CVPR SEATTLE, WA JUNE 17-21, 2024

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

Minje Kim, Tae-Kyun Kim

KAIST

Imperial College London

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

KCVL

Introduction

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

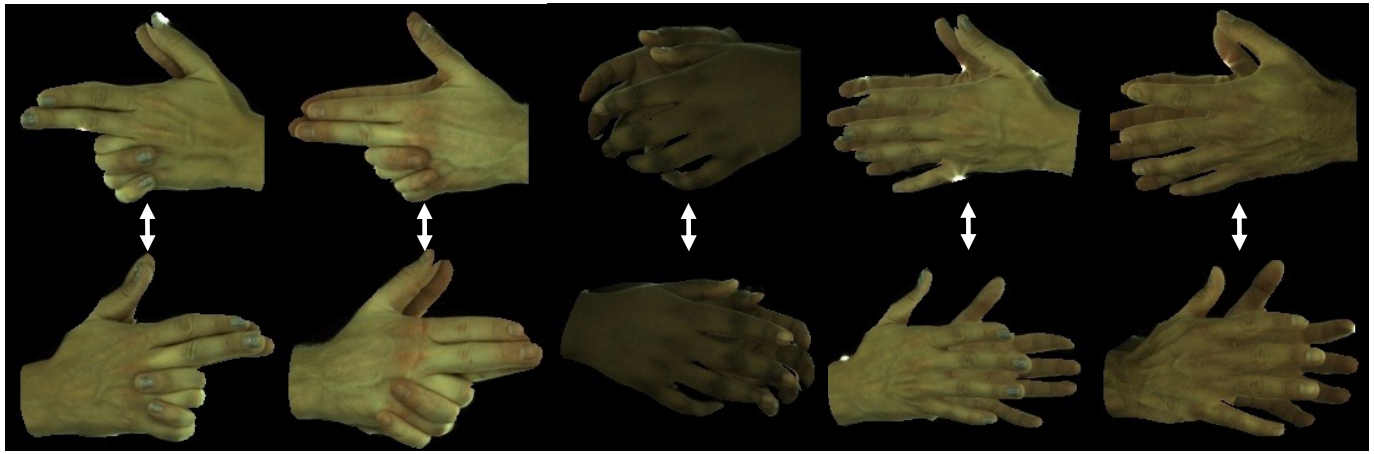
- **BiTT** : The first method for reconstructing both hand textures from a single image

① Symmetric information of left ↔ right hand

② Hand Texture Parametric Model

Motivation

- Both hands usually has similar appearance.
- By using both hand similarity, it is able to recover occluded parts.

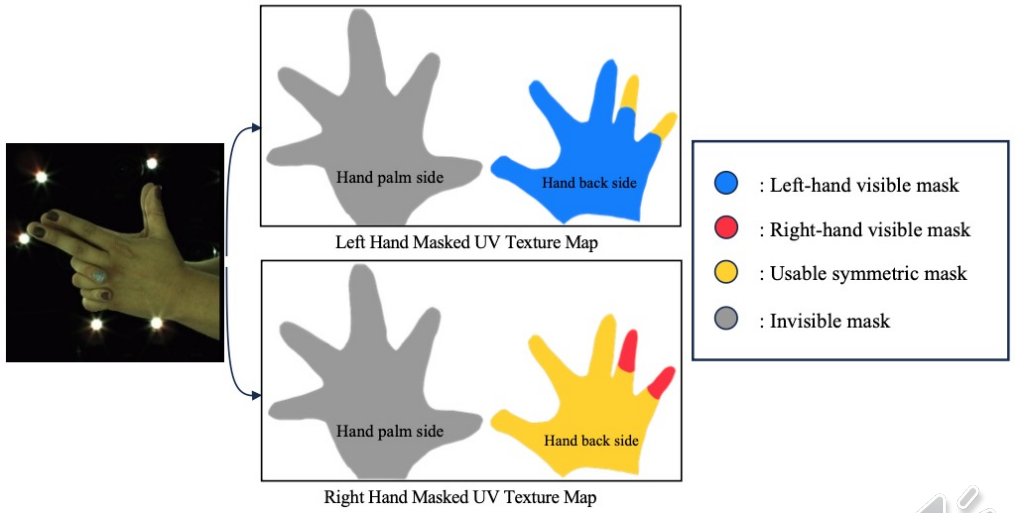


Motivation

- Percentage of visible texture for each hand part

Dataset	InterHand2.6M [28]	RGB2Hands [40]
Left-hand visible texture	35.40%	39.72%
Right-hand visible texture	36.44%	39.06%
Usable symmetric texture	24.29%	10.21%
Invisible texture	19.89%	24.95%

- Using usable symmetric texture,
  - → Can acquire up to 50~60% of the hand texture



Related Work

- 3D Mesh Representation
  - Neural Radiance Field
    - Learning implicit color field, and shape.
    - Requires lots of images for training the model.
    - Requires additional mesh extracting process for applying at general environment situation.
  - Mesh-based Representation
    - Representing through mesh, texture UV map (or vertex colors)
    - Easy to handle and manage interactions with other objects.
    - No need for additional process for utilization.
- Hand Texture Parametric Model (HTML)
    - Our method

Comparing with prior works

- Prior works mainly focus on reconstructing hand appearance **on a single hand** from **multiple images** (Multiview, Monocular video, ~ thousands of image)

