SRHand

Super-Resolving Hand Images and 3D Shapes via View/Pose-aware Neural Image Representations and Explicit 3D Meshes

Minje Kim and Tae-Kyun Kim

NIPS 2025

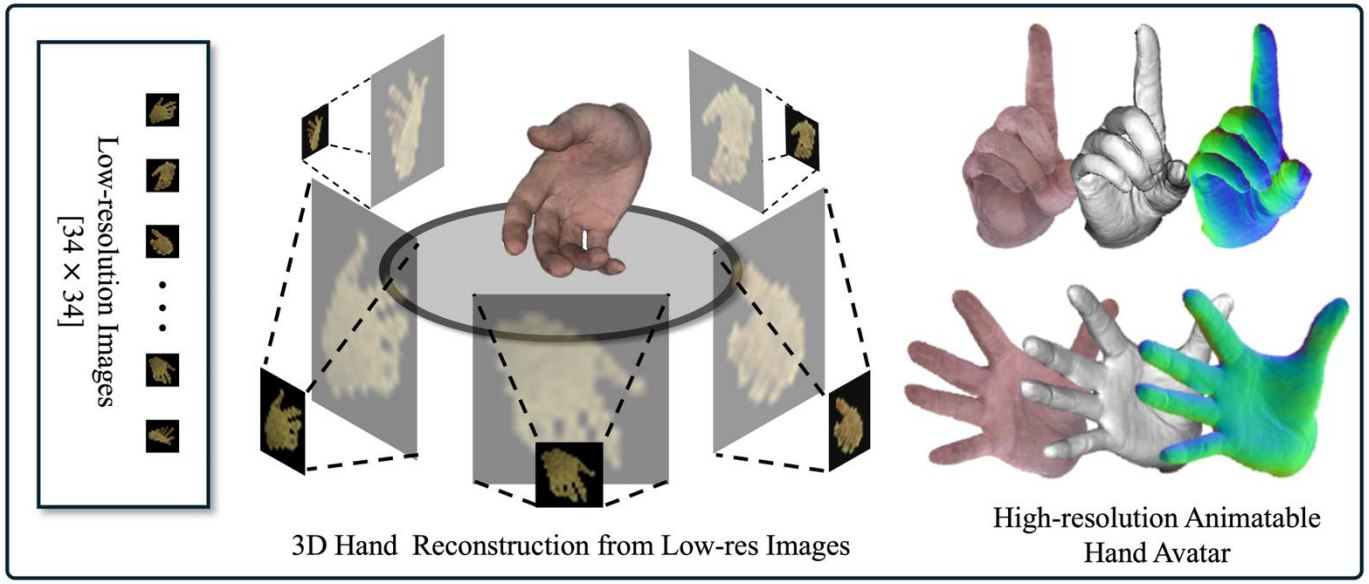
KAIST

NEURAL INFORMATION PROCESSING SYSTEMS

KCVL

Introduction

Full-body Captured Multi-view Images [334 × 512]



Motivation

1. Hands usually occupies less than 1.5% of the captured full body image.

2. For detail reconstruction, high resoultion images are essential

3. In multi-view setup, the resolution of the captured hand region varies (16x16 ~ 48 x48) with human poses and camera positions, making the problem more challenging

KCVL

Contribution

- We introduce **SRHand**, a novel framework that **super-resolves both 2D images and 3D shapes** by integrating implicit image representations with explicit 3D meshes

- We propose **a geometric-aware implicit image function (GIIF),** that conditions implicit neural representations **on normal maps** while leveraging **adversarial learning** to enhance texture fidelity.

- We **jointly fine-tune** the hand **SR module** with the **3D reconstruction process** to enforce multi-view/pose consistency.

