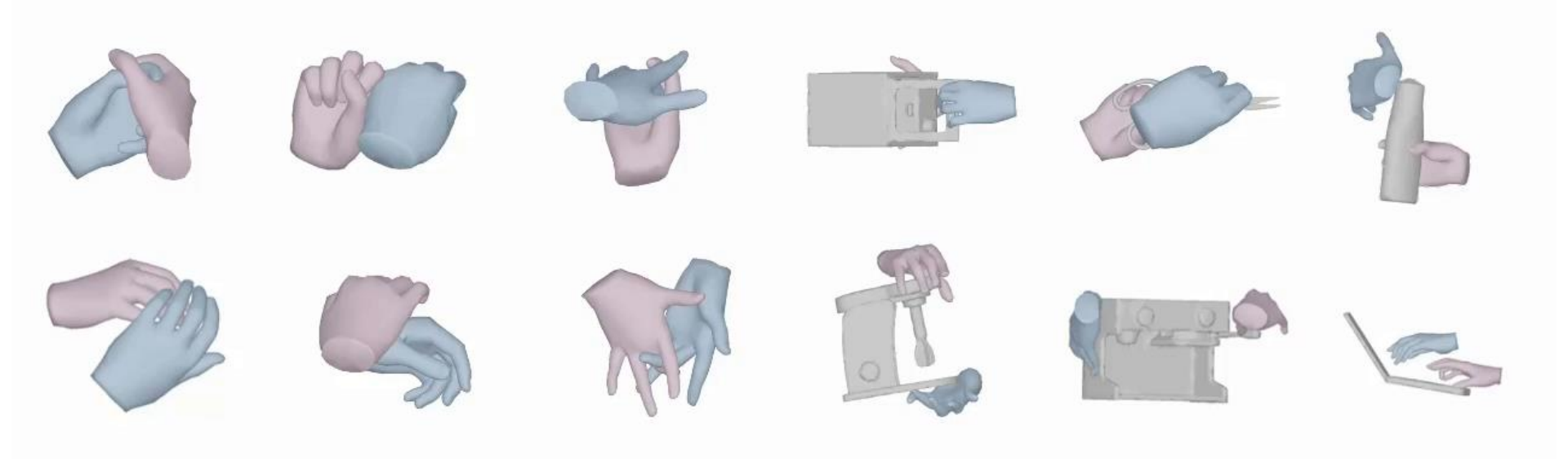

Giljoo Nam [2]


Minhyuk Sung [1]

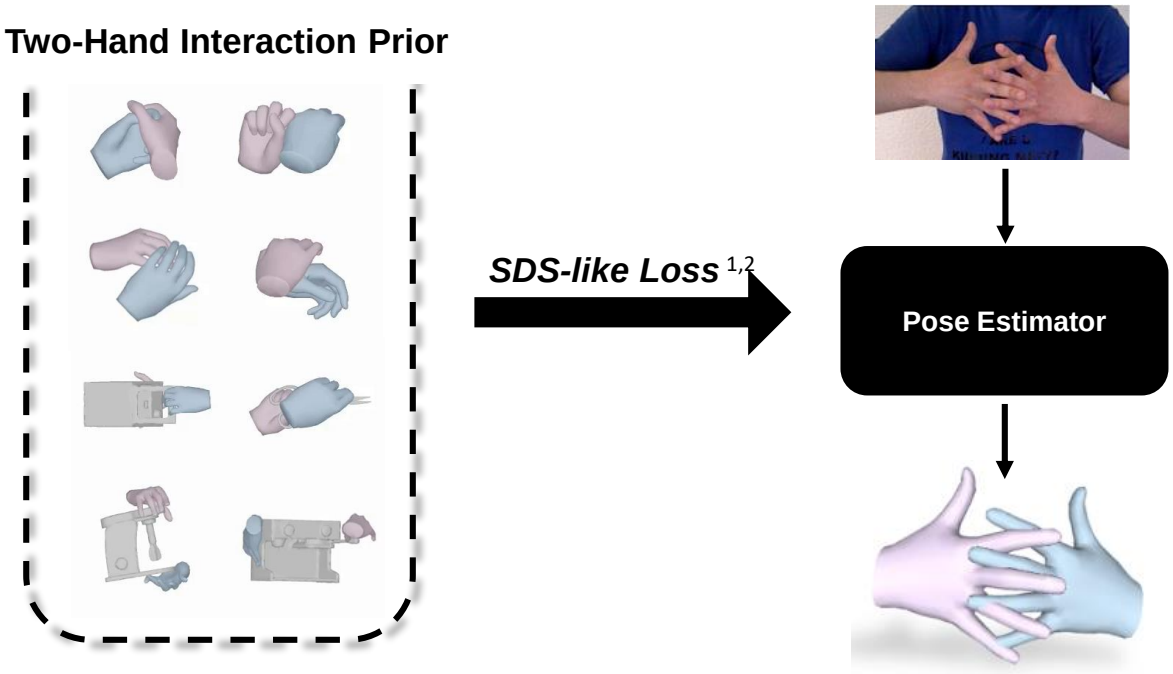

Tae-Kyun (T-K) Kim [1,3]

KAIST [1] Meta [2] Imperial College London [3]

Overview



Overview

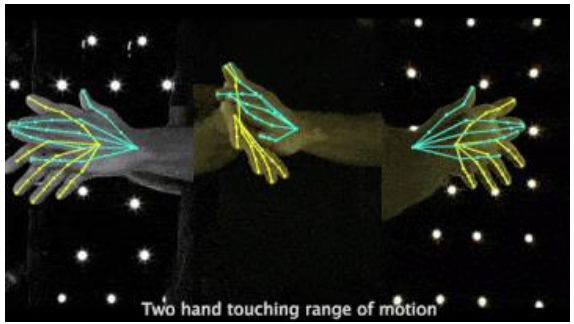


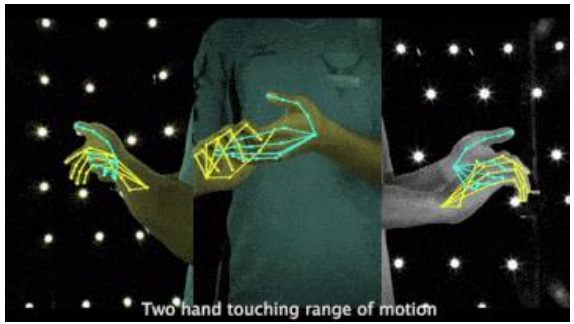
[1] Poole *et al.*, Dreamfusion: Text-to-3d using 2d diffusion. In ICLR, 2022.
[2] Müller *et al.*, Generative proxemics: A prior for 3d social interaction from images. In CVPR, 2024.

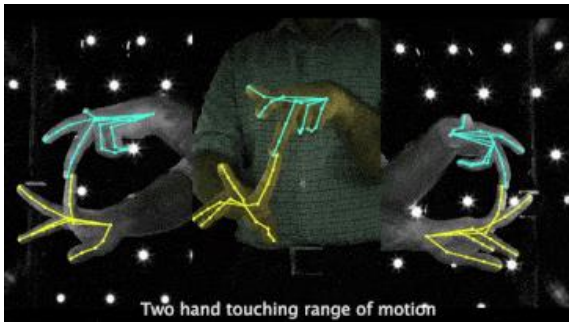
Motivation

*“Two-hand interaction **generation** (cf. reconstruction) remains under-explored.”*

InterHand2.6M [1]



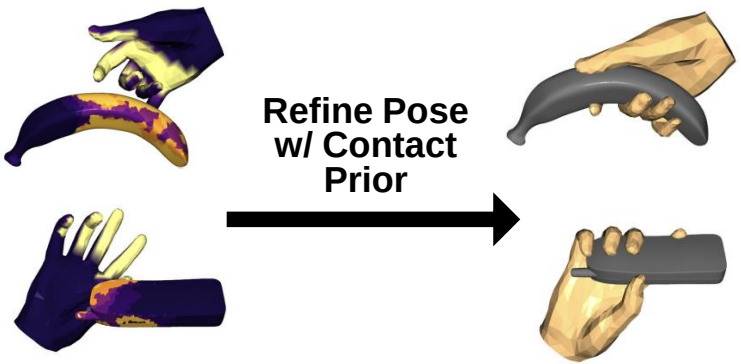




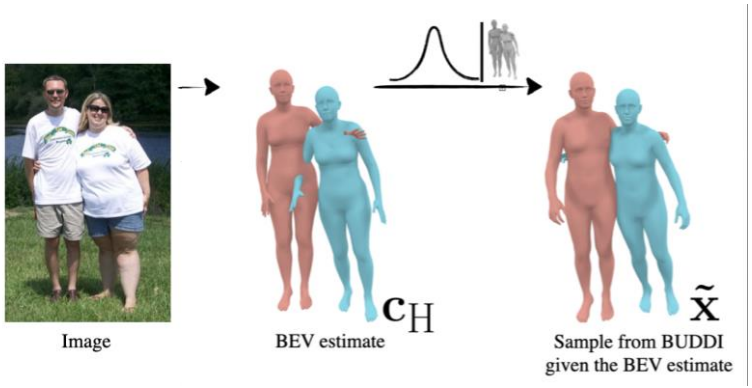
[1] Moon *et al.*, Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In ECCV, 2020.

Motivation

ContactOpt [1]



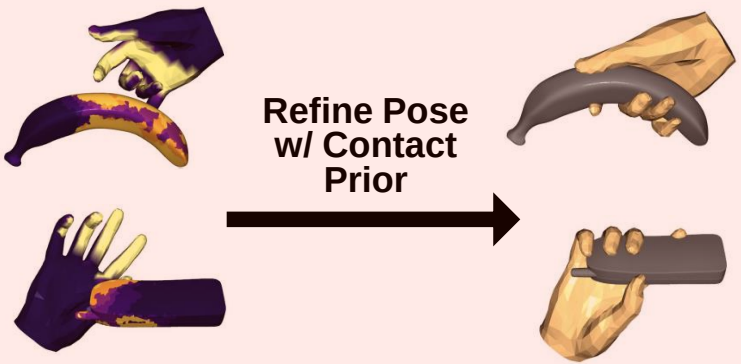
BUDDI [2]



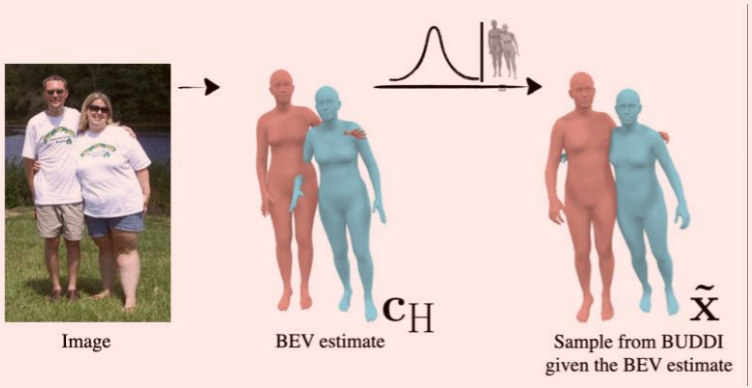
[1] Grady *et al.*, ContactOpt: Optimizing Contact to Improve Grasps. In CVPR, 2021.
[2] Müller *et al.*, Generative proxemics: A prior for 3d social interaction from images. In CVPR, 2024.

Motivation

ContactOpt [1]



BUDDI [2]



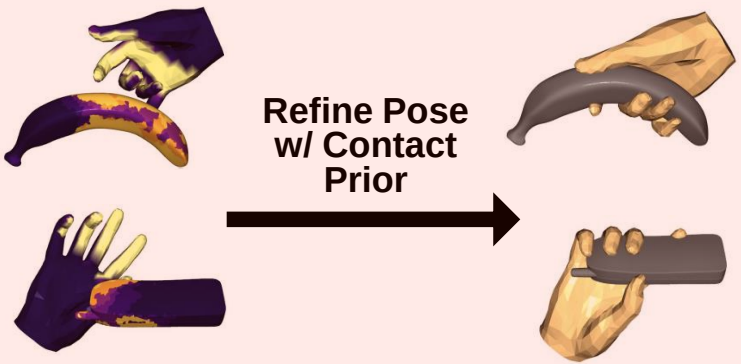
→ ***Adapt these for generating two-hand interaction?***

🤔

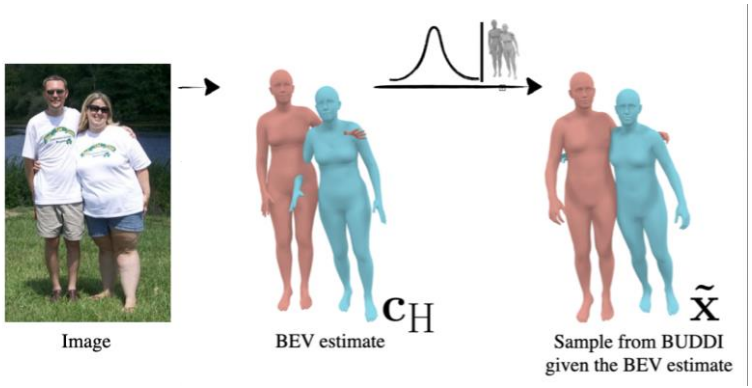
[1] Grady *et al.*, ContactOpt: Optimizing Contact to Improve Grasps. In CVPR, 2021.
[2] Müller *et al.*, Generative proxemics: A prior for 3d social interaction from images. In CVPR, 2024.

Motivation

ContactOpt [1]



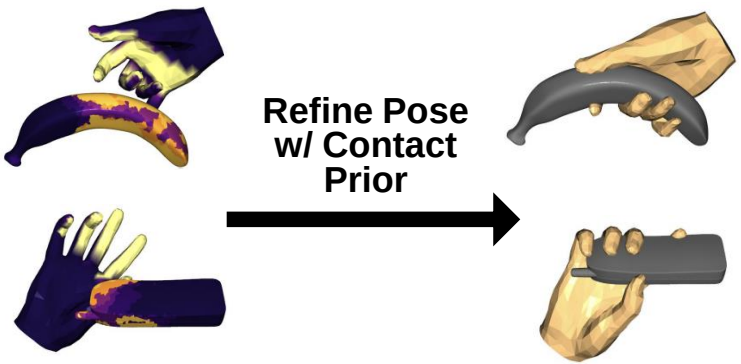
BUDDI [2]



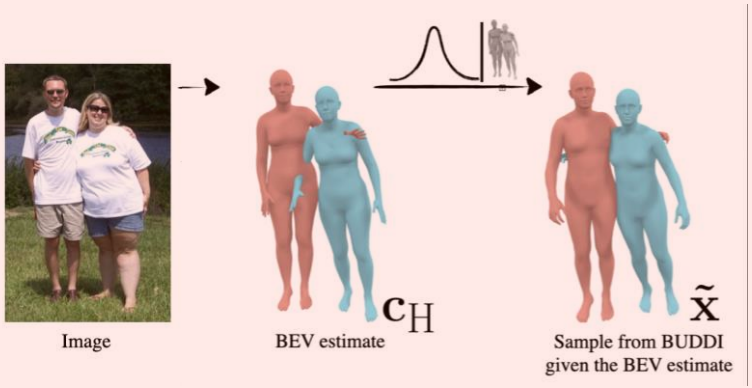
[1] Grady *et al.*, ContactOpt: Optimizing Contact to Improve Grasps. In CVPR, 2021.
[2] Müller *et al.*, Generative proxemics: A prior for 3d social interaction from images. In CVPR, 2024.