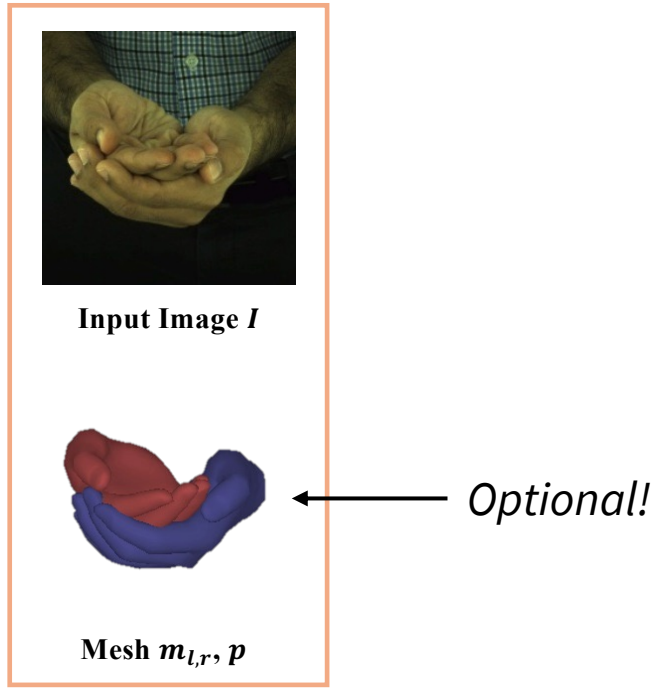
- BiTT reconstructs hand appearance of **both hands** from **a single image**

- **Why reconstruct from a single image?**
  1. Obtaining sufficient data to generalize can take a long time.
  2. Multiple image does not guarantee to produce full visible part.
  3. Occluded texture reconstruction is necessary.

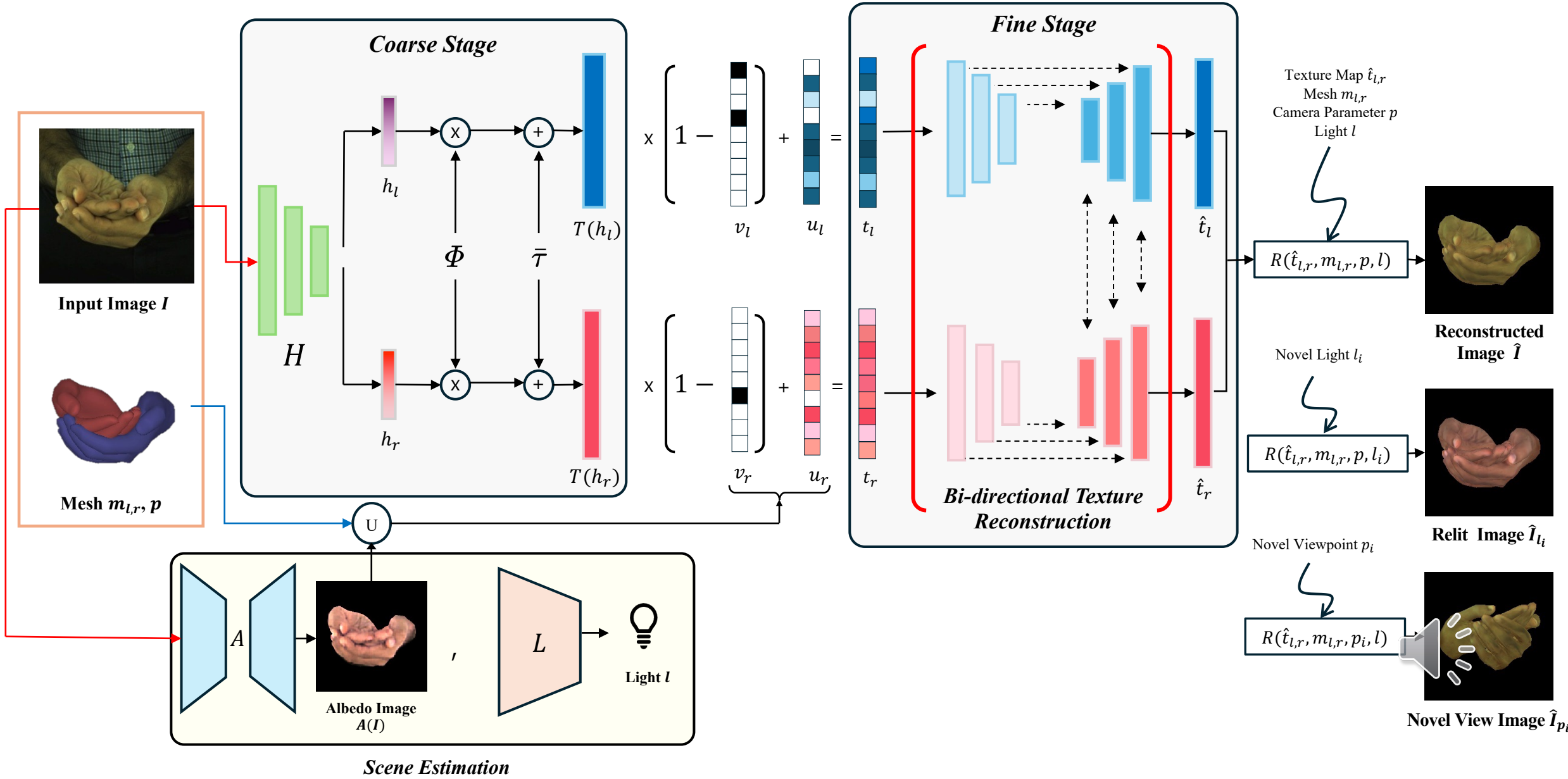
→ We demonstrate that a single image is sufficient for reconstructing both hand texture.