KCVL

Related Work

- **Implicit Image Representation**
  - Coordinate-based neural functions, enabling resolution-agnostic inference

$$LIIF(I_{lr}) = f_\theta(z, [x, c]), \text{where } z = E_\varphi(I_{lr})$$

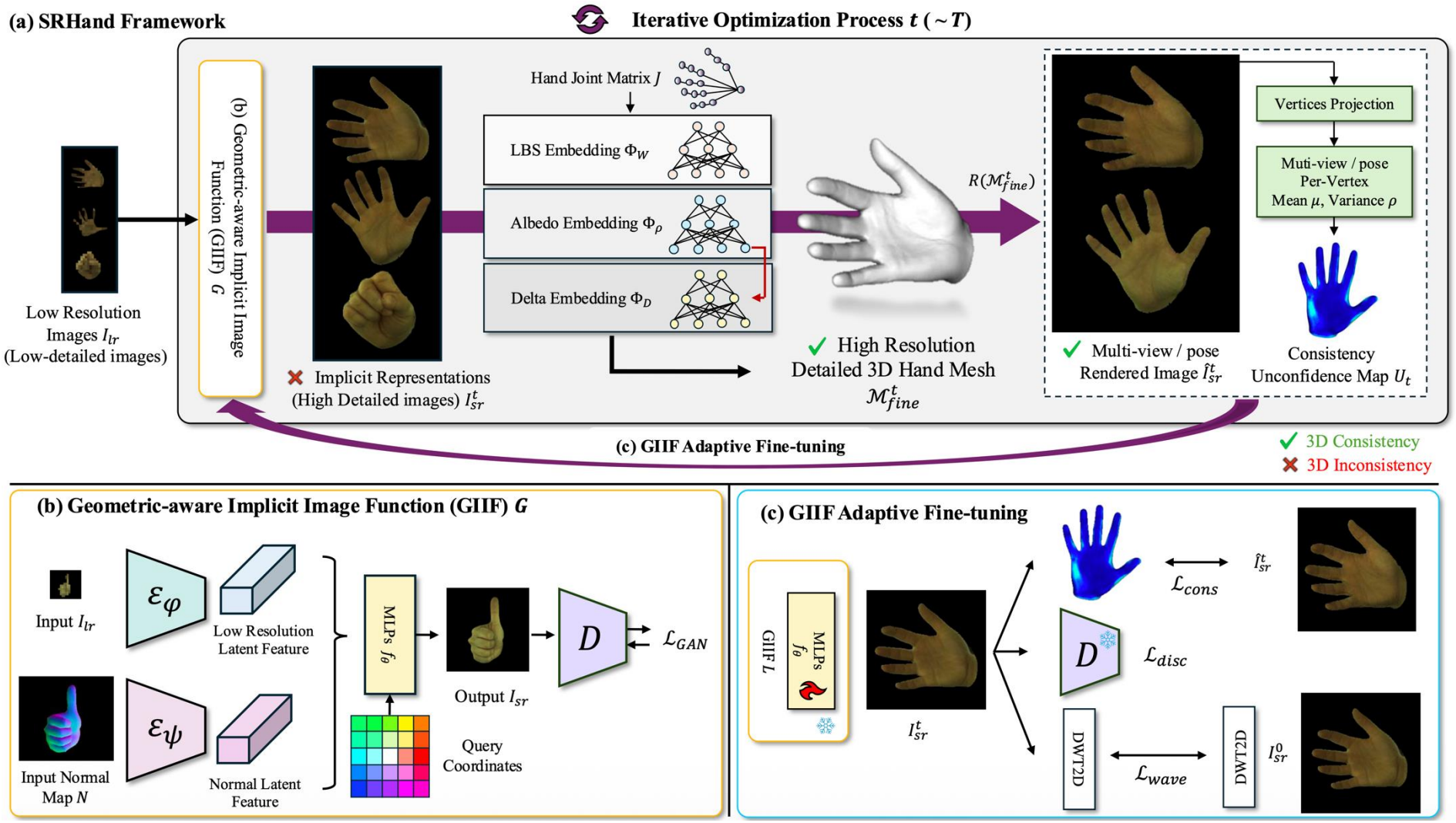
- **Explicit 3D Mesh**
  - Our mesh representation is based on [2], following below formulation.

$$R(\pi^i | \mathcal{M}_{fine}, J) = \Phi_\rho(J) \cdot SH(Y, \mathcal{N}'), \text{where } \mathcal{M}_{fine} = \Omega(\bar{\mathcal{M}}' + \Phi_D(J), \Phi_W(J), \theta, \beta)$$