Motivation

ContactOpt [1]

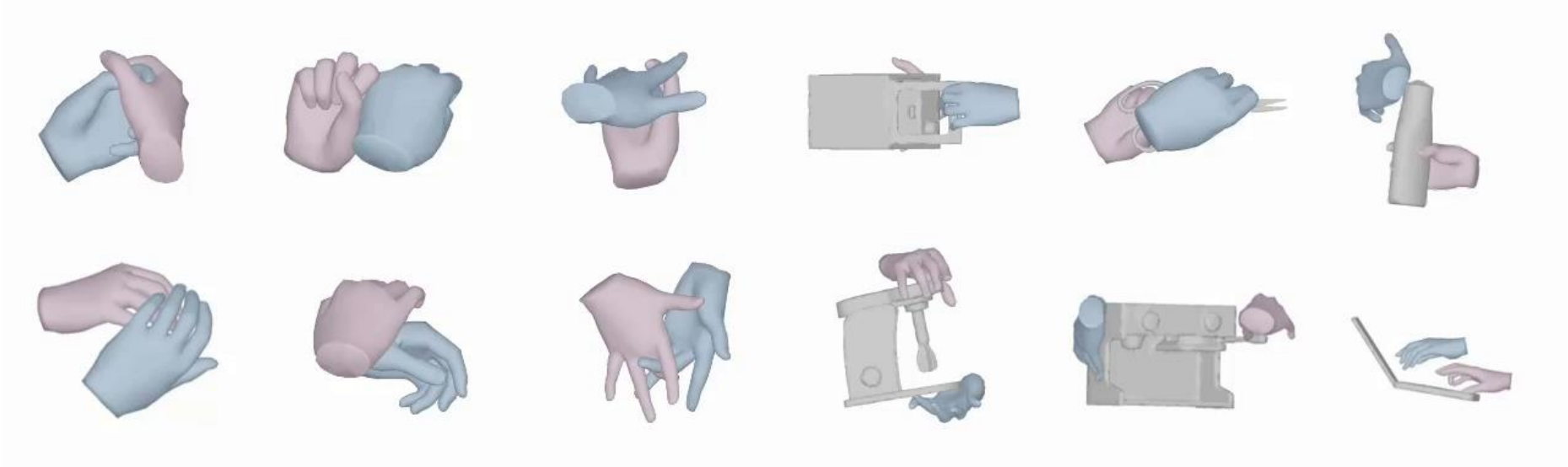


BUDDI [2]

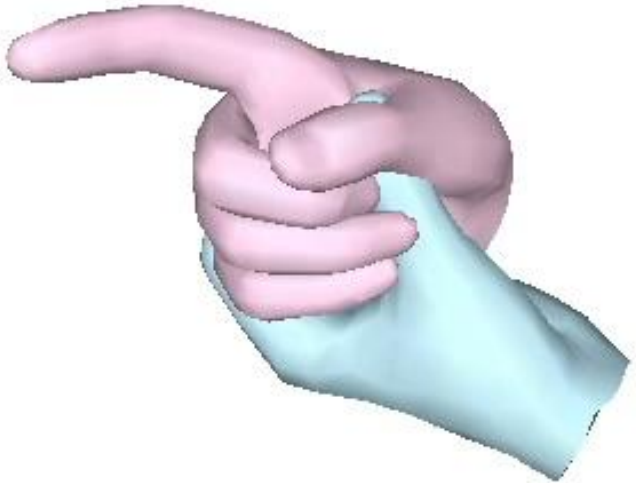


[1] Grady *et al.*, ContactOpt: Optimizing Contact to Improve Grasps. In CVPR, 2021.
[2] Müller *et al.*, Generative proxemics: A prior for 3d social interaction from images. In CVPR, 2024.

InterHandGen



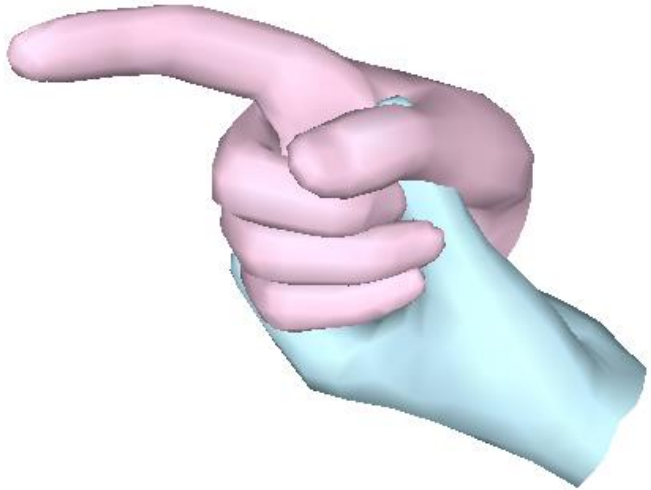
Technical Challenge



Combination of two hands with diverse articulations

Key Idea: Distribution Reformulation

$$p_\phi(\mathbf{x}_l, \mathbf{x}_r) = p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r | \mathbf{x}_l)$$



Key Idea: Distribution Reformulation

$$p_\phi(\mathbf{x}_l, \mathbf{x}_r) = p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r | \mathbf{x}_l)$$

Extension to object-conditional two-hand generation

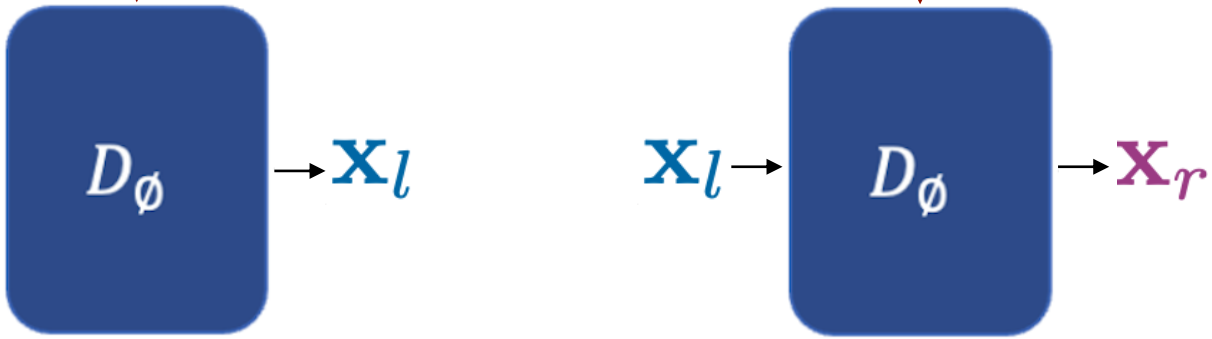
$$p_\phi(\mathbf{x}_l, \mathbf{x}_r | \mathbf{c}) = p_\phi(\mathbf{x}_l | \mathbf{c})\, p_\phi(\mathbf{x}_r | \mathbf{x}_l, \mathbf{c})$$

Object condition (PointNet++[1] feature)

[1] Qi *et al.*, PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In NeurIPS, 2017.

Training: “Conditioning Hand Dropout”

$$p_\phi(\mathbf{x}_l, \mathbf{x}_r) = p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$

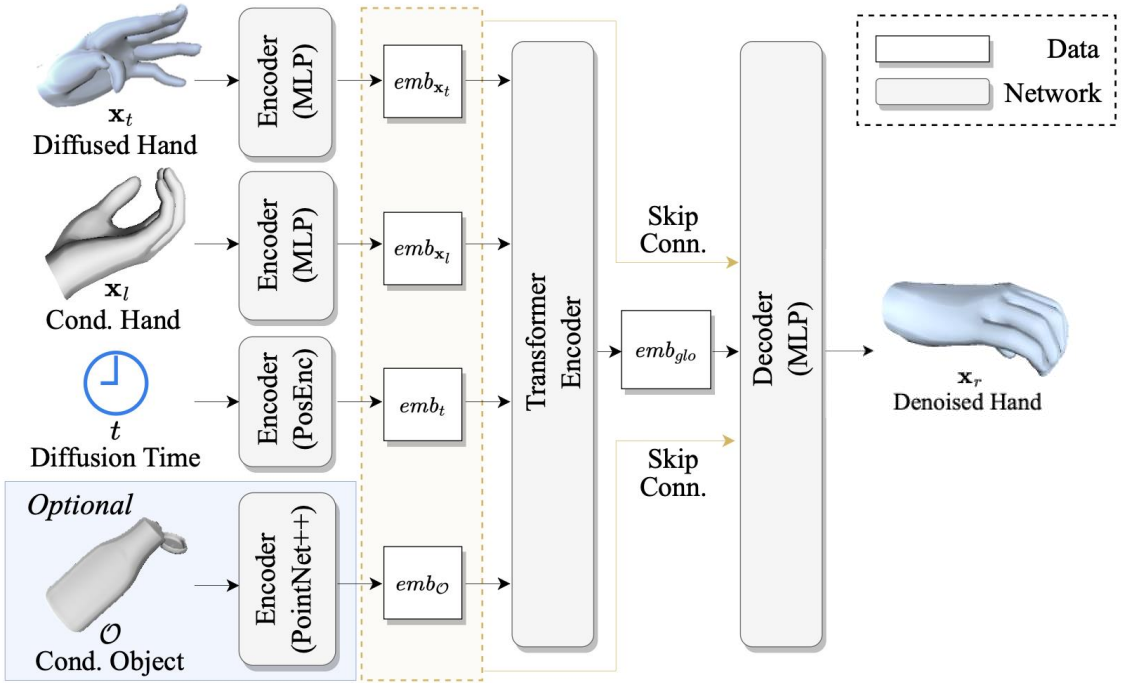


Training: “Conditioning Hand Dropout”

$$\mathbf{x}_r \; p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r | \mathbf{x}_l)$$

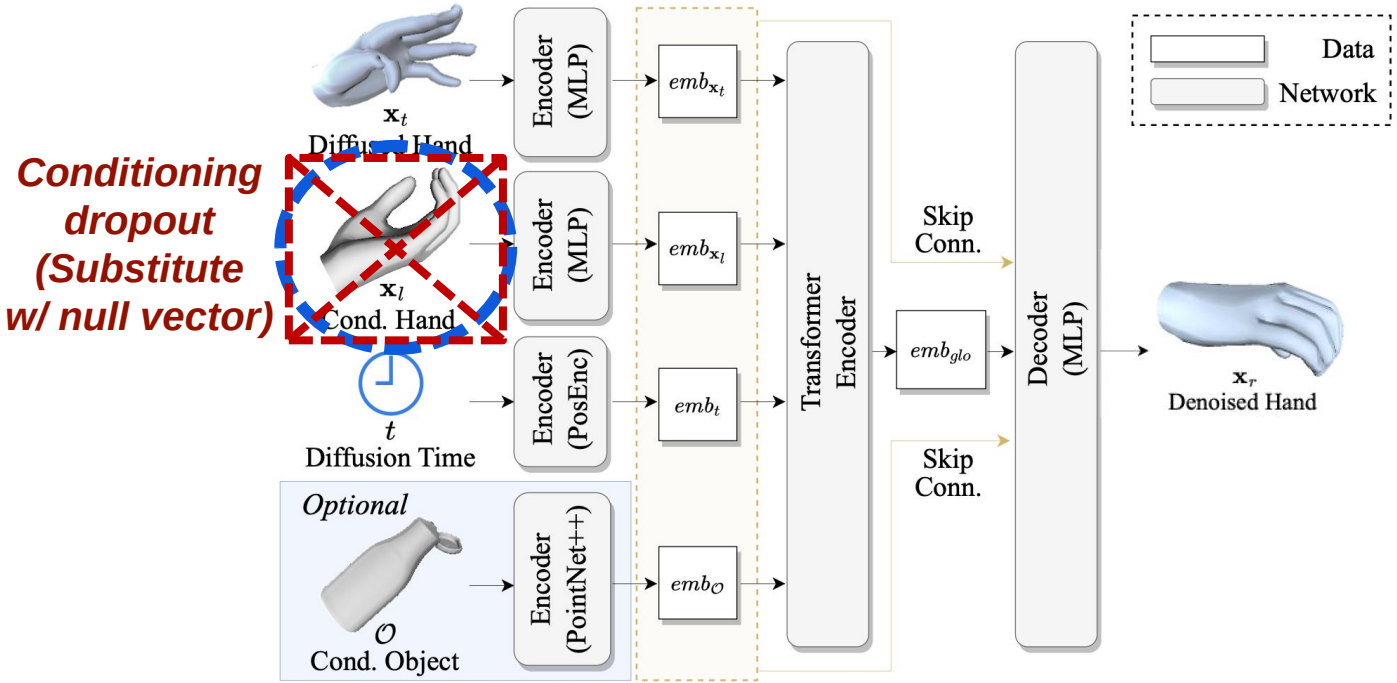
Training: “Conditioning Hand Dropout”

$$p_\phi(\mathbf{x}_r) \qquad p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$



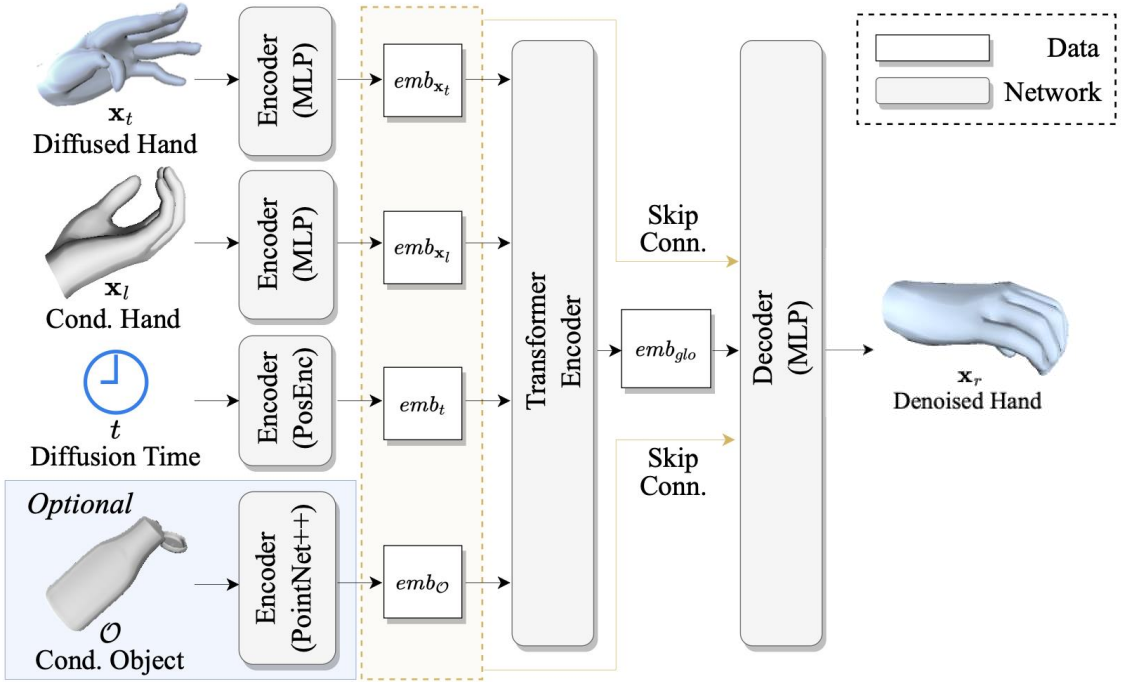
Training: “Conditioning Hand Dropout”

$$p_\phi(\mathbf{x}_r) \qquad p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$



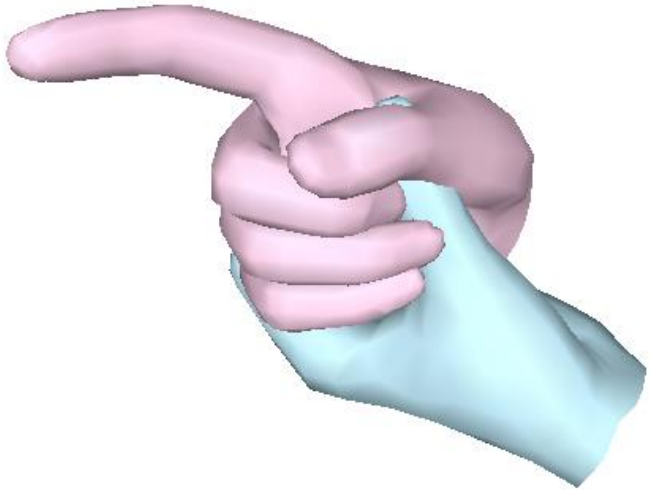
Loss Function

$$\nabla_\phi \left\| \mathbf{x}_r - D_\phi(\mathbf{x}_t, \mathbf{x}_l, t)) \right\|^2$$



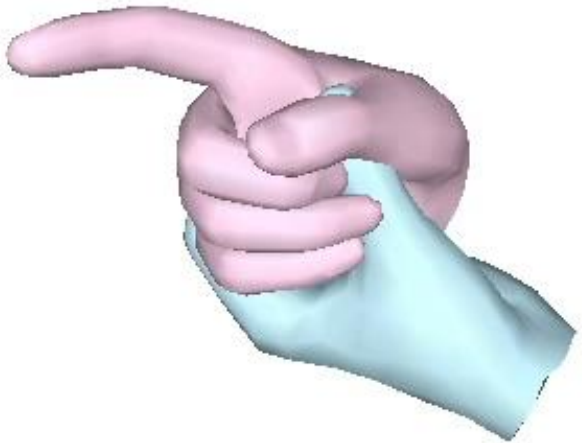
Inference: “Cascaded Hand Denoising”

$$p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$



Inference: “Cascaded Hand Denoising”

$$p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$



Classifier-Free Guidance [1]

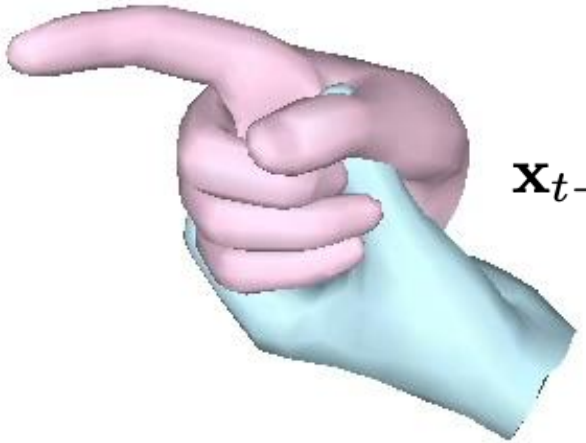
$$\hat{\epsilon} \leftarrow (1 + w_{cfg})\hat{\epsilon}_{cond} - w_{cfg}\hat{\epsilon}_{uncond}$$

→ ***To control fidelity and diversity***

[1] Ho and Salimans, Classifier-free diffusion guidance. In NeurIPS Workshops, 2021.