Methods

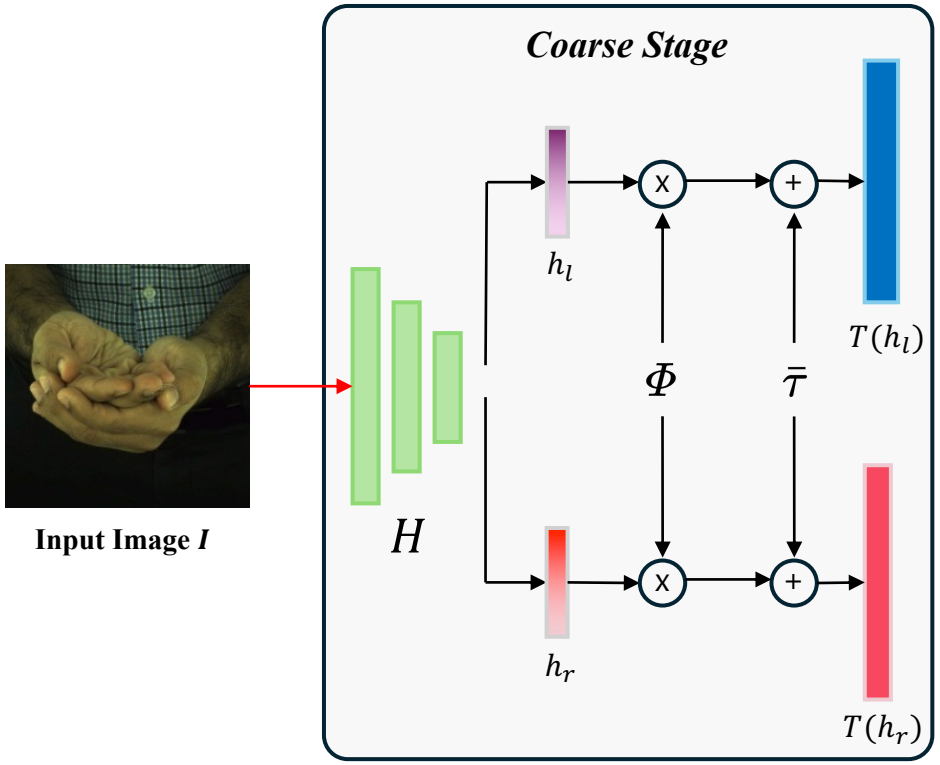
- Training framework
  - BiTT is based on per-scene optimization process (~ NeRF)
  - NeRF requires multiple image of the same scene
  - ↔ BiTT requires single image of the scene (hand)
  - Training time takes less than 7 minutes for each hands.