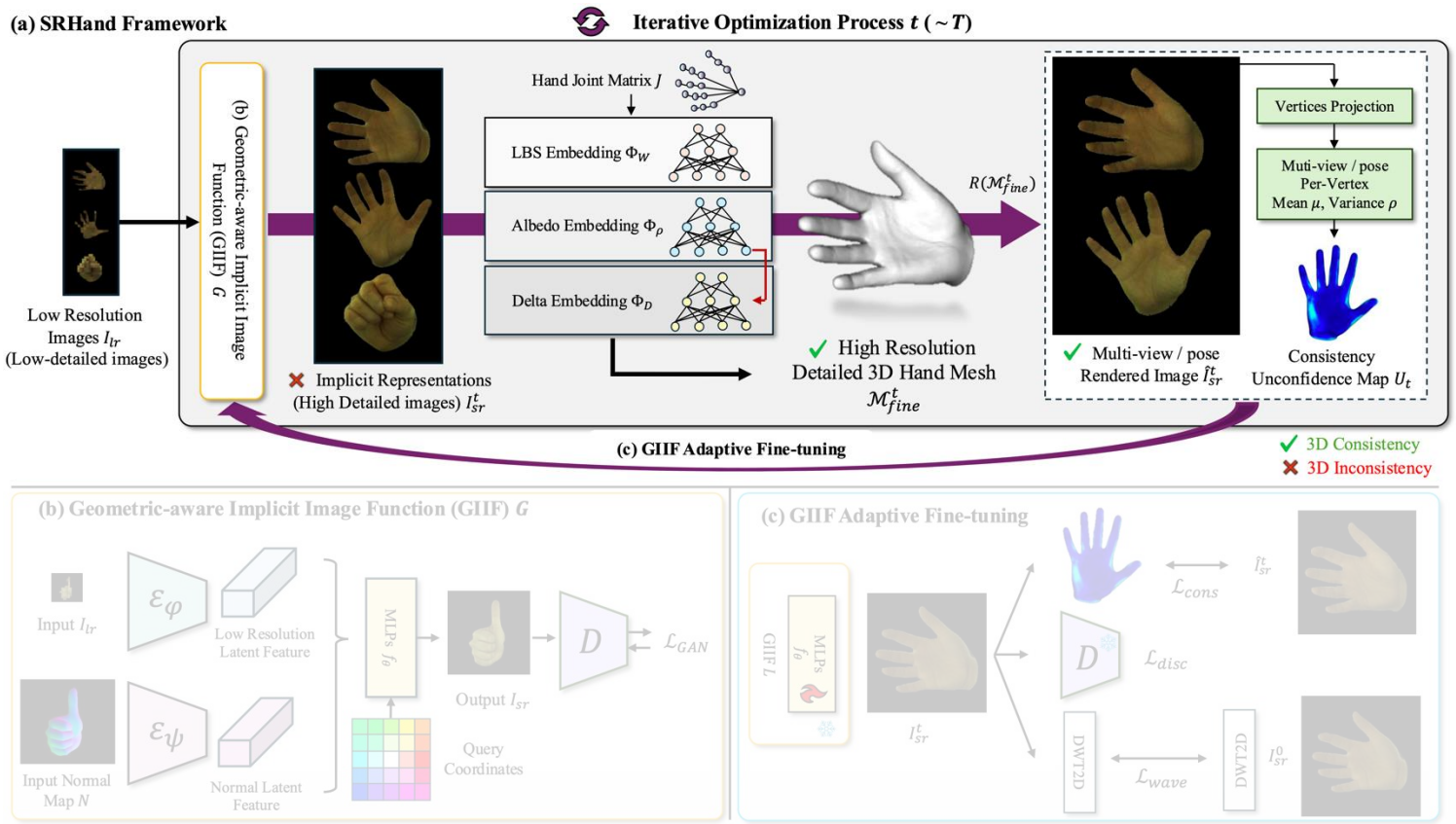
[1] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In CVPR, 2021.
[2] Qijun Gan, Zijie Zhou, and Jianke Zhu. Xhand: Real-time expressive hand avatar. arXiv:2407.21002, 2024.