Inference: “Cascaded Hand Denoising”

$$p_\phi(\mathbf{x}_l)\, p_\phi(\mathbf{x}_r|\mathbf{x}_l)$$



Anti-penetration guidance

$$\mathbf{x}_{t-1} \leftarrow \mathbf{x}_{t-1} - w_{pen} \nabla_{\mathbf{x}_{t-1}} \mathcal{L}_{pen}(\mathbf{x}_{t-1}, \mathbf{x}_l)$$

→ ***To avoid penetration***

Results on Two-Hand Synthesis

(a) **Comparisons on two-hand interaction generation.**

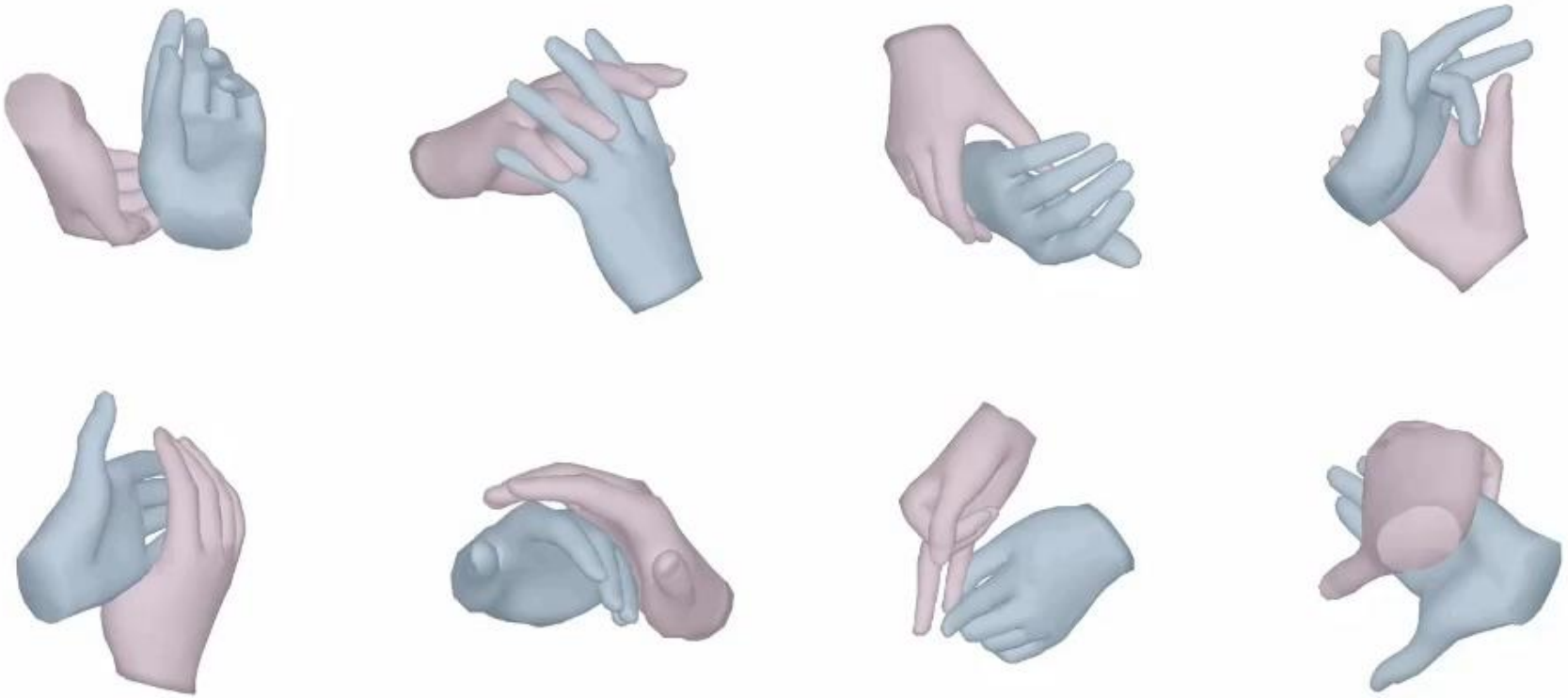
Method	FHID ↓	KHID $(\times 10^{-1})$ ↓	Diversity ↑	Precision ↑	Recall ↑	PenVol (cm^3) ↓
VAE [6]	8.18	6.23	2.32	0.55	0.02	7.32
BUDDI* [2]	3.48	4.10	2.71	0.56	0.47	0.82
Ours	**1.00**	**0.15**	**3.59**	**0.86**	**0.85**	**0.76**

(b) **Comparisons on object-conditioned two-hand interaction generation.**

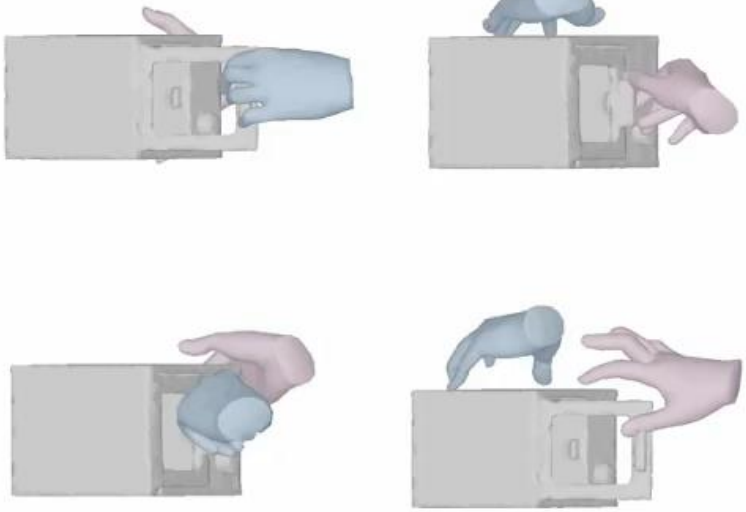
Method	FHID ↓	KHID $(\times 10^{-1})$ ↓	Diversity ↑	Precision ↑	Recall ↑	PenVol (cm^3) ↓
ContactGen* [7]	22.56	1.58	6.70	0.21	0.37	1.80
VAE [6]	21.75	2.12	5.29	0.60	0.17	4.98
BUDDI* [2]	22.51	1.35	6.50	0.28	0.36	1.38
Ours	**12.91**	**0.55**	**6.77**	**0.71**	**0.67**	**1.33**

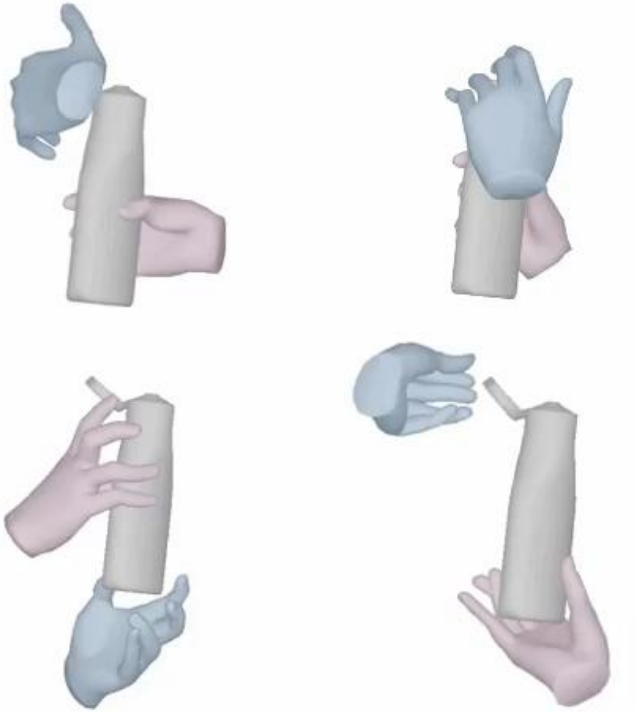
→ *FHID and KHID denote FID and KID measured using hand interaction features.*

Results on Two-Hand Synthesis



Results on Two-Hand Synthesis





Application: In-the-Wild Two-Hand Reconstruction

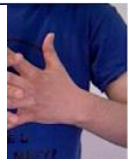
Two-Hand In

(a) **Results on InterHand2.6M [8].**

Method	MPVPE ↓	MPJPE ↓	MPRPE ↓
InterWild [7]	13.01	14.83	29.29
InterWild [7] + Ours	**12.10**	**14.53**	**26.56**

(b) **Results on HIC [9].**

Method	MPVPE ↓	MPJPE ↓	MPRPE ↓
InterWild [7]	15.70	16.17	31.35
InterWild [7] + Ours	**15.04**	**15.45**	**26.63**



timator

Application: In-the-Wild Two-Hand Reconstruction

Input Video

Reconstruction
InterWild [7]

Reconstruction
InterWild [7] + Our Prior

InterHandGen: Two-Hand Interaction Generation via Cascaded Reverse Diffusion

CVPR 2024

Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

CVPR 2024

J Kim, TK Kim

KAIST

Imperial College London

Arbitrary-scale upscaling

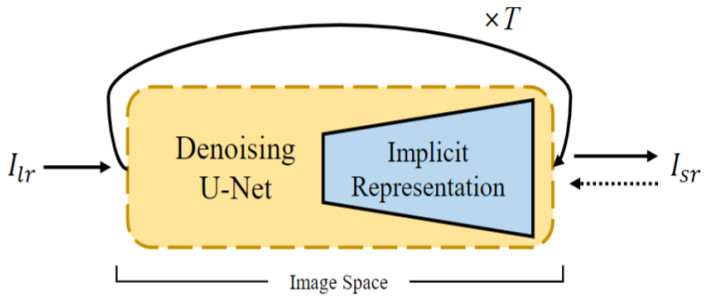
- Most existing SR and image generation methods, however, generate images only at fixed-scale magnification and suffer from over-smoothing and artifacts.
- Additionally, they do not offer enough diversity of output images nor image consistency at different scales.

(a) Arbitrary-Scale Image Generation

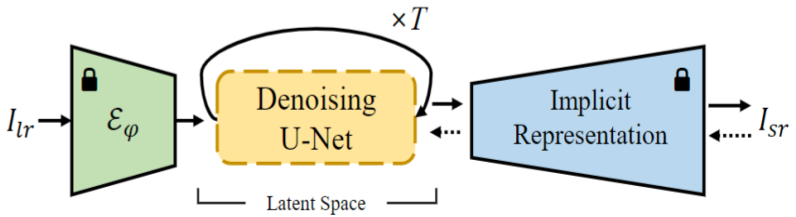
(b) Arbitrary-Scale Super-Resolution

Proposed method

- We propose a novel pipeline that can super-resolve an input image or generate from a random noise a novel image at arbitrary scales.
- The method consists of a pre-trained auto-encoder, a latent diffusion model, and an implicit neural decoder, and their learning strategies.
- The proposed method adopts diffusion processes in a latent space, thus efficient, yet aligned with output image space decoded by MLPs at arbitrary scales.



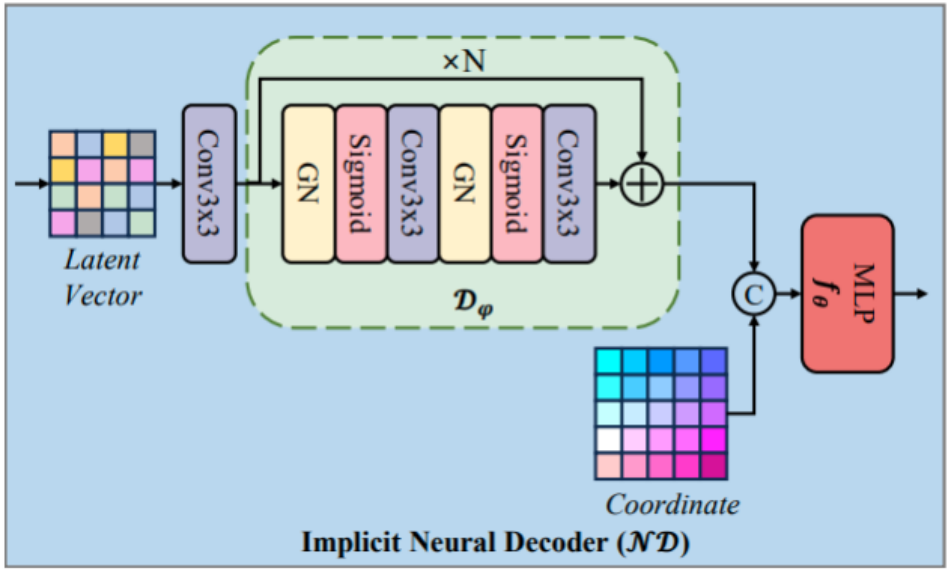
(a) IDM



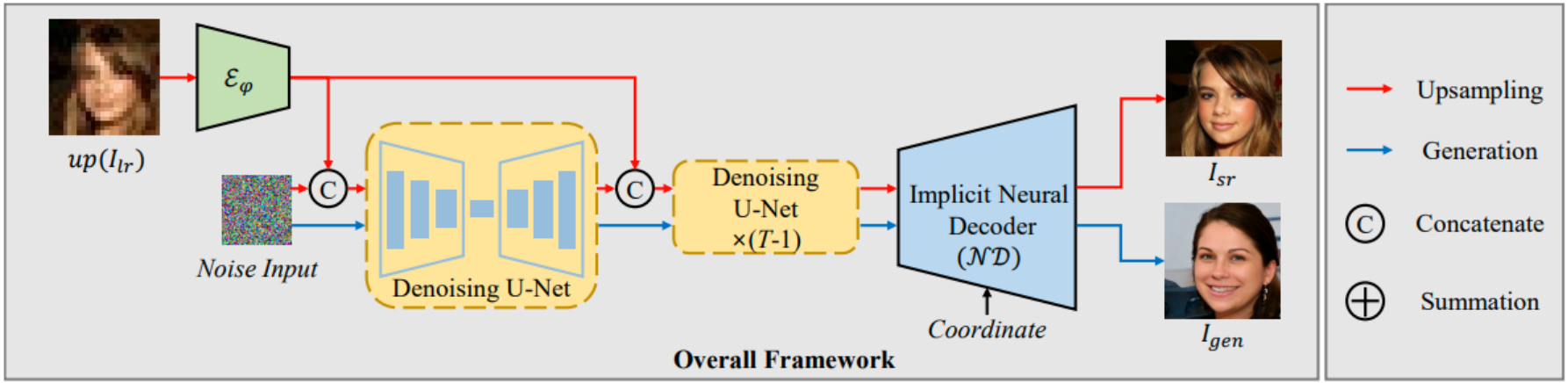
(b) Ours

Proposed method

- More specifically, our arbitrary-scale decoder is designed by the symmetric decoder w/o up-scaling from the pre-trained auto-encoder, and Local Implicit Image Function (LIIF) in series.