Methods - Overall Architecture



Methods – Course Stage

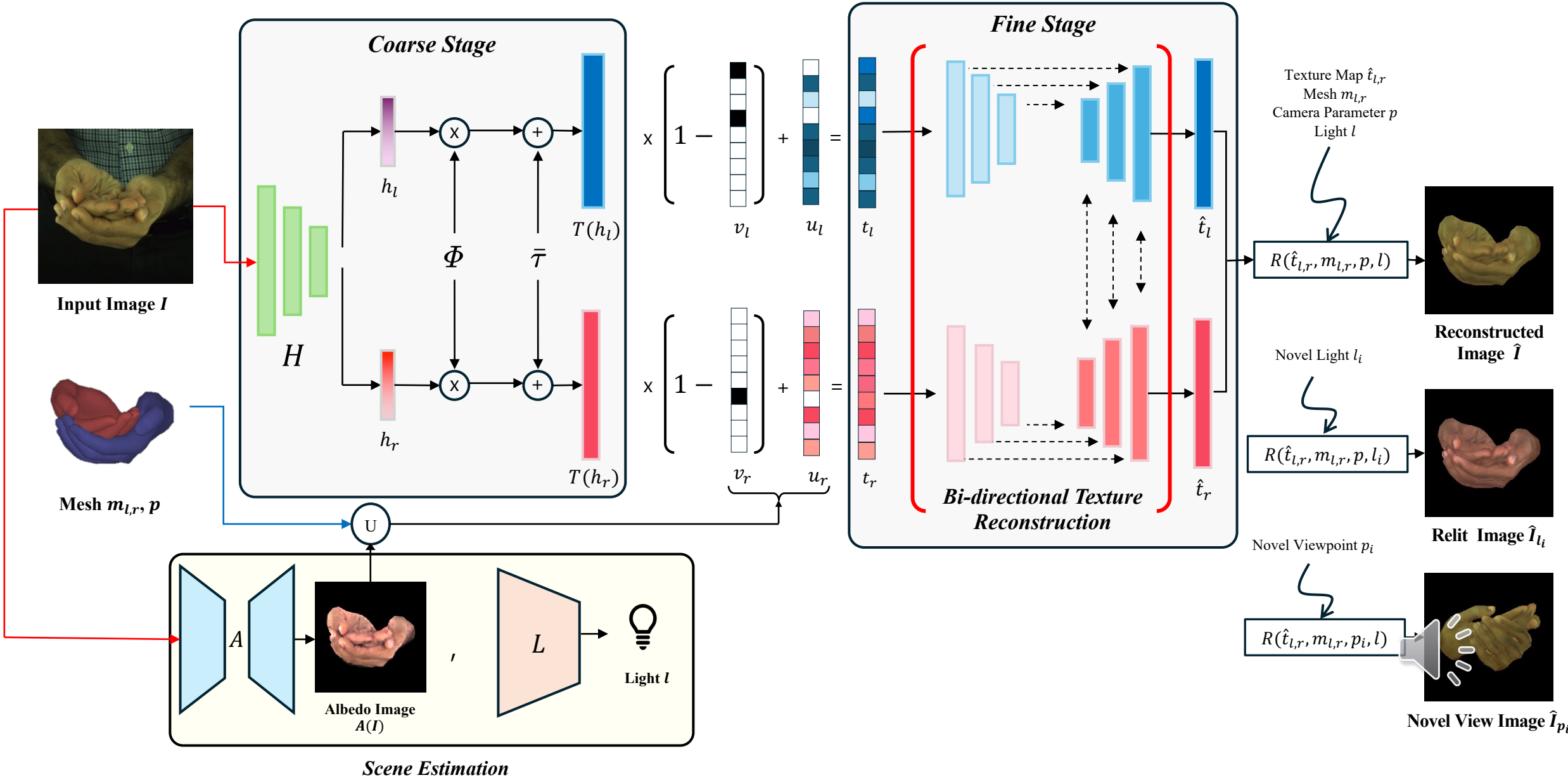


$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\tau_i$$

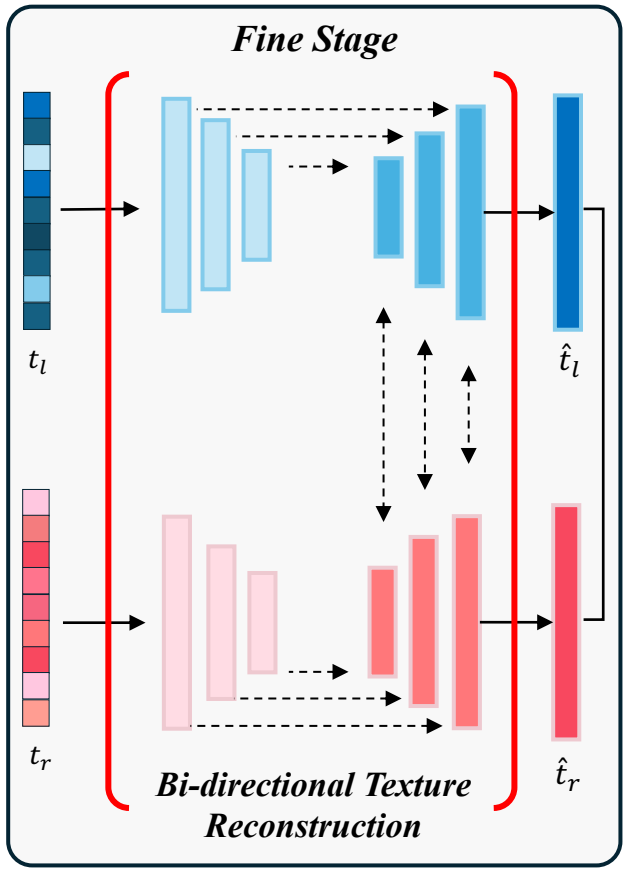
$$T(h_i) = \bar{\tau} + \Phi h_i, \text{ where } h_l, h_r = H(I), i = l, r$$