KCVL

Methodology



Methodology



SRHand Pipeline

0) LR Images

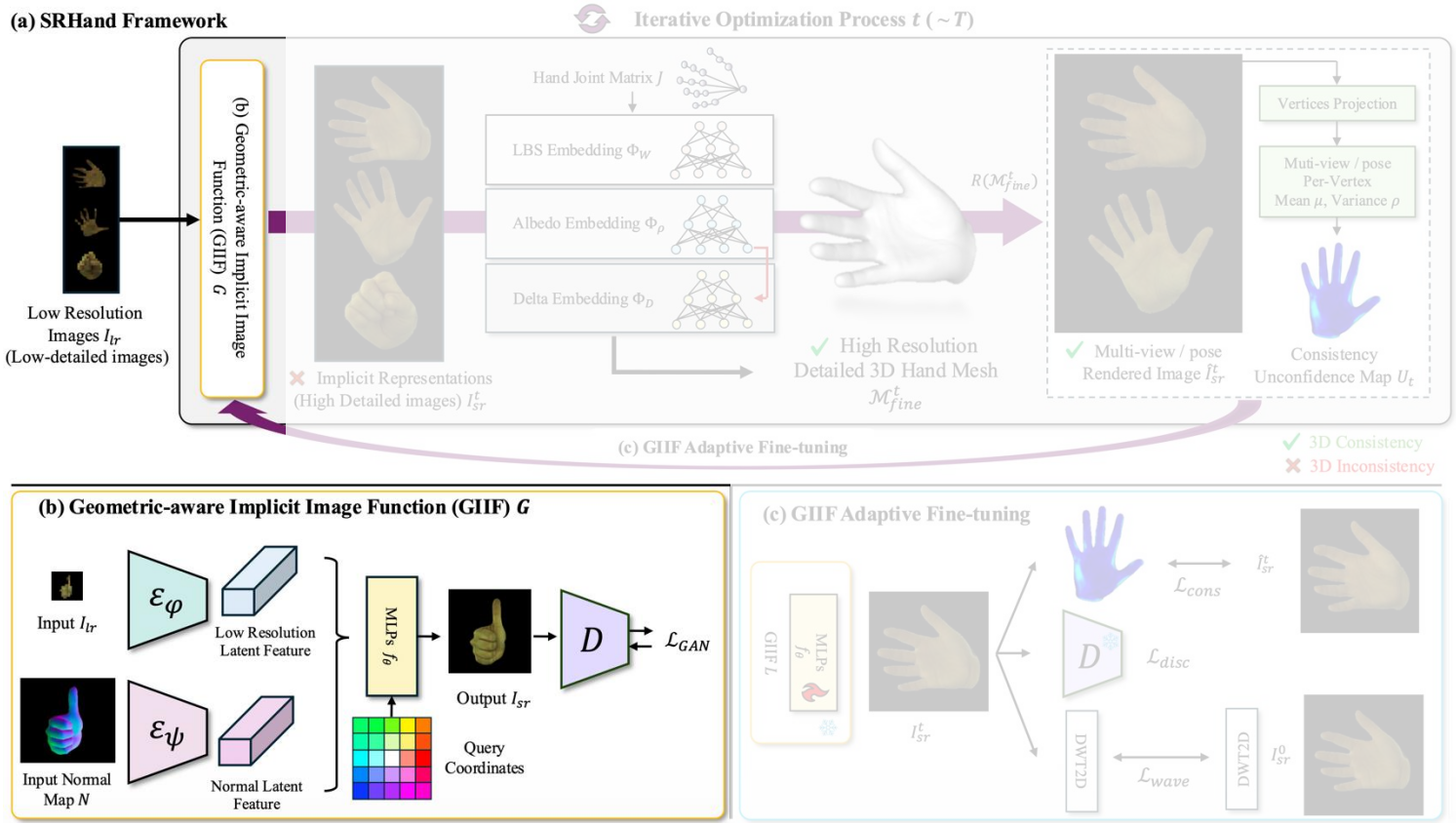
1) Forward SR Module (GIIF)