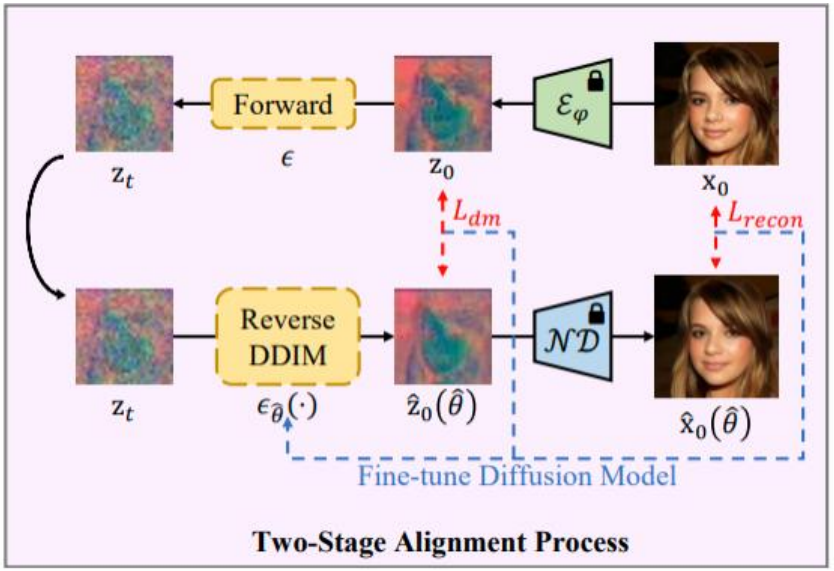


$$I(c) = \mathcal{ND}(\mathrm{z}, c^*) = f_\theta(\mathcal{D}_\varphi(\mathrm{z}), c^*)$$


$up(I_{lr})$
$\mathcal{E}_\varphi$
Noise Input
C
Denoising U-Net
C
Denoising U-Net ×(T-1)
Implicit Neural Decoder ($\mathcal{ND}$)
Coordinate
I_{sr}
I_{gen}
Upsampling
Generation
Concatenate
Summation
Overall Framework

Proposed method

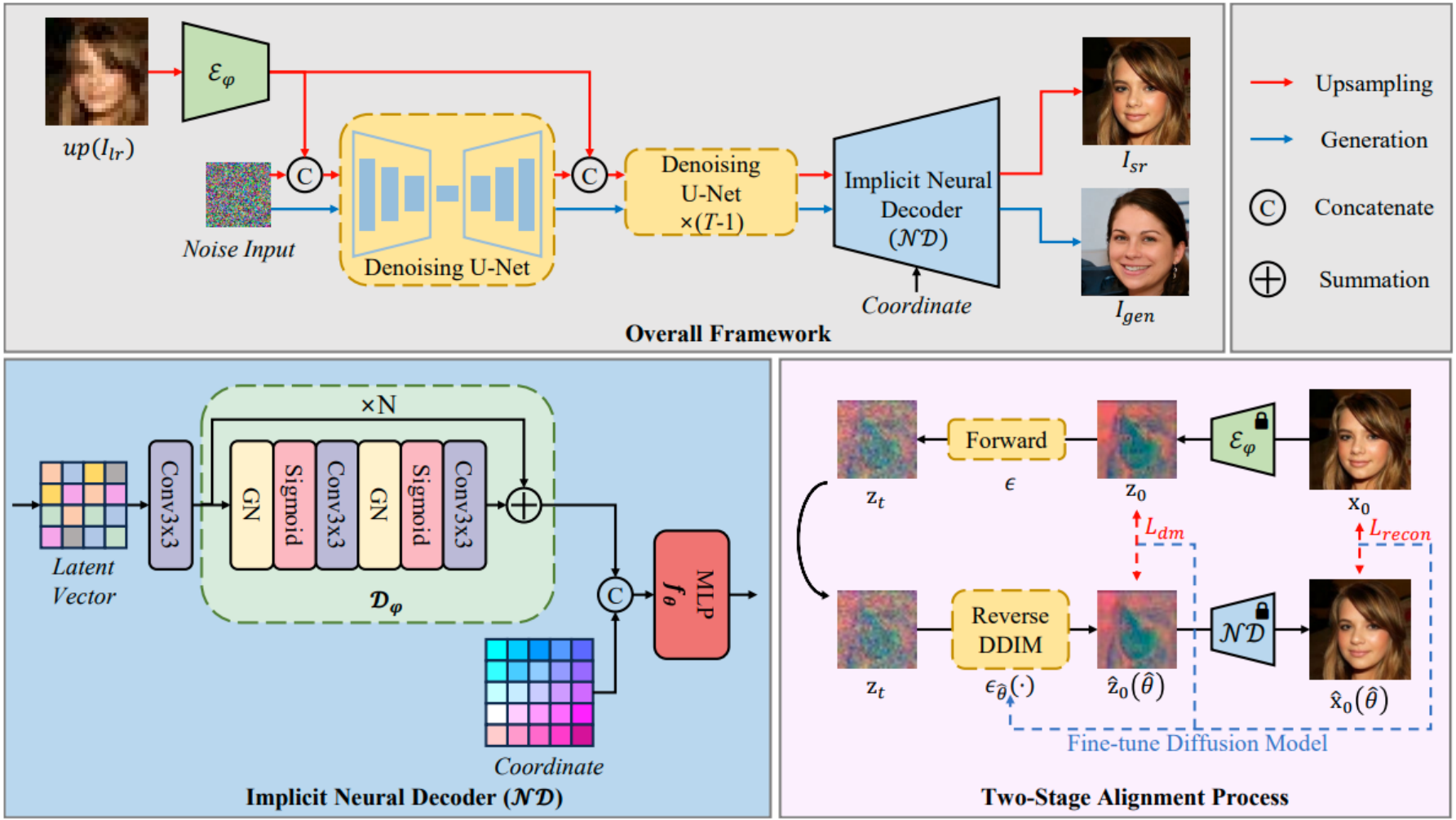
- The latent diffusion process is learnt by the denoising and the alignment losses jointly.
- Errors in output images are backpropagated via the fixed decoder, improving the quality of output images.



$$L_{align} = \lambda_1 L_{dm} + \lambda_2 L_{recon}$$

$$L_{recon} = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \left\| \mathrm{x}_0 - \hat{\mathrm{x}}_0 \right\|_2^2$$

$$\begin{aligned} L_{dm} &= \left\| \epsilon - \epsilon_\theta \left(\mathrm{z}_t \right) \right\|_2^2 \\ &= \left\| \frac{1}{\sqrt{1-\bar{\alpha}_t}} \left(\mathrm{z}_t - \sqrt{\bar{\alpha}_t} \mathrm{z}_0 \right) - \frac{1}{\sqrt{1-\bar{\alpha}_t}} \left(\mathrm{z}_t - \sqrt{\bar{\alpha}_t} \hat{\mathrm{z}}_0 \right) \right\|_2^2 \\ &= \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \left\| \mathrm{z}_0 - \hat{\mathrm{z}}_0 \right\|_2^2 \end{aligned}$$


$up(I_{lr})$
$\mathcal{E}_\varphi$
C
Noise Input
Denoising U-Net
Denoising U-Net ×(T-1)
Implicit Neural Decoder ($\mathcal{ND}$)
Coordinate
I_{sr}
I_{gen}
Overall Framework
Upsampling
Generation
Concatenate
Summation
×N
Latent Vector
Conv3x3
GN
Sigmoid
Conv3x3
GN
Sigmoid
Conv3x3
$\mathcal{D}_\varphi$
MLP f_θ
Coordinate
Implicit Neural Decoder ($\mathcal{ND}$)
Forward
ϵ
z_t
z_0
x_0
L_{dm}
L_{recon}
Reverse DDIM
$\epsilon_{\hat{\theta}}(\cdot)$
$\hat{z}_0(\hat{\theta})$
$\hat{x}_0(\hat{\theta})$
Fine-tune Diffusion Model
Two-Stage Alignment Process

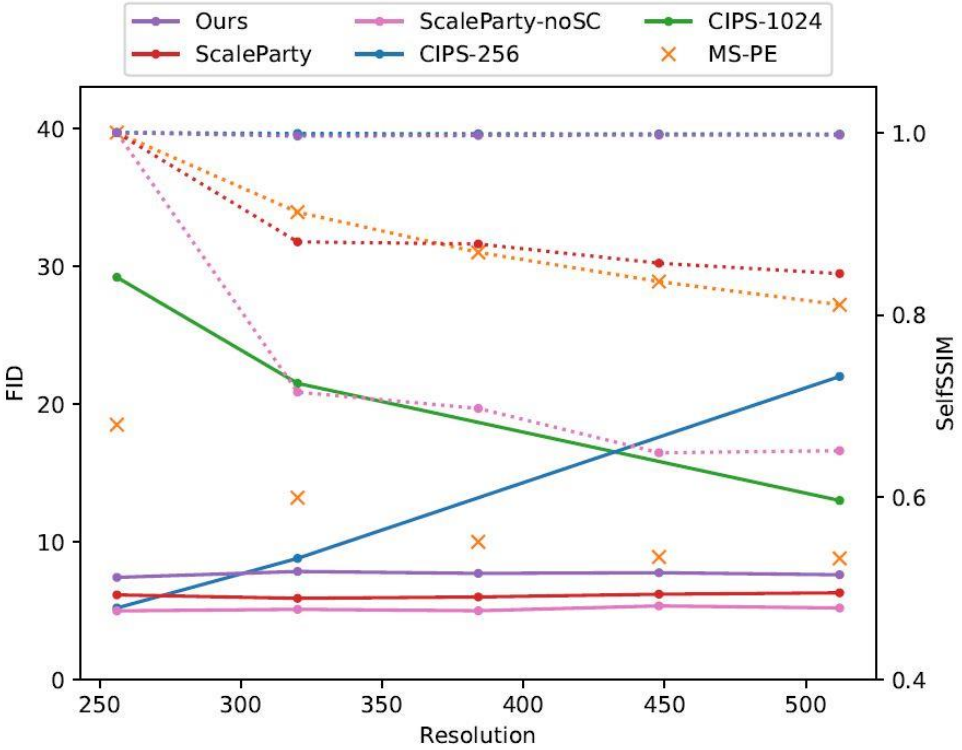
Experiments

- In the extensive experiments using multiple public benchmarks on the two tasks i.e. image super-resolution and novel image generation at arbitrary scales, the proposed method outperforms relevant methods in metrics of image quality, diversity and scale consistency.
- It is significantly better than the relevant prior-art in the inference speed and memory usage.

Datasets used

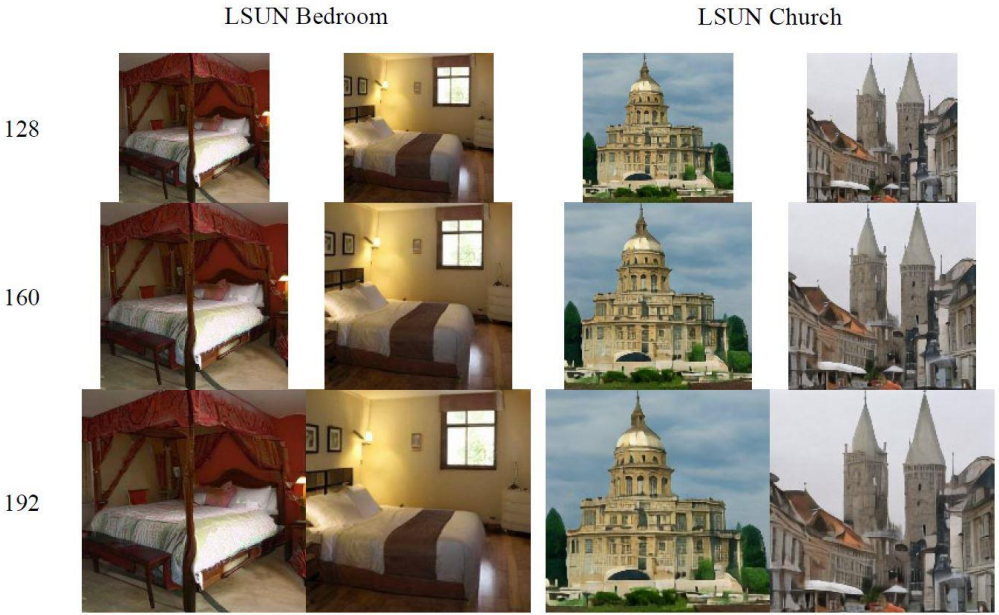
- The Human Face Dataset contains two sub-datasets: Flick-Faces-HQ (FFHQ) [15] and CelebA-HQ [14]. These datasets consist of 70K and 30K different human face images, respectively.
- We used LSUN [36] for general scenes. The LSUN database is divided into various subcategory images, with the smaller size of image is 256x256 pixels.
- To demonstrate the upsampling potential of our model on ultra-high-resolution images, we used the wild datasets DIV2K and Flickr2K.

- We compare our model quantitatively with FID and SelfSSIM scores on the FFHQ datasets.
- The solid lines represent the FID scores of the methods that generate images of arbitrary scale, while the dotted lines indicate the SelfSSIM scores.
- The ‘×’ symbol indicates the method that only generates images of a fixed scale.
- Our model demonstrates competitive performance in both evaluation metrics.



- Comparison of quantitative results on LSUN Bedroom datasets.