$$\hat{I}_{coarse} = \{R(T(h_i), m_i, p, l)\}_{i=l,r}$$

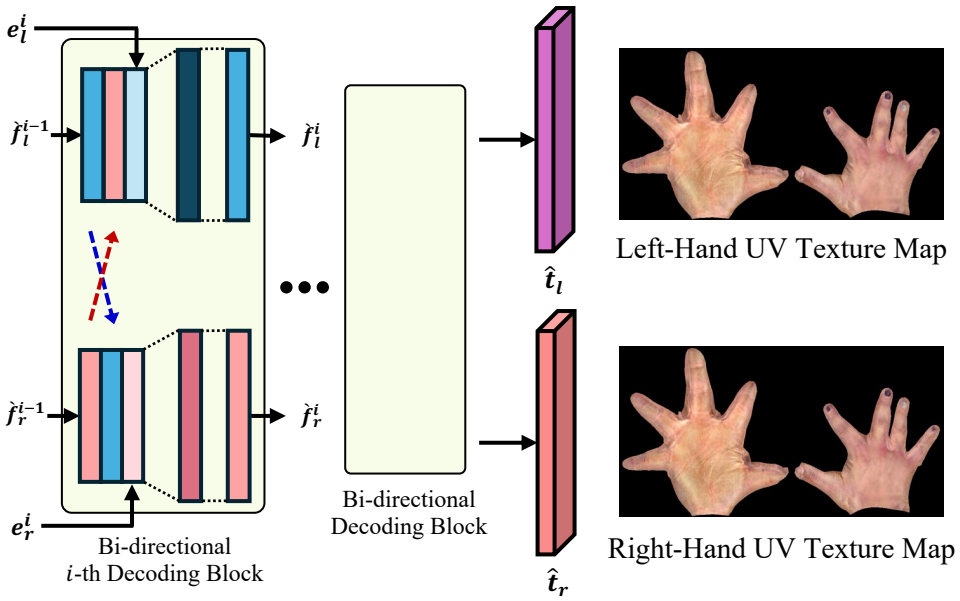
Methods - Overall Architecture



Methods - Fine Stage



Bi-directional Decoding block



Left-Hand UV Texture Map

Right-Hand UV Texture Map

$$\hat{f}_l^i = \sigma(\mathcal{N}([e_l^i, \hat{f}_l^{i-1}, \hat{f}_r^{i-1}]))$$

$$\hat{f}_r^i = \sigma(\mathcal{N}([e_r^i, \hat{f}_r^{i-1}, \hat{f}_l^{i-1}]))$$

Please check the details (ex. manicure, vessels, wrinkles) represented on the texture UV map!

Methods – Loss Functions

$$\mathcal{L}_{rec} = \lambda_{rec}\|(I - \hat{I})\|_1 + \lambda_{rec}^{coarse}\|I - \hat{I}_{coarse}\|_1 + \lambda_{rec}^{albedo}\|I - \hat{I}_{albedo}\|_1$$

$$\mathcal{L}_{nv} = \sum_{i=l,r} \|((T(h_i) - \hat{t}_i)(1 - v_i)\|_1$$

$$\mathcal{L}_{alb} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\|A(\hat{I}_{coarse,\hat{l}_i}) - A(\hat{I}_{coarse,\hat{l}_j})\|_1].$$

$$\mathcal{L}_{sym} = \lambda_{sym}\|(\hat{t}_l - \hat{t}_r)\|_1$$

$$\boxed{\mathcal{L} = \mathcal{L}_{rec} + \lambda_{nv}\mathcal{L}_{nv} + \lambda_{alb}\mathcal{L}_{alb} + \lambda_{sym}\mathcal{L}_{sym}}$$

Experiment

- Quantitative Results of InterHand2.6M dataset

(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**

Experiment

- Quantitative Results of RGB2Hands dataset

Evaluation	Method	L1↓	LPIPS↓	PSNR↑	SSIM↑	MS-SSIM↑
Appearance Reconstruction	S2Hand [6]	0.0179	0.0601	25.72	0.9459	0.9286
	HTML [34]	0.0203	0.0923	24.42	0.8927	0.9075
	HARP [17]	0.0155	**0.0433**	25.63	0.9309	**0.9344**
	BiTT(ours)	**0.0148**	0.0683	**26.02**	**0.9501**	0.9323
Novel Poses	S2Hand	0.0222	0.0778	24.22	0.9326	0.8991
	HTML	0.0233	0.0961	23.25	0.8829	0.8900
	HARP	0.0208	**0.0758**	23.88	0.9043	0.9042
	BiTT(ours)	**0.0196**	0.0774	**24.54**	**0.9352**	**0.9046**

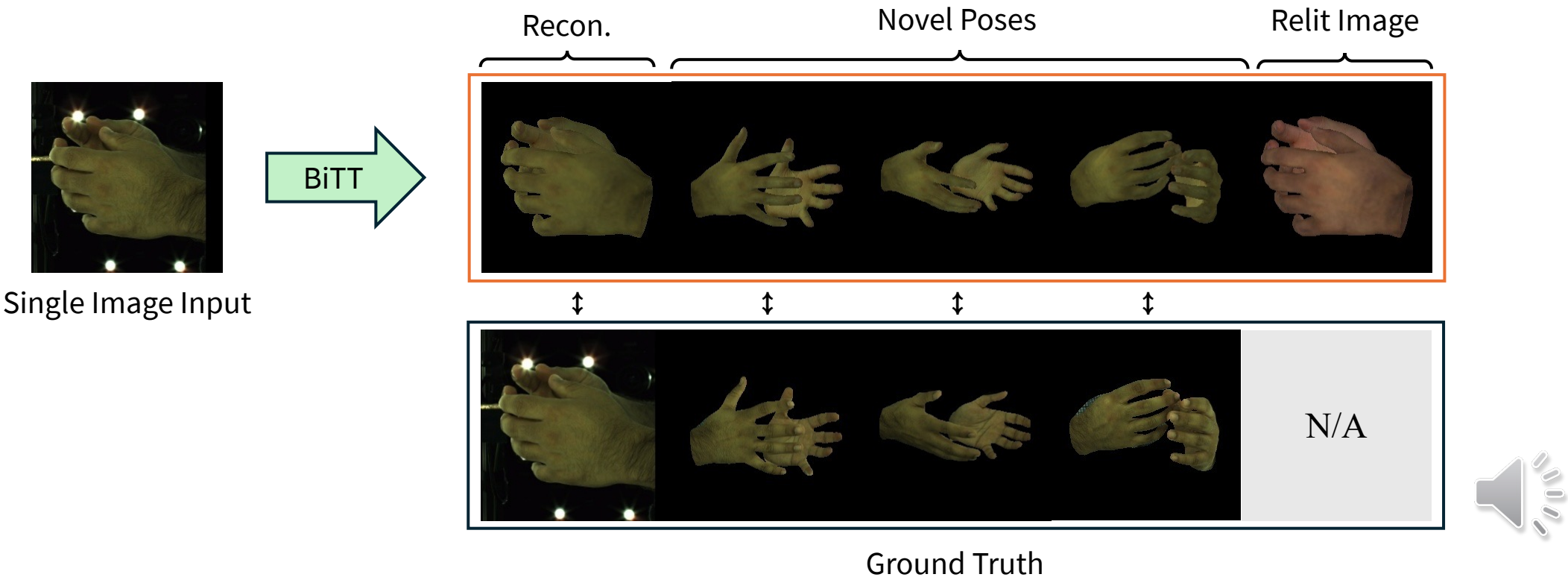
Dataset	InterHand2.6M [28]	RGB2Hands [40]
Left-hand visible texture	35.40%	39.72%
Right-hand visible texture	36.44%	39.06%
Usable symmetric texture	24.29%	10.21%
Invisible texture	19.89%	24.95%