2) 3D Hand Reconstruction

3) GIIF Adaptive Fine-tuning

4) Re-start 1) process

Methodology



GIIF (Geometric –aware IIF)

- Why Implicit Image Function?
  - Input resolution can vary!
  - IIF can cover arbitrary scale

- LR latent feature + Normal latent feature

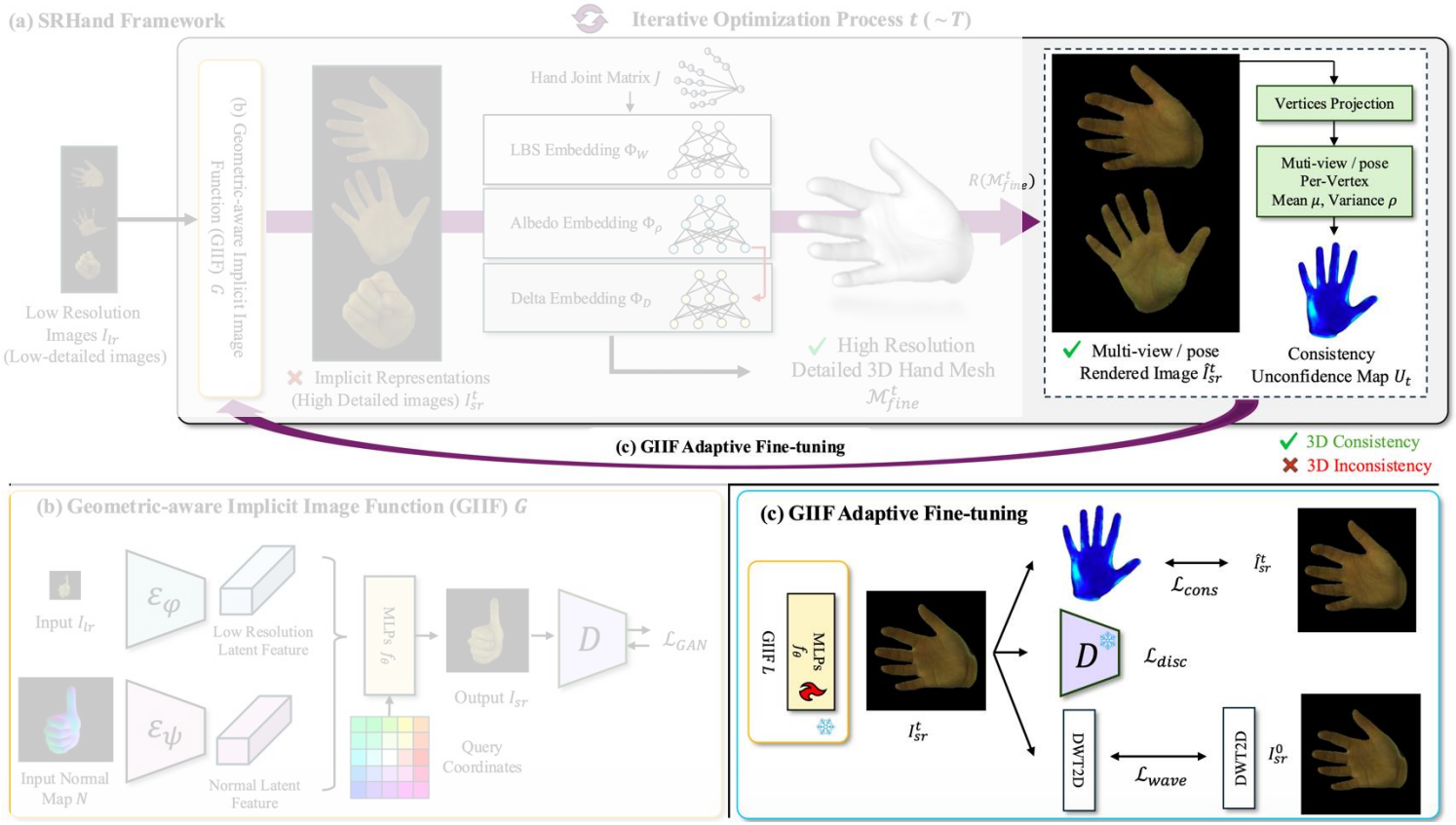
$$\mathbf{f}_{fused} = \mathcal{E}_\varphi(I_{lr}). \oplus \mathcal{E}_\psi(N)$$