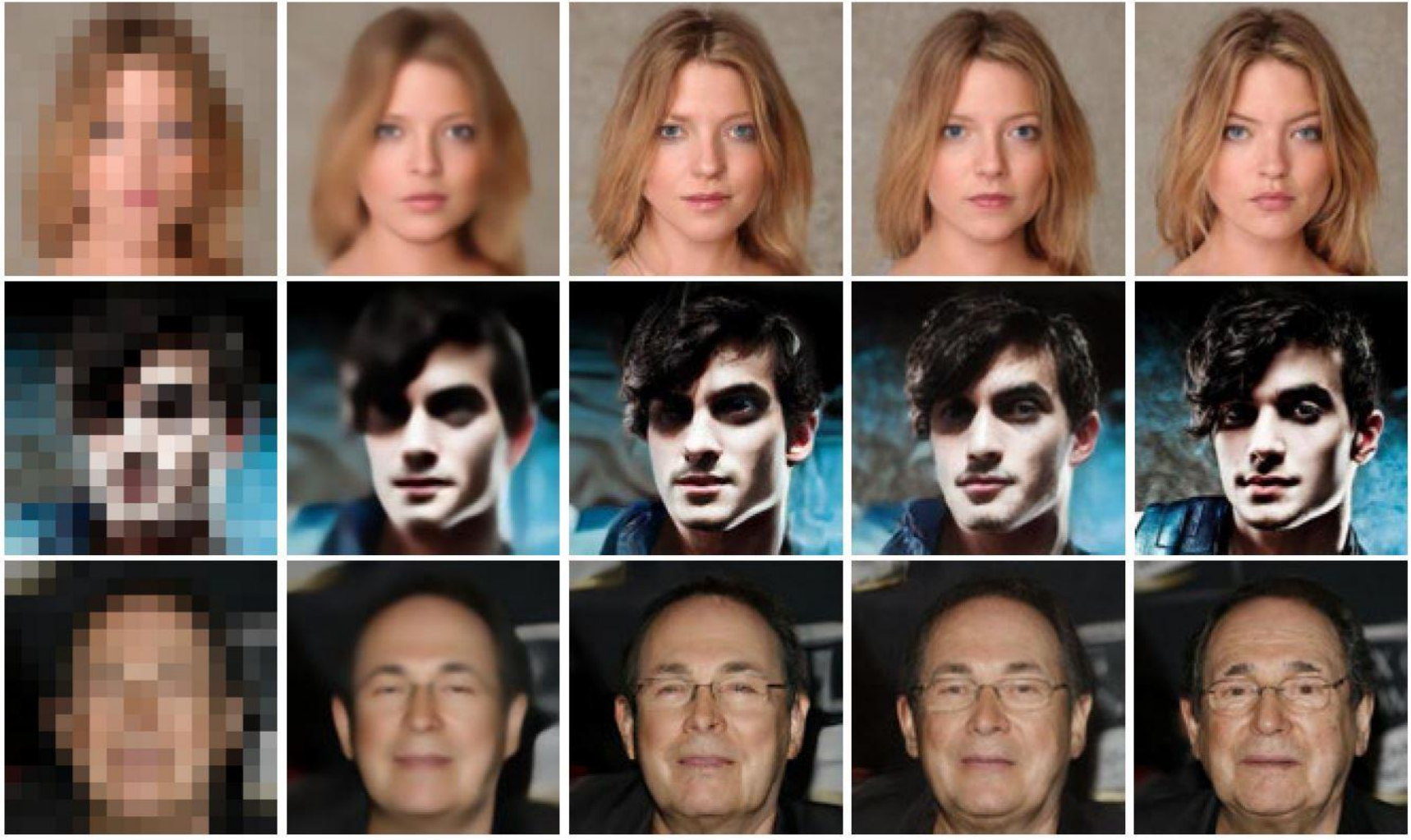
Dataset:	LSUN Bedroom						
Method	Res	FID↓	Prec↑	Rec↑	SelfSSIM (5k)↑		
MSPIE	128	11.39	**66.45**	26.97	1.00	0.10	0.10
	160	16.45	63.84	23.09	0.10	1.00	0.12
	192	12.65	58.10	25.93	0.10	0.12	1.00
Scaleparty	128	10.15	62.50	20.63	1.00	0.94	0.92
	160	9.85	**64.14**	22.02	0.92	1.00	0.95
	192	9.91	**64.77**	21.10	0.89	0.94	1.00
Ours	128	**7.20**	59.69	**38.26**	1.00	**0.98**	**0.99**
	160	**7.43**	58.52	**32.12**	**0.96**	1.00	**0.99**
	192	**7.73**	59.57	**27.98**	**0.95**	**0.97**	1.00



- Quantitative results of arbitrary-scale super-resolution on CelebA-HQ and LSUN Bedroom datasets. For each method, PSNR↑/LPIPS↓ scores are reported.

Dataset:	CelebA-HQ				
Method	5.3×	7×	10×	10.7×	12×
LIIF [7]	**27.52** / 0.1207	**25.09** / 0.1678	22.97 / 0.2246	22.39 / 0.2276	21.81 / 0.2332
SR3 [27]	-	21.15 / 0.1680	20.25 / 0.2856	-	19.48 / 0.3947
IDM [9]	23.34 / 0.0526	23.55 / 0.0736	23.46 / 0.1171	23.30 / 0.1238	23.06 / 0.1800
Ours	24.66 / **0.0455**	24.13 / **0.0690**	**23.64 / 0.1110**	**23.62 / 0.1183**	**23.52 / 0.1427**

Dataset:	Lsun Bedroom	LSUN Tower
Method	16×	
PULSE [22]	12.97 / 0.7131	13.62 / 0.7066
GLEAN [6]	19.44 / 0.3310	18.41 / 0.2850
IDM [9]	**20.33** / 0.3290	19.44 / 0.2549
Ours	20.08 / **0.3269**	**21.24 / 0.1897**


Low-Resolution
LIIF
IDM
Ours
Ground-Truth

- Quantitative comparison of 4× super-resolution using in the wild datasets. D and F refer to the DIV2k and Flickr2k.

Method		Datasets	PSNR↑	SSIM↑
Reg.-based	EDSR	D+F	28.98	0.83
	LIIF	D+F	**29.00**	**0.89**
GAN-based	ESRGAN	D+F	26.22	0.75
	RankSRGAN	D+F	26.55	0.75
Flow-based	SRFlow	D+F	27.09	0.76
Flow+GAN	HCFlow++	D+F	26.61	0.74
Diffusion	IDM	D	27.10	0.77
	IDM	D+F	27.59	0.78
	Ours	D	**27.61**	**0.81**

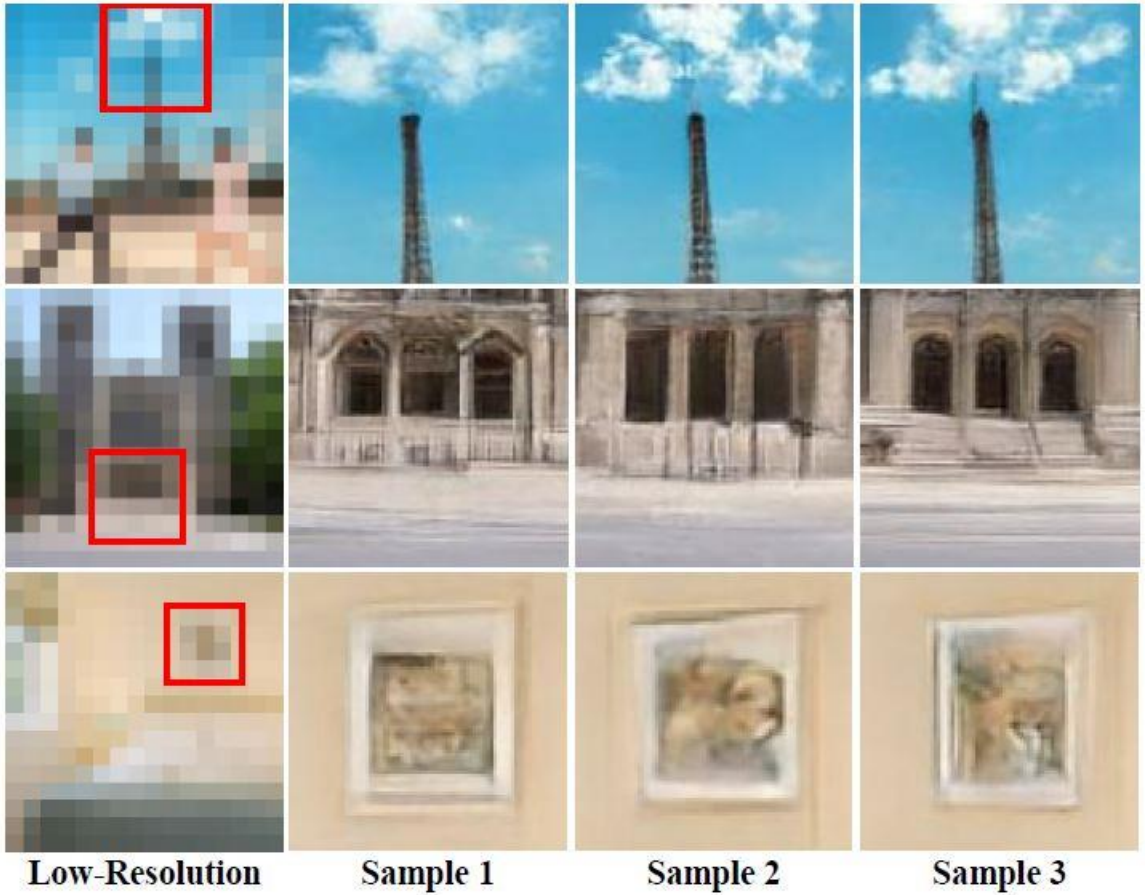
- Comparison (PSNR↑ / LPIPS↓) on the DIV2K at out-of-distribution scales.

Method	8×	12×	17×
LIIF	**23.97** / 0.4790	22.28 / 0.5900	21.23 / 0.6560
Ours	23.82 / **0.4265**	**22.73 / 0.5463**	**21.83 / 0.6225**


Low-Resolution
LIIF
Ours
Ground-Truth

- top: Qualitative results of the proposed method for arbitrary-scale upsampling on face datasets. bottom: Comparison of scale consistency on face dataset.

- Visualization of result diversity in super-resolution tasks



- Comparison of inference Speed in terms of FPS (Frames Per Second). ‘-’ indicates that the model did not work in our environment due to memory overflow.

Method	FPS↑				
	8×	12×	30×	100×	200×
IDM	0.0202	0.0200	6.54e-3	3.98e-4	-
Ours	**0.2568**	**0.2510**	**0.2473**	**0.1833**	**0.0982**

`iEdit`: Localised Text-guided Image Editing with Weak Supervision

CVPR 2024 workshop

Rumeysa Bodur, Erhan Gundogdu, Binod Bhattarai, Tae-Kyun Kim, Michael Donoser, Loris Bazzani

Introduction

- Text-to-image models have seen significant advancements, however, their *controllability* for *localised* edits remains limited.

Problem:

- Current models struggle with preserving image fidelity while performing text-guided edits.

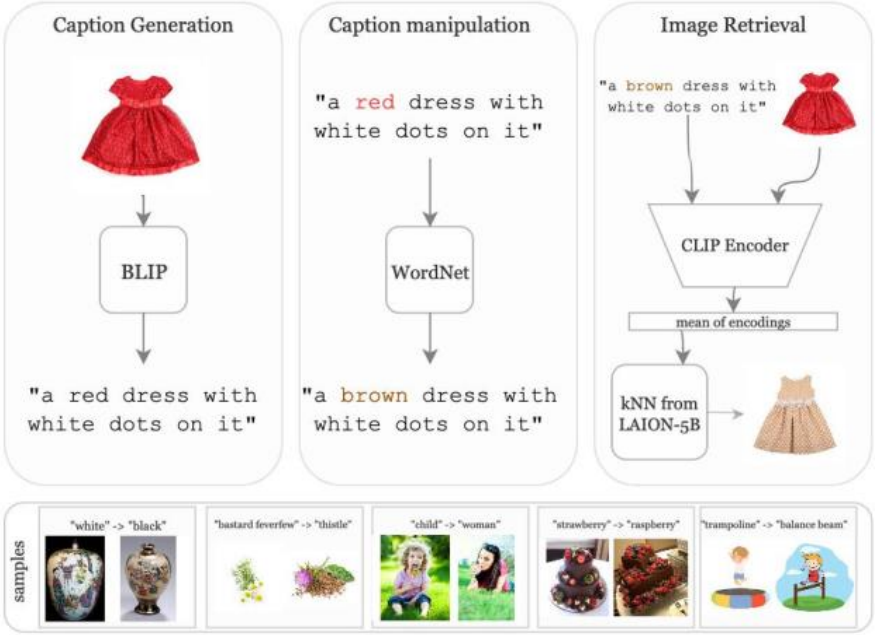
Proposed Solution:

- We present **iEdit**, a novel framework for text-guided image editing using LDMs and weak supervision, creating descriptive target prompts and retrieving pseudo-target images.

Method

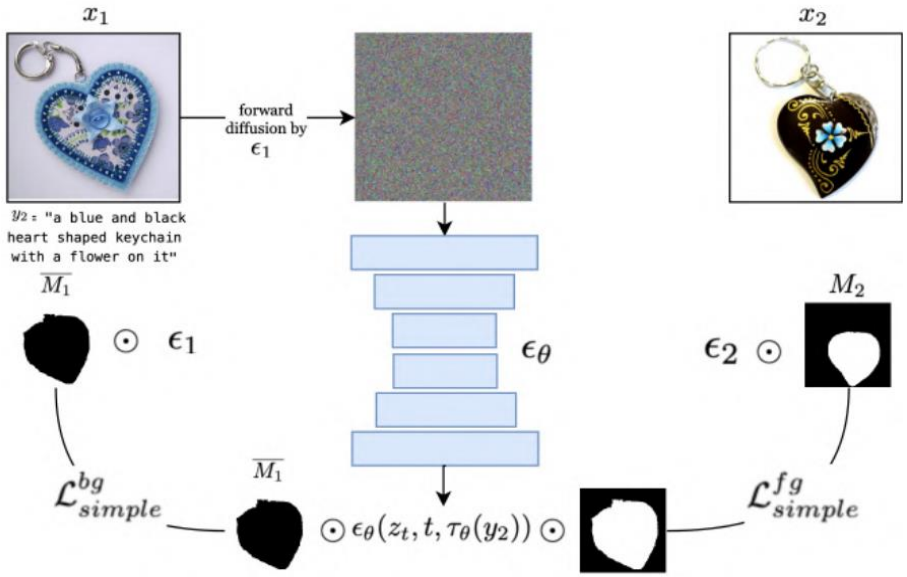
Automatic Dataset Construction: Leveraged LAION-5B to create pseudo-target images and descriptive edit prompts.

- BLIP for generating clean captions
- WordNet for caption manipulation to obtain target captions
- Using CLIP encodings to retrieve pseudo-target images from LAION-5B



iEdit with Location Awareness:
Incorporate masks to encourage to region of interest to align with that of the pseudo-target image, while keeping the rest unchanged.

- CLIPSeg for automatically generating masks from prompt differences
- UNet is finetuned for predicting the noise regarding the target image in the mask area, and the source image in the inversed mask area

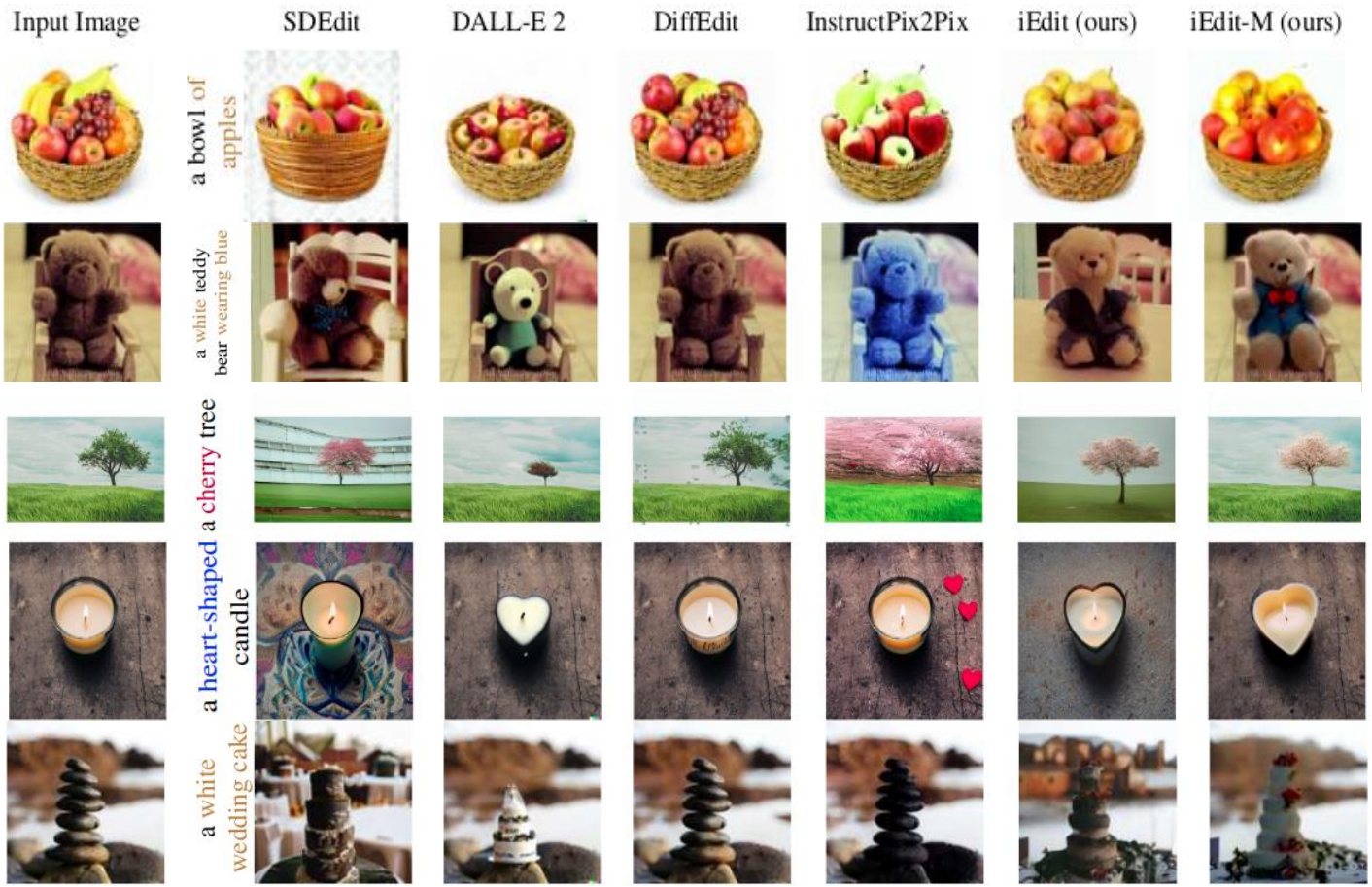


Quantitative Results

Method	Generated Images			
	CLIPScore (%) ↑	**FID↓**	**SSIM-M(%)**	**SSIM-$\overline{M}$(%) ↑**
SDEdit [28]	62.58	171	82.44	50.64
DALL-E 2 [32]	65.44	143	82.45	**94.76**
DiffEdit [7]	60.31	**95**	89.31	78.14
InstructPix2Pix [4]	65.12	108	88.62	76.43
`iEdit` (iP2P dataset)	63.99	106	84.73	76.66
`iEdit`-M (iP2P dataset)	63.04	<u>100</u>	84.87	77.33
`iEdit` (ours)	<u>65.76</u>	158	82.70	52.02
`iEdit`-M (ours)	**66.36**	114	83.08	<u>78.18</u>