Experiment - Qualitative Results

HTML S2Hand HARP BiTT(ours) Ground Truth

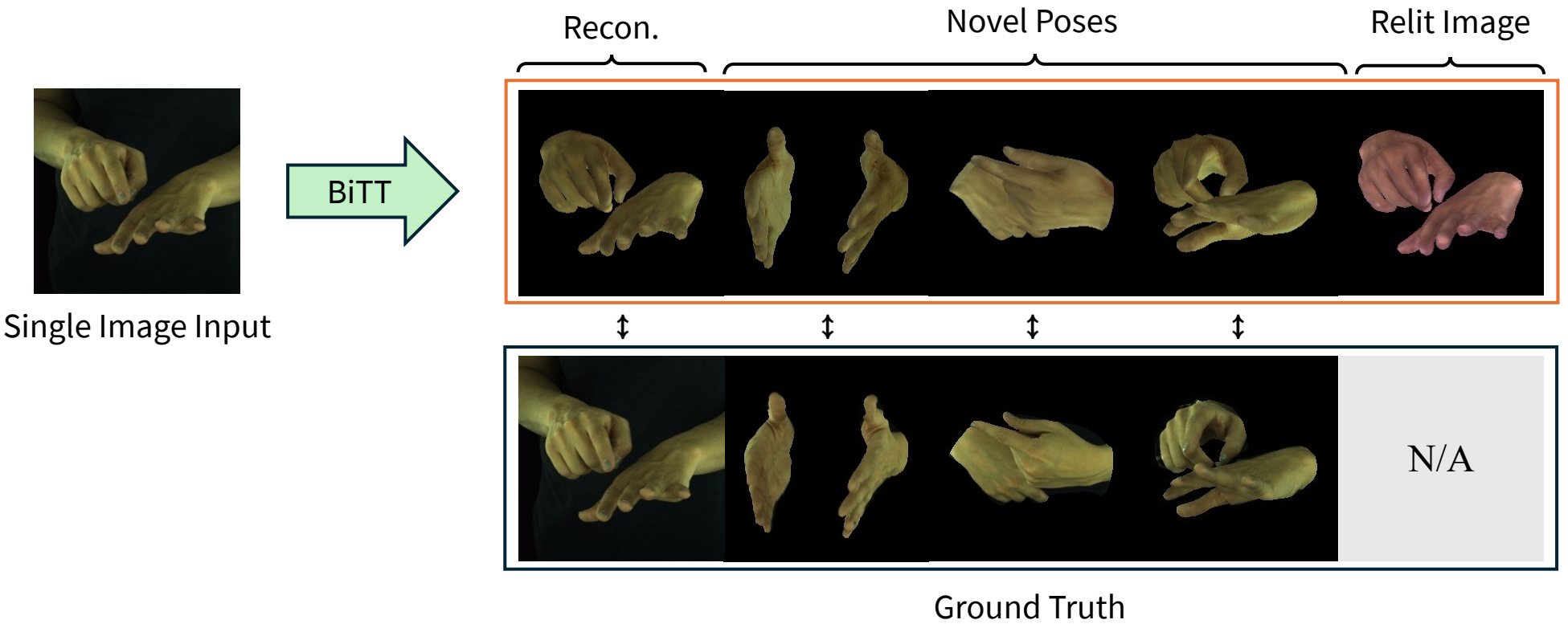
Experiment

- Qualitative Results



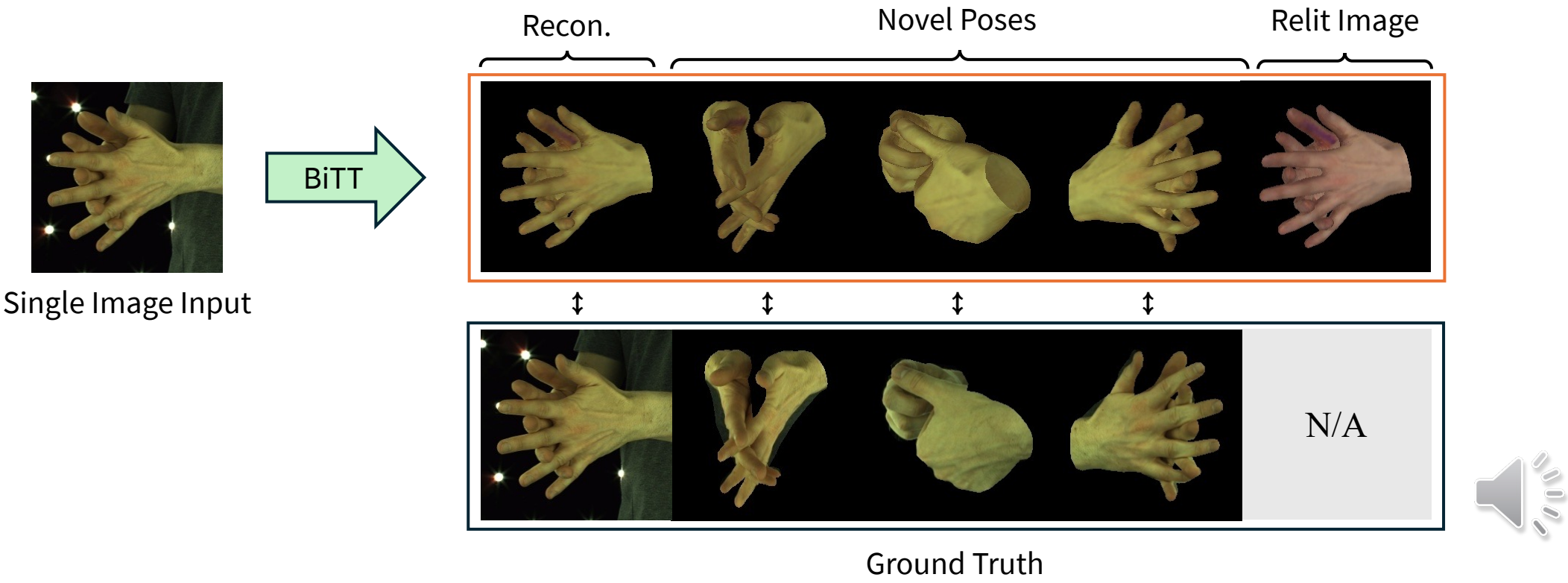
Experiment

- Qualitative Results



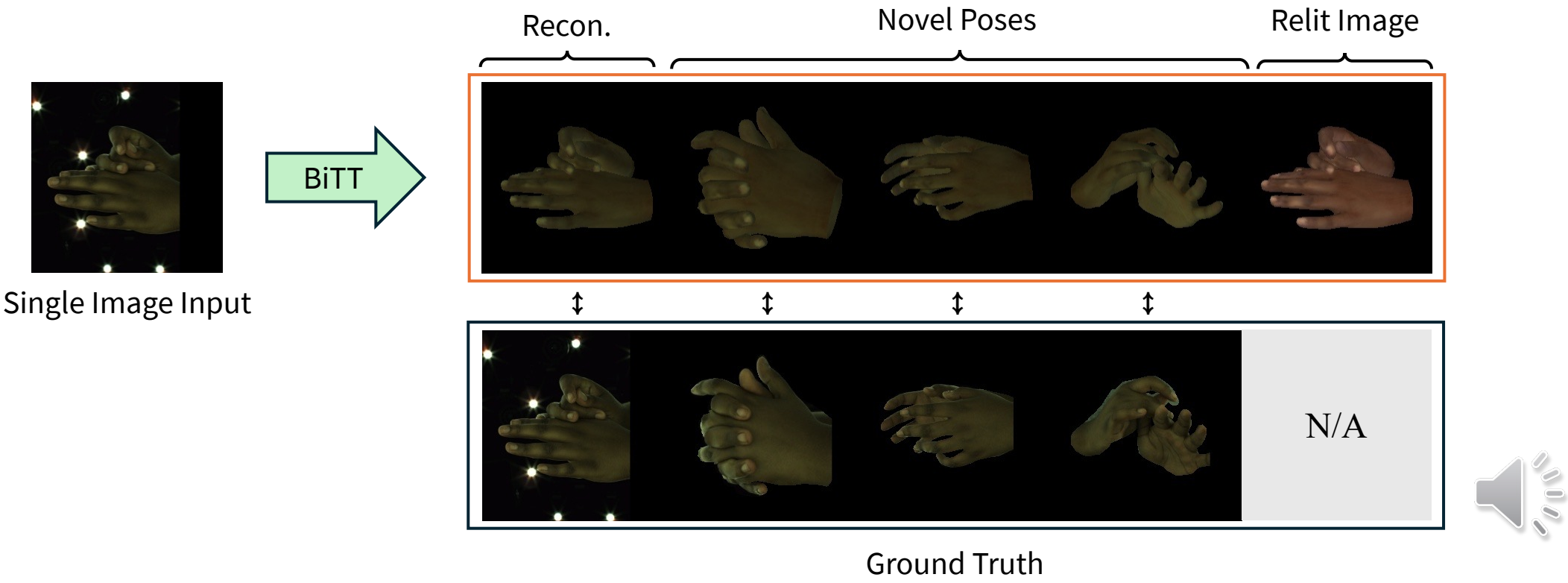
Experiment

- Qualitative Results



Experiment

- Qualitative Results



Experiment - Qualitative Results

- Results on Re:InterHand Dataset

Experiment - Qualitative Results