- Decoding through cell-coordinate base

$$I_{sr} = G(I_{lr}, N) = \mathcal{F}_\theta\left(\mathbf{f}_{fused}, [x, c]\right)$$

KCVL

Methodology



GIIF Adaptive Fine-tuning

- Consistency Loss

$$\mathcal{L}_{cons} = \mathcal{L}_1((R(\mathcal{M}_{fine}) \cdot U^t(Q), I_{sr} \cdot U^t(Q))$$

- Frequency Maintenance Loss
  - Through Wavelet Decomp.

$$\mathcal{L}_{wave} = \mathcal{L}_1(\tilde{\phi}(I_{sr}^t), \tilde{\phi}(I_{sr}^0))) + \mathcal{L}_1(\sum \hat{\phi}(I_{sr}^t), \sum \hat{\phi}(I_{sr}^0))$$

- Discriminator Loss

$$\mathcal{L}_{disc} = log(1 - D(I_{sr}^t))$$

Experiments

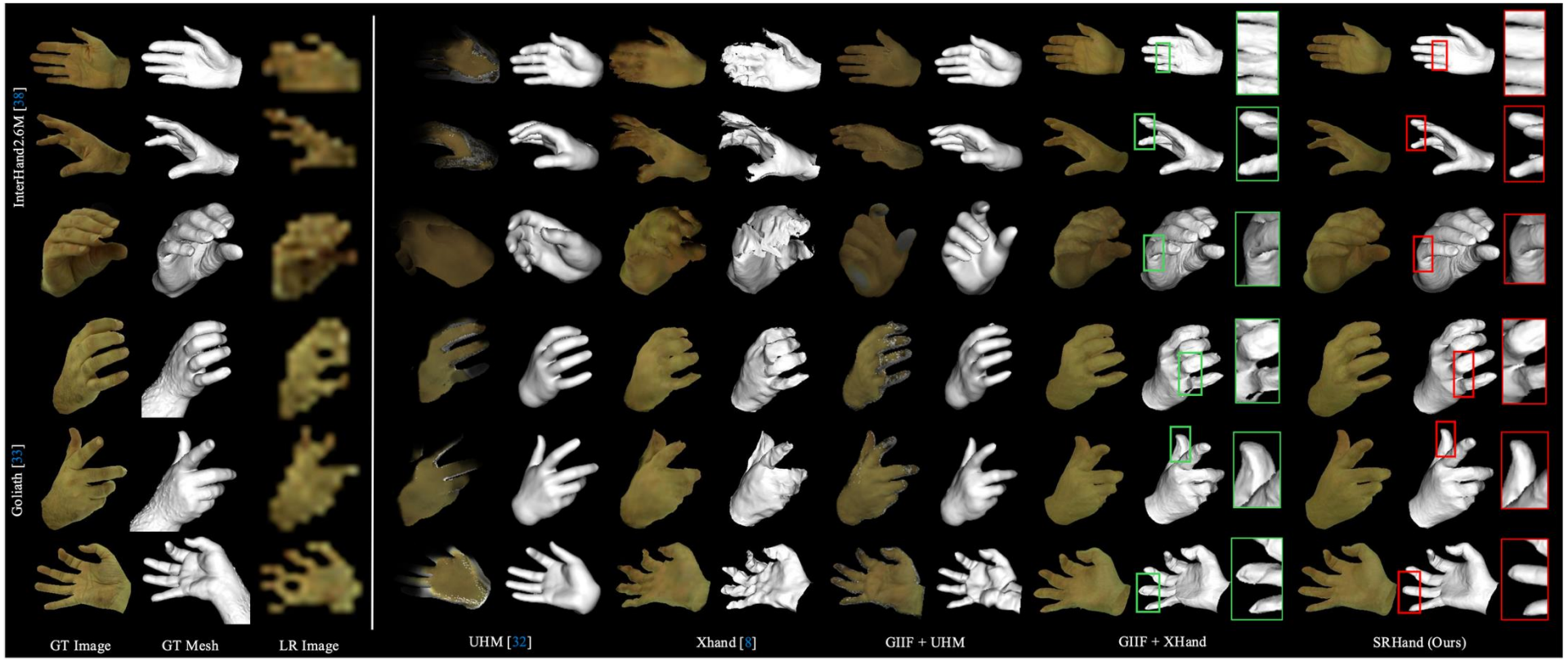
- 3D Reconstruction from SR Images

SR Module	3D Recon. Methods	InterHand2.6M [38]					Goliath [33]				
		PSNR / LPIPS (SR)	PSNR	LPIPS	P2P (*mm*)	Incon.	PSNR / LPIPS (SR)	PSNR	LPIPS	P2P (*mm*)	Incon.
Bicubic	Ours	22.23 / 0.2645	26.44	0.0895	4.01	0.0131	19.17 / 0.3244	22.52	0.1377	5.80	0.0128
LIIF [6]	XHand [10]	27.47 / 0.1063	27.36	0.0691	4.32	0.0151	24.87 / 0.1459	23.64	0.0984	3.34	0.0146
	Ours		27.11	0.0755	3.39	0.0151		22.87	0.1123	4.12	0.0140
GIIF (w/o ftd.)	UHM [37]	29.96 / 0.0305	22.33	0.1522	72.55	-	27.91 / 0.0497	23.85	0.1319	24.29	-
	XHand		27.71	0.0507	3.43	0.0067		22.76	0.1118	3.70	0.0084
	Ours		29.17	0.0404	3.09	0.0058		23.50	0.0783	3.49	0.0070
GIIF (w/ ftd.)	XHand	30.03 / 0.0303	28.75	0.0443	3.45	0.0052	28.07 / 0.0495	21.95	0.1139	3.60	0.0082
	Ours	30.06 / 0.0302	29.88	0.0362	2.16	0.0050	28.09 / 0.0495	24.31	0.0813	3.50	0.0069