Method	Real Images			
	CLIPScore (%) ↑	**FID↓**	**SSIM-M(%)**	**SSIM-$\overline{M}$(%) ↑**
SDEdit [28]	65.84	180	74.36	64.60
DALL-E 2 [32]	65.46	162	74.41	**93.97**
DiffEdit [7]	64.39	**100**	82.06	<u>91.88</u>
InstructPix2Pix [4]	66.91	145	80.59	79.92
`iEdit` (iP2P dataset)	65.62	132	81.25	79.13
`iEdit`-M (iP2P dataset)	65.93	<u>125</u>	80.82	80.18
`iEdit` (ours)	<u>67.02</u>	166	74.59	70.09
`iEdit`-M (ours)	**67.44**	147	74.98	80.44

Qualitative Results



Prompt Augmentation for Self-supervised Text-guided Image Manipulation

CVPR 2024

Rumeysa Bodur, Binod Bhattarai, Tae-Kyun Kim

Introduction

- **Text-guided image editing** enables intuitive and powerful manipulation of images using natural language descriptions.

Problem:

- Existing methods struggle with *coherent image transformation* while *preserving original content* that remains contextually relevant.

- **Existing Methods:**

- Use manual masks (DALL-E 2, Imagen Editor)
- Fine-tuned for specific subjects or domains (DreamBooth)
- Use synthetic targets (InstructPix2Pix)

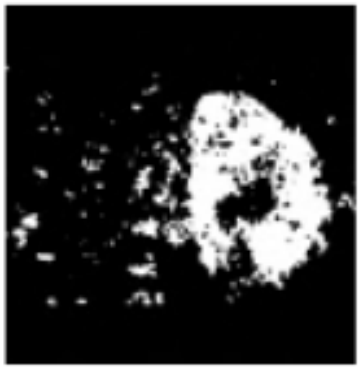
Objective:

Introducing **Prompt Augmentation**:

- **Expand** a single input prompt into several target prompts.
- **Enhance** textual context and enable more precise, localised image editing through the proposed contrastive loss.

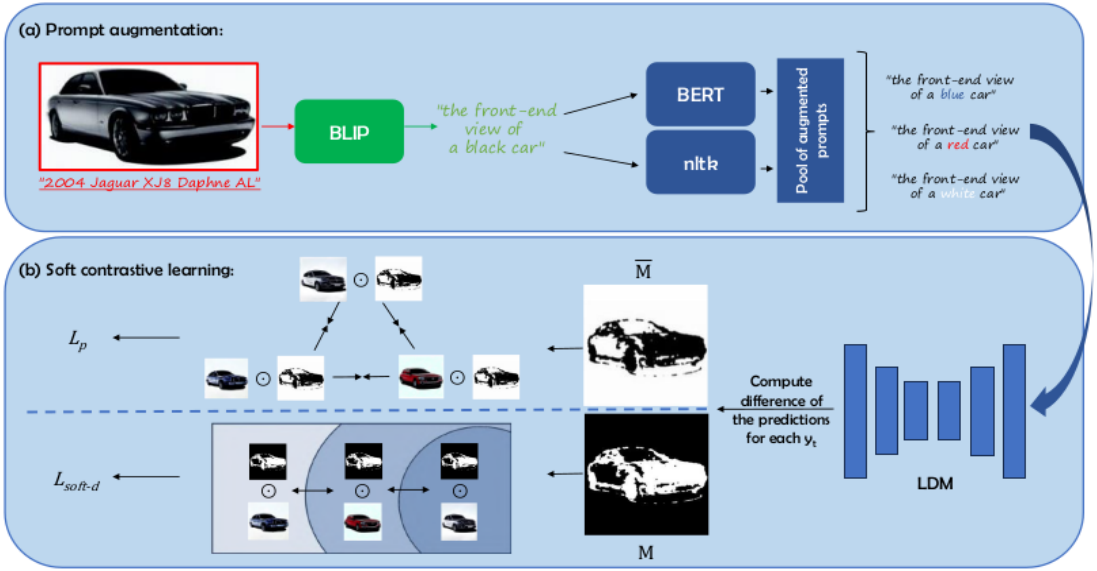
An owl in the snow

An *eagle* in the snow
A *goose* in the snow
An *elf* in the snow



Method

- **Prompt Augmentation:** - BLIP for generating clean captions
- BERT to create multiple target prompts by masking and replacing nouns/adjectives, and NLTK for synonyms, antonyms, co-hyponyms
- **Mask Extraction:** Automatically generate masks by differentiating noise estimates for each augmented prompt.



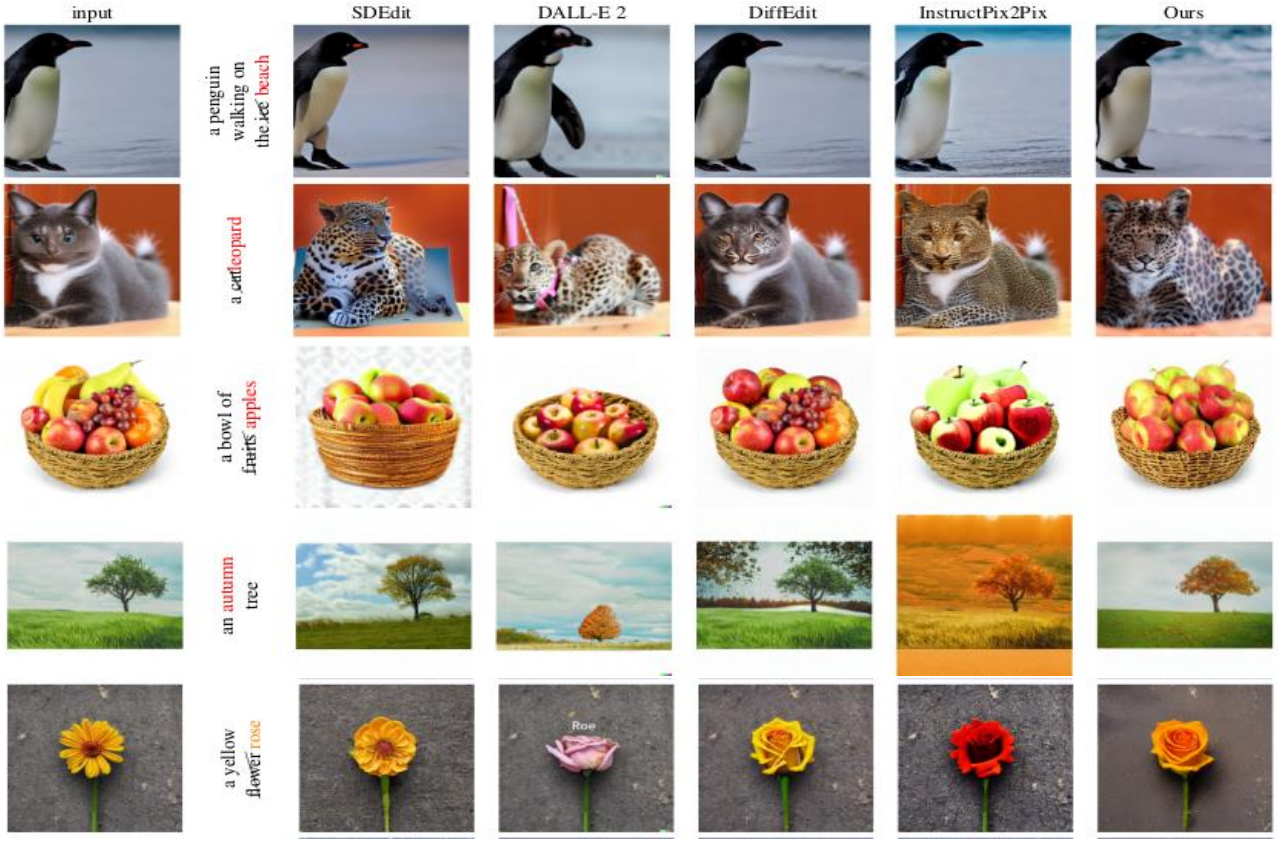
- **Contrastive Loss:** Pushes edited regions while pulling preserved regions towards their original state and each other.
- **Soft Contrastive Loss:** Incorporates prompt similarity for nuanced interaction.

$$L_d^{soft} = \frac{1}{N_p} \sum_{i,j} \left(1 - |(M_i \odot z_i) - (M_j \odot z_j)|_2^2 \cdot \gamma(y_i, y_j)\right)$$

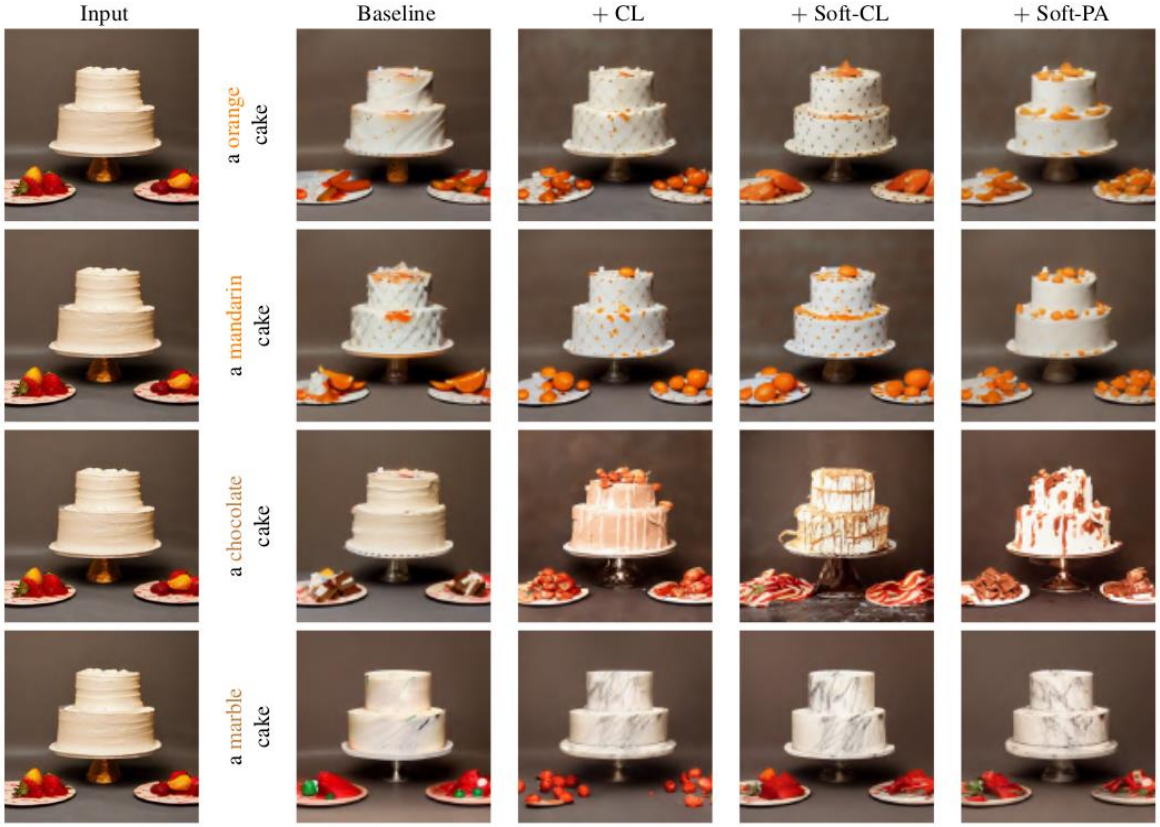
Quantitative Results

Method	CLIPScore (%) ↑	FID ↓	SSIM-$\overline{M}$(%) ↑	CLIP-R-Precision(%) ↑	Human Study
SDEdit [27]	76.76	174	63.07	75.91 ± 1.5	76.3% - 23.7%
DiffEdit [11]	72.77	85	85.14	67.60 ± 1.7	73.84% - 26.16%
DALL-E 2 [31]	78.47	151	96.74	81.83 ± 1.4	68.3% - 31.7%
InstructPix2Pix [8]	77.66	123	81.71	77.45 ± 1.6	62.15% - 37.85%
Our Method	78.19	133	70.39	80.69 ± 1.45	-

Qualitative Results



Ablation Study



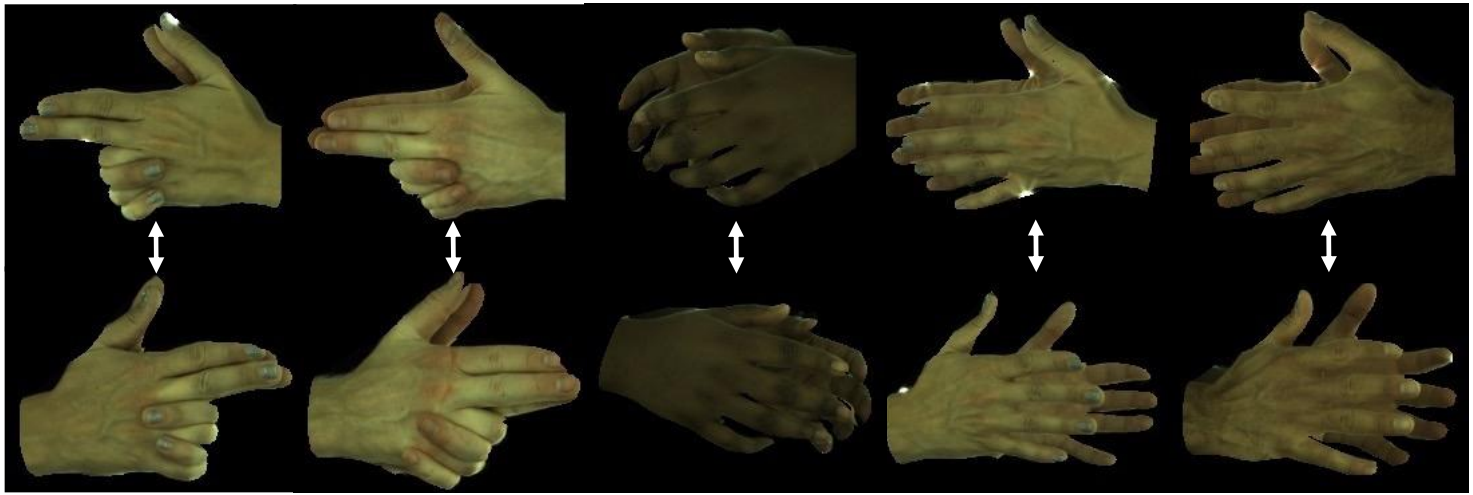
BiTT: Bi-directional Texture Reconstruction of Interacting Two Hands from a Single Image

CVPR 2024

Minje Kim, Tae-Kyun Kim

Motivation & Challenges

- Two hands usually has **similar texture appearances**.
- Hand reconstruction from multi-frames **does not** guarantee **all aspect of the hands** visible and makes it **slow**.
- In a single image, interacting two hands has **more occlusions** than a single hand.