Conclusion

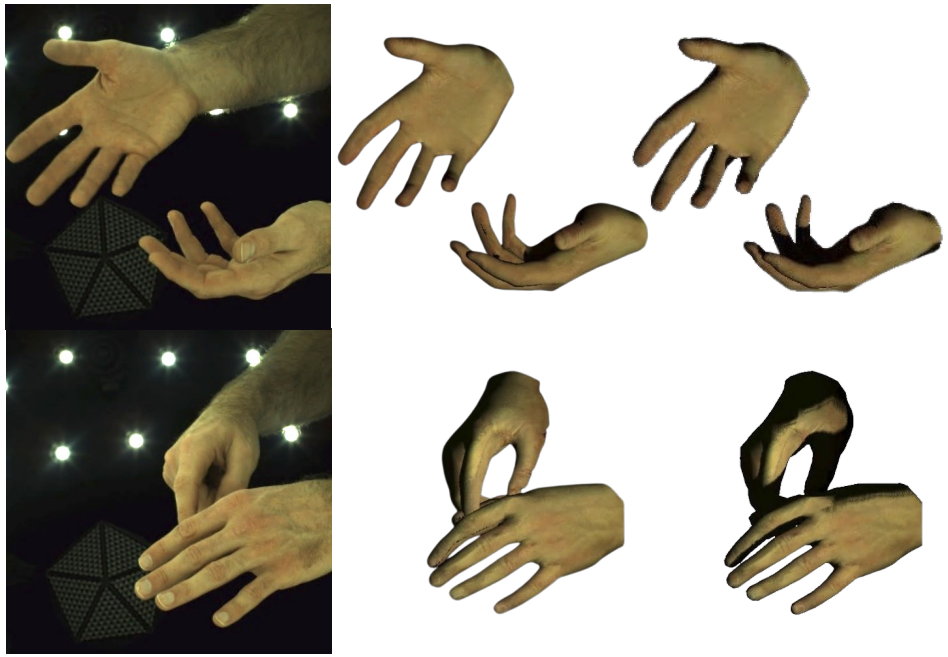
- We present **BiTT**, the first method for reconstructing both hands full textures with given single image.
  - High-fidelity
  - Efficient, Easy to train
  - Easy to modify (free viewpoint, pose, light scene)

- Demonstrates **utilizing paramatric model**, bi-directional reconstruction of texture between both hand is appropriate approach.