KCVL

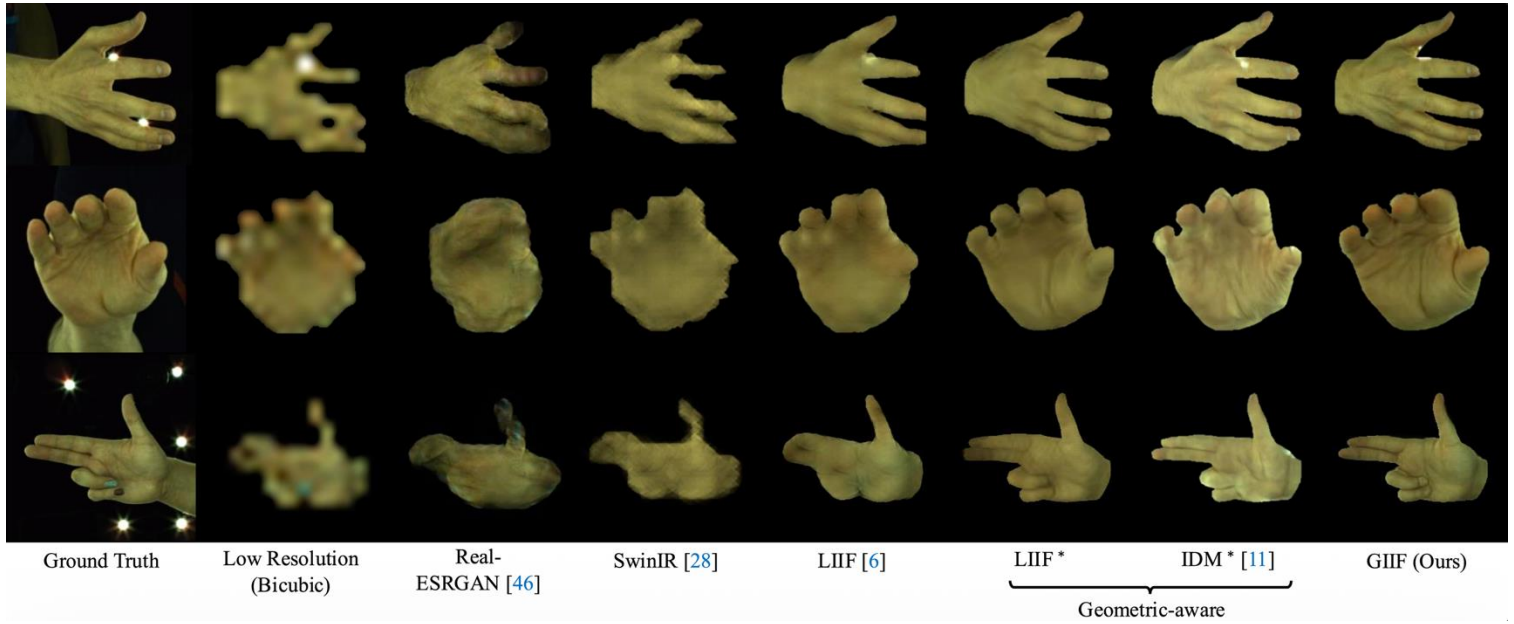
Experiments

- 3D Reconstruction from SR Images



Experiments

- Hand Image Super-Resolution Results



(a) Experiments are performed with $\times 16$ upscaling. (* denotes the model has been modified with normal map conditioning.)

Methods		PSNR ↑	LPIPS ↓
Real-ESRGAN [46]		22.39	0.2287
SwinIR [28]		25.51	0.1552
LIIF [6]		25.76	0.1848
IDM [11]		14.58	0.3603
Geometric-aware	LIIF*	29.85	0.0996
	IDM*	21.49	0.0970
	GIIF	**31.60**	**0.0637**

(b) Results in (PSNR / LPIPS) of continuous scale trained on $\times$ 16 factor. GIIF achieves best performance in all scaling factors.

Methods	Upscaling factor		
	$\times 8$	$\times 21.3$	$\times 32$
LIIF [6]	28.62/0.1319	23.91/0.2071	21.46/0.2693