Overview

- We propose **BiTT**, the first method to **reconstruct realistic both hands textures** from a **single image** in **end-to-end** framework**.**
- We utilize **symmetric information** between both hands, and **hand texture parametric model** [1].
- It is **easily trainable** (only a single image!), **free-pose**, **free-view** and **relightable** method.

Input: Single Image of Interacting Two Hands

Two hands reconstruction via
① **Symmetric information of left and right hand**,
② **Hand texture parametric model**

BiTT

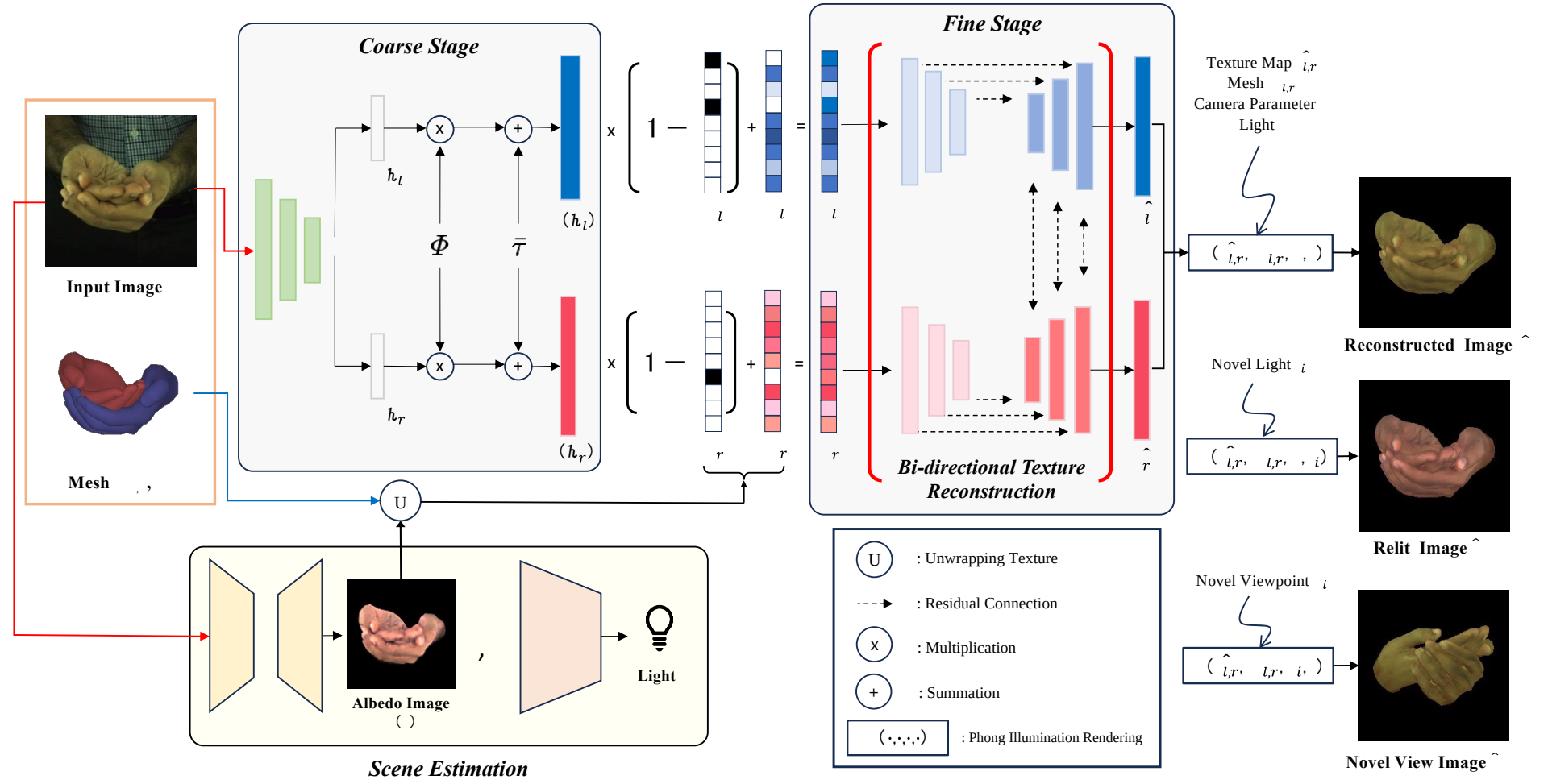
Free View, Free Pose, Relightable Two Hand Reconstruction

Methodology – Overview of the BiTT Model Architecture

- We propose **a novel coarse-to-fine model framework** for reconstructing two hand textures.
  1) ***Coarse Stage*:** We estimate the full texture using **hand texture parametric model** [1], and **albedo image** and **lighting condition** of the scene.
  2) ***Fine Stage*:** We finely **optimize** two hand texture using **symmetric information** of both hands.

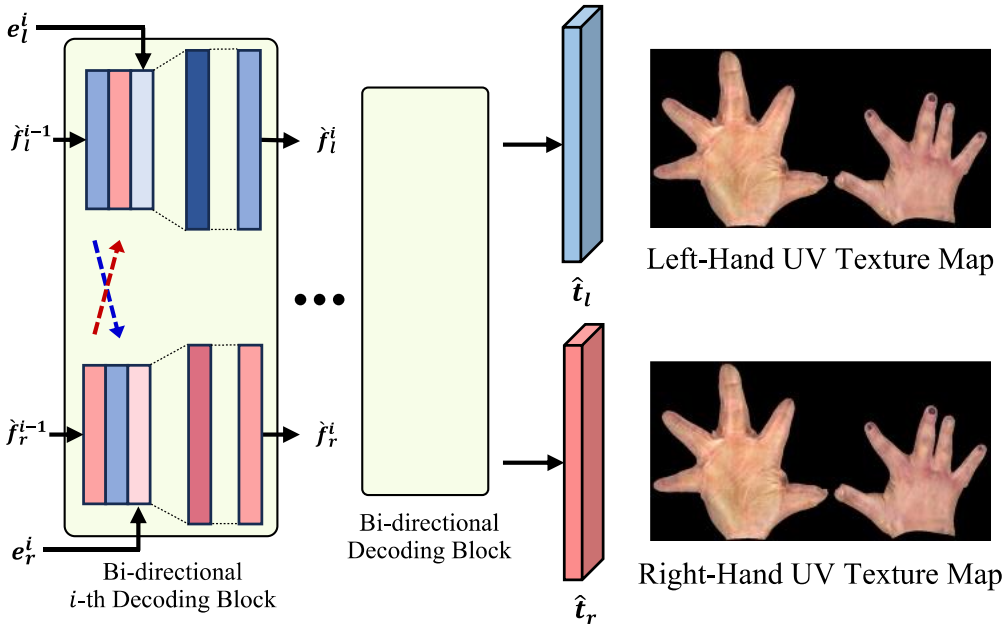
- We also propose **Albedo Consistency Loss** by relighting reconstructed hand.

- Please check details in the paper and supp. materials!


Coarse Stage
Fine Stage
Input Image
Mesh
Albedo Image
Light
Scene Estimation
Bi-directional Texture Reconstruction
Texture Map
Mesh
Camera Parameter
Light
Reconstructed Image
Novel Light
Relit Image
Novel Viewpoint
Novel View Image
U : Unwrapping Texture
: Residual Connection
: Multiplication
: Summation
: Phong Illumination Rendering

Bi-directional Texture Reconstruction – Fine Stage

- Given coarse estimated both hands textures, the **bi-directional decoding block** optimizes the entire hand texture feature using symmetric information.



Quantitative Results

- BiTT achieves SOTA reconstruction results on InterHand2.6M [2].
- BiTT remains robust to geometric misalignments, which is the advantage of the parametric model.

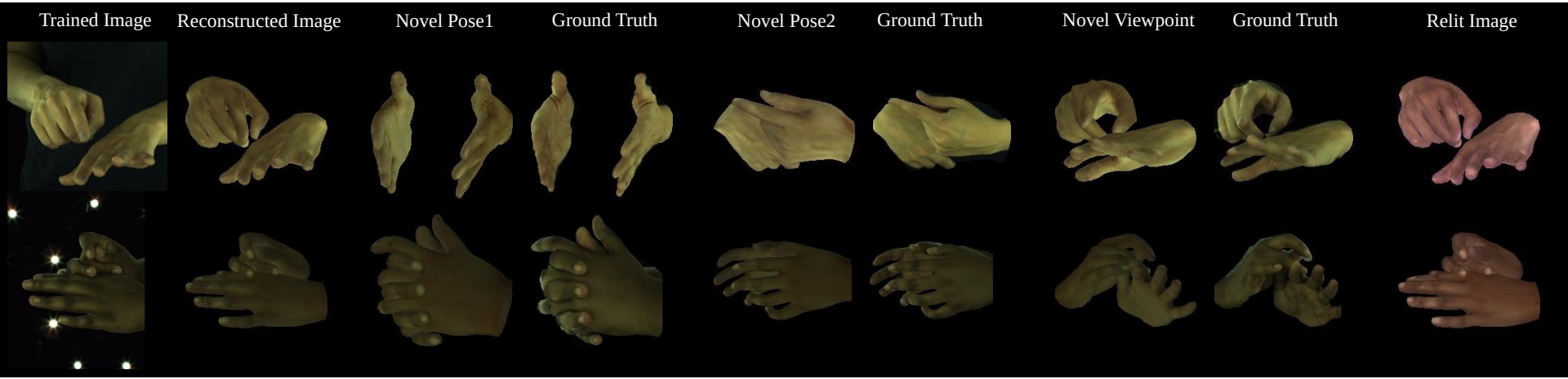
(a) Using GT mesh in all methods.

Evaluation	Method	L1↓	LPIPS↓	PSNR↑	MS-SSIM↑
Appearance Reconstruction	S2Hand [6]	0.0206	0.1340	26.39	0.8570
	HTML [34]	0.0256	0.1292	24.72	0.8152
	HARP [17]	0.0157	**0.0696**	28.11	0.9061
	BiTT(ours)	**0.0101**	0.1019	**30.41**	**0.9349**
Novel Poses	S2Hand	0.0221	0.1343	25.70	0.8507
	HTML	0.0255	0.1291	24.49	0.8153
	HARP	0.0239	0.1266	25.79	0.8546
	BiTT(ours)	**0.0209**	**0.1261**	**26.54**	**0.8564**
Different Views	S2Hand	0.0217	0.1320	25.73	0.8484
	HTML	0.0254	0.1282	24.42	0.8133
	HARP	0.0234	0.1189	25.97	0.8346
	BiTT(ours)	**0.0204**	**0.1092**	**27.79**	**0.8843**

(b) Without using GT mesh in all methods.

Evaluation	Method	L1↓	LPIPS↓	PSNR↑	MS-SSIM↑
Appearance Reconstruction	S2Hand [6]	0.0264	0.1214	25.72	0.8897
	HTML [34]	0.0268	0.1207	24.48	0.8545
	HARP [17]	0.0237	0.1047	25.17	0.8697
	BiTT(ours)	**0.0131**	**0.1044**	**28.40**	**0.9093**
Novel Poses	S2Hand	0.0280	0.1525	23.06	0.8092
	HTML	0.0310	0.1299	23.46	0.8281
	HARP	0.0256	0.1410	24.32	0.8419
	BiTT(ours)	**0.0223**	**0.1228**	**25.12**	**0.8423**
Different Views	S2Hand	0.0244	0.1512	24.22	0.8335
	HTML	0.0291	0.1297	24.22	0.8375
	HARP	0.0251	0.1367	24.49	0.8507
	BiTT(ours)	**0.0210**	**0.1273**	**26.34**	**0.8674**

Qualitative Results



Qualitative Results

Input Image

Reconstructed Image

Relit Image

Novel Pose

Input Image

Reconstructed Image

Relit Image

Novel Pose

Further study

- As BiTT is based on **mesh rendering**, we can apply **self-shadow rendering** using traditional concepts in computer graphics.

Input Image

w/o Shadow

w/ Shadow

References

[1] N. Qian, and et al. HTML: A Para- metric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *ECCV*, 2020.

[2] G. Moon, and et al. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020.

[3] M. Li, and et al. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022.

[4] K. Karunratanakul, and et al. Harp: Personalized hand reconstruction from a monocular rgb video. In *CVPR*, 2023.

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Dense Hand-Object(HO) GraspNet with Full Grasping Taxonomy and Dynamics

ECCV 2024

Woojin Cho[1], Jihyun Lee[1], Minjae Yi[1], Minje Kim[1], Taeyun Woo[1], Donghwan Kim[1], Taewook Ha[1], Hyokeun Lee[3], Je-Hwan Ryu[4], Woontack Woo[1], Tae-Kyun Kim[1,2]

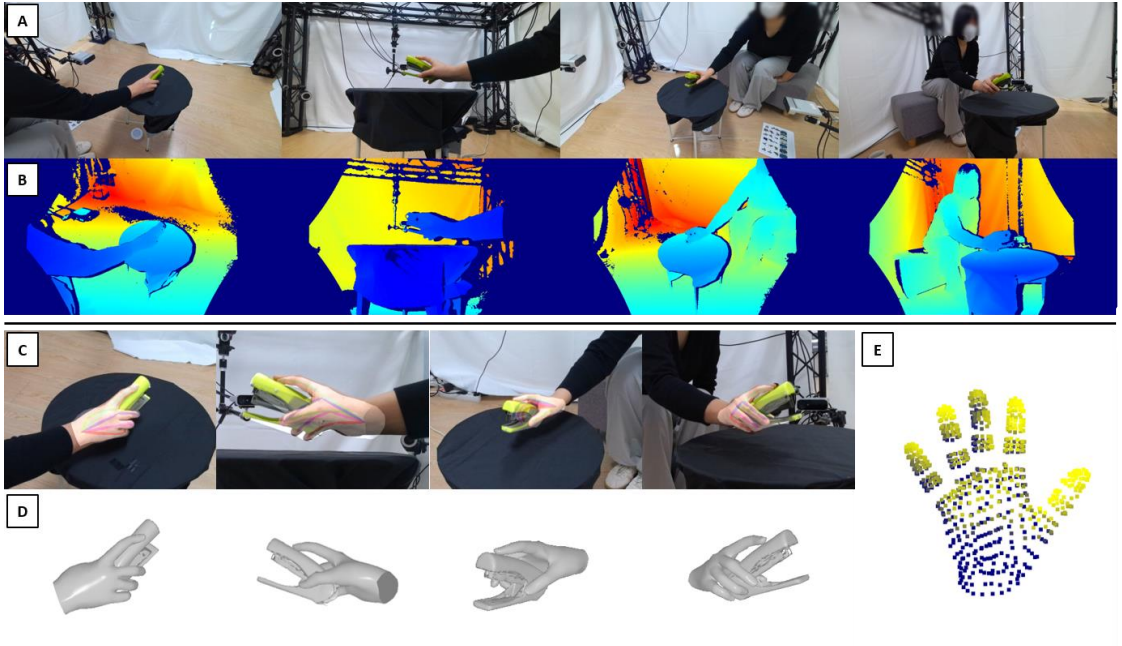
Dataset page

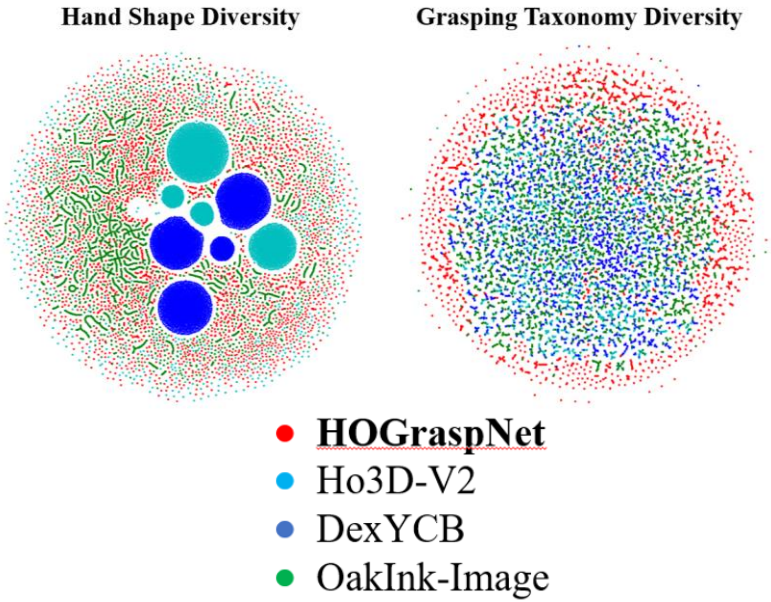
1 KAIST
2 IMPERIAL
3 KWANGWOON UNIVERSITY
4 SUPERMIND

Introduction

- **HOGraspNet**, an extensive multi-view RGBD training dataset for hand, object, and their interaction with grasp annotations.
- Existing 3D hand-object interaction datasets are limited either in data size, interaction variety, or annotation quality.
- Focused on **covering all grasp taxonomies**, including grasp labels and a wide range of intraclass variations.