- Experiments has been conducted through all identities in dataset and achieved state-of-art performance.

Conclusion

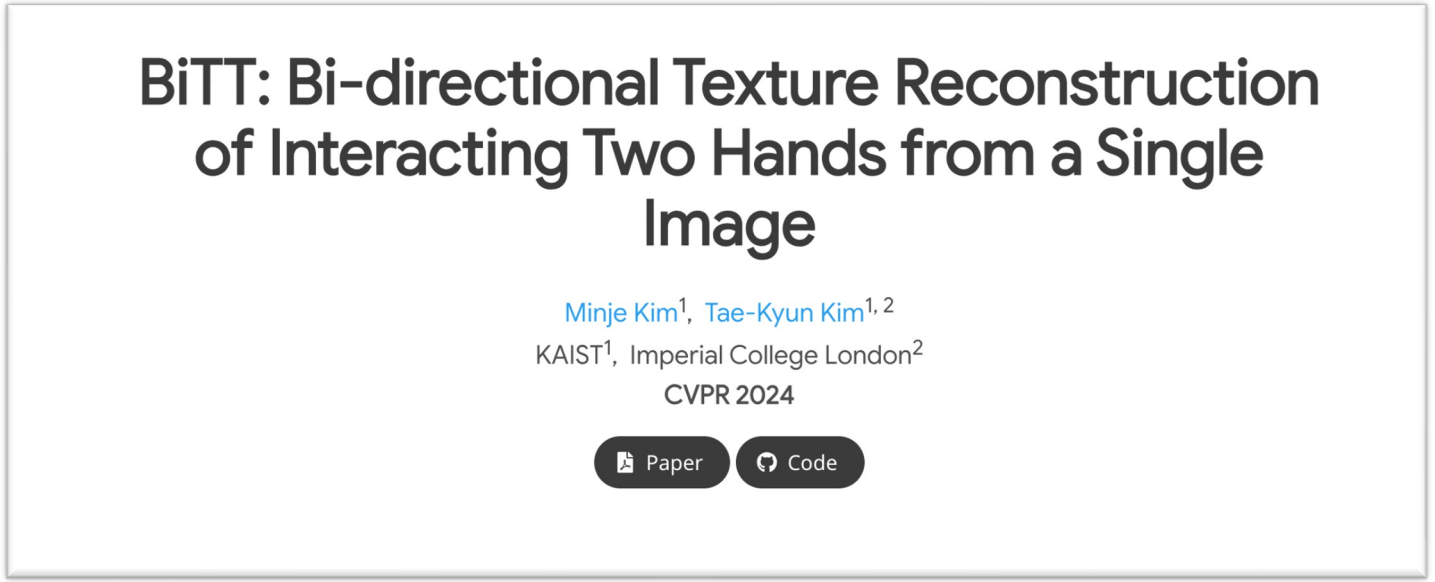
- Advantage of Explicit Representation
  - Can be easily compatible to traditional computer graphics
  - We applied self-occlusion aware shadow reconstruction



Input Image w/o Shadow w/ Shadow

Project Website & Code

- Project Website



- Code : https://github.com/yunminjin2/BiTT

Email Address: minjekim@kaist.ac.kr