IDM [11]	25.46/0.1215	21.28/0.1796	19.02/0.2426
LIIF*	30.20/0.0812	29.17/0.0855	28.55/0.0885
IDM*	22.68/0.0696	22.71/0.0780	22.87/0.0878
GIIF	**32.70/0.0533**	**31.53/ 0.0606**	**30.01/0.0640**

KCVL

Conclusion

- We present SRHand that integrates view/pose-aware implicit neural representations with explicit 3D mesh reconstruction.
- Our approach leverages a geometric-aware implicit image function (GIIF) to super-resolve low-detailed hand images in arbitrary scale.
- Achieves state-of-the-art performance in qualitatively and quantitatively.

KCVL

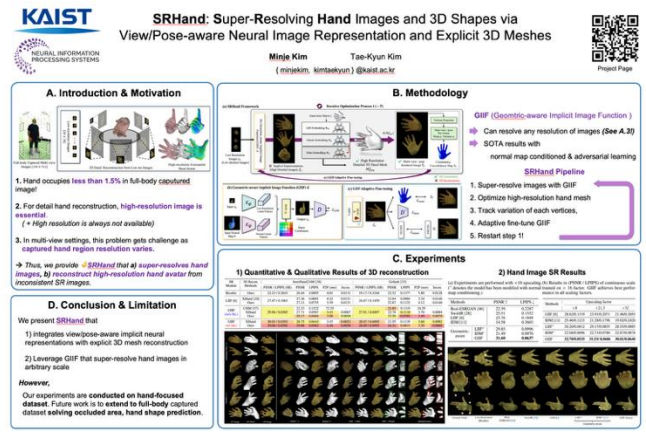
Thank you 🙏

Contact : minjekim@kaist.ac.kr

Paper

Project Page

yunminjin2/SRHand



Code

Poster