Dataset Overview





Summary

- **1,489,112 RGB-D** frames with **4 viewpoints**
- 3 Grasps per **30 Objects**, a total **28 grasp classes**
- **99 participants** aged **10 to 74**

Annotation types

- 3D hand pose for 21 joints
- **Grasp class**, 6D object pose, and contact map
- **MANO**[1], **HALO**[2] mesh annotation

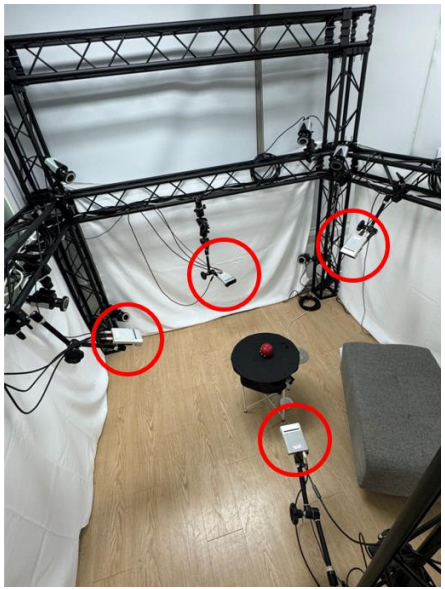
Split Protocols

Generated five distinct train/test splits based on key components.

- S0 (default)
- S1 (unseen subjects)
- S2 (unseen views)
- S3 (unseen objects)
- S4 (unseen taxonomy)

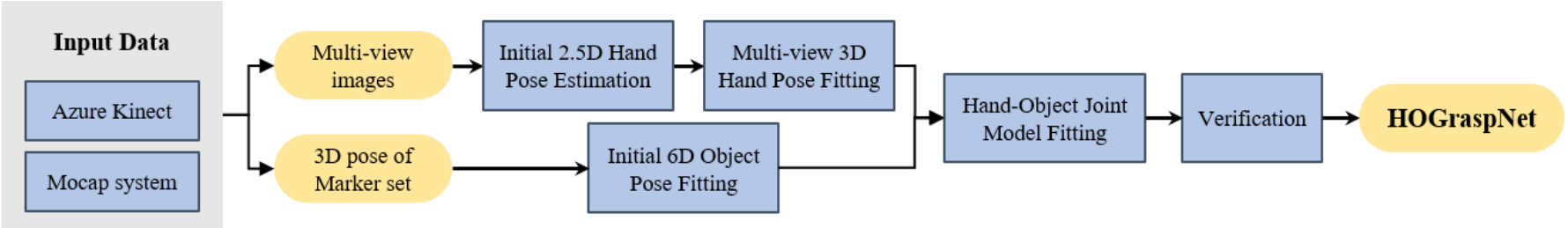
Hardware Setup

- 4 synchronized RGB-D cameras (Azure Kinect)
- 3mm optical markers on each object with 8 IR cameras



Annotation

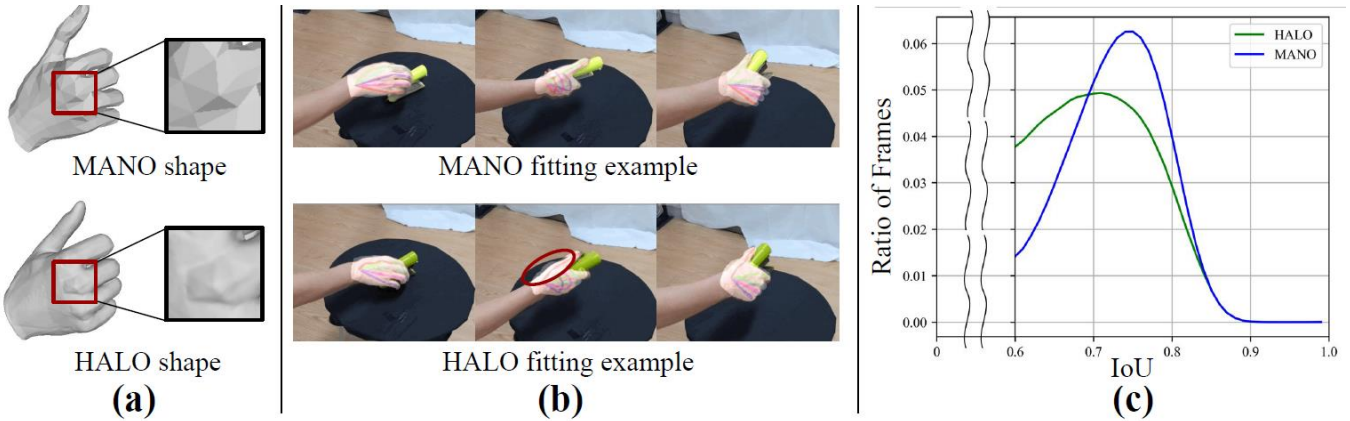
- We designed automatic annotation and verification pipeline inspired by prior studies [**3,4**].



- Overall loss function for the MANO pose $\theta \in \mathbb{R}^{48}$ and shape $\beta \in \mathbb{R}^{10}$ and the object 6D pose $\phi \in \mathbb{R}^{6}$ as follows:

$$\mathcal{L} = \lambda_h^{2D}\mathcal{L}_h^{2D} + \lambda_o^{3D}\mathcal{L}_o^{3D} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{phy}\mathcal{L}_{phy}$$

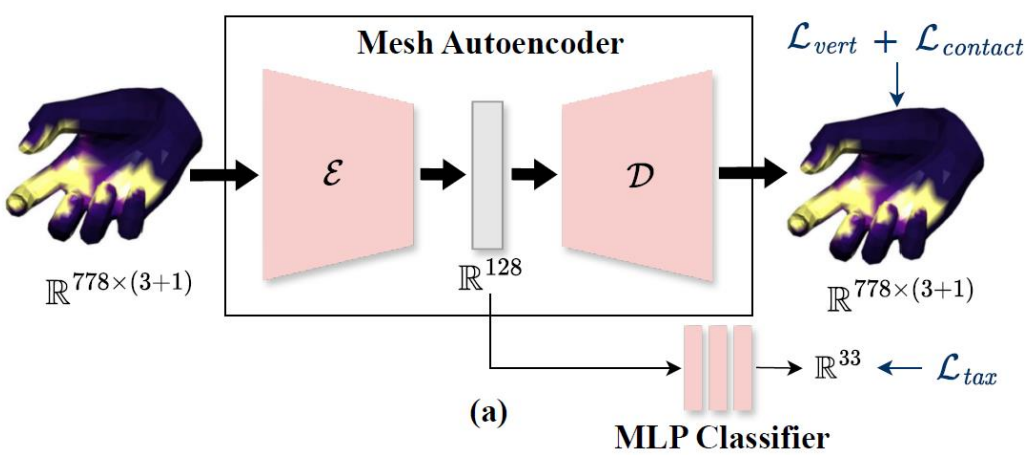
- Conduct both Intersections over Union-based automatic verification and manual verification through crowdsourcing.
- For hand shape annotation, we additionally provide the hand implicit surface based on HALO[**2**], which parameterizes an articulated occupancy field.

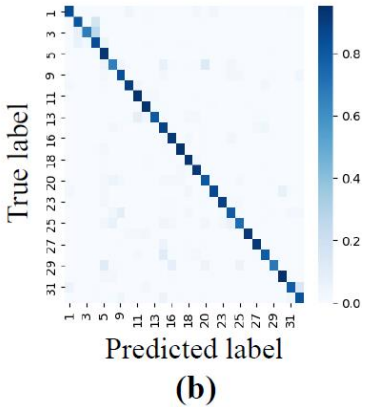


Experiments

Grasp Classification

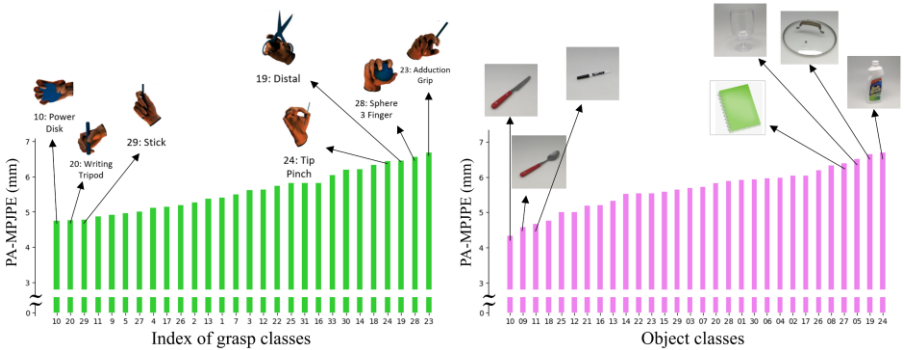
- We evaluate grasp classification performance using S0 split with convolutional mesh autoencoder CoMA[**5**].





Hand-Object Pose Estimation

- Present the benchmarking results on hand-object pose estimation on HOGraspNet S0 split with HFL-Net[6] as a baseline.



	ADD-0.1D(↑)		ADD-0.1D(↑)
1: cracker_box	88.79	16: golf_ball	46.27
2: potted_meat_can	59.47	17: credit_card	34.06
3: banana	58.21	18: dice	2.44
4: apple	75.74	19: disk_lid	99.29
5: wine_glass	97.13	20: smartphone	52.22
6: bowl	94.64	21: mouse	41.40
7: mug	72.04	22: tape	62.34
8: plate	99.42	23: master_chef_can	88.75
9: spoon	50.60	24: scrub_cleanser_bottle	89.80
10: knife	38.17	25: large_marker	34.59
11: small_marker	28.29	26: stapler	61.40
12: spatula	67.16	27: note	88.70
13: flat_screwdriver	63.52	28: scissors	54.34
14: hammer	82.99	29: foldable_phone	25.02
15: baseball	73.72	30: cardboard_box	77.80
Avg	63.61		

- Cross-validation results on hand pose estimation with same setup on HO3D[4] and DexYCB[3].

Train Set	Test Set	MPJPE (mm)	PA-MPJPE (mm)
HO3D [4]	DexYCB [3]	57.31	10.31
HOGraspNet	DexYCB [3]	**42.65**	**9.36**

References

[1] Romero et al. Embodied hands: Modeling and capturing hands and bodies together. In: ACM TOG (2017)

[2] Karunratanakul et al. A skeleton-driven neural occupancy representation for articulated hands. In: 3DV (2021)

[3] Chao et al. Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR (2021)

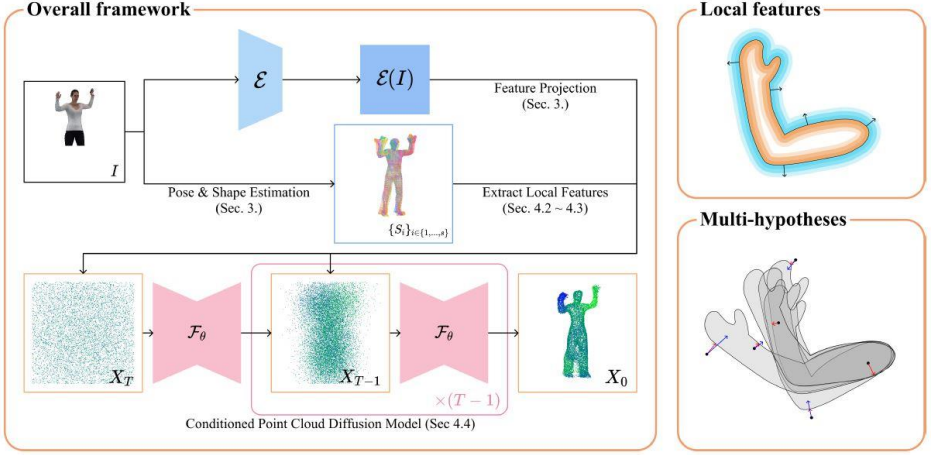
[4] Hampali et al. Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)

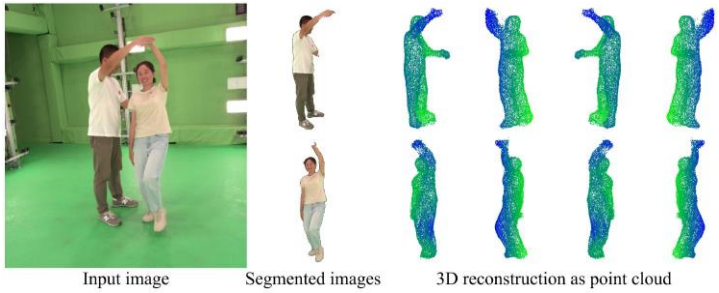
[5] Ranjan et al. Generating 3d faces using convolutional mesh autoencoders. In: ECCV (2018)

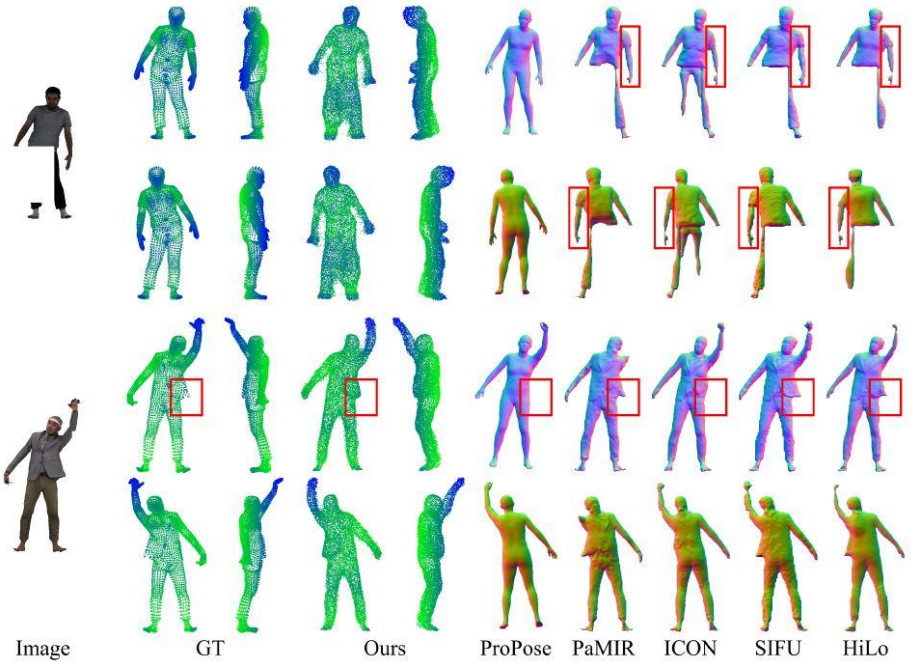
[6] Lin et al. Harmonious feature learning for interactive hand-object pose estimation. In: CVPR (2023)

[5

D. Kim, T-K. Kim, Multi-hypotheses Conditioned Point Cloud Diffusion for 3D Human Reconstruction from Occluded Images, NeurIPS 2024.







MGHanD: Multi-modal Guidance for authentic Hand Diffusion (under review)