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

- Supplementary Material -

1 In this supplementary material, we provide more ablation experiments and discussions that are not
2 included in the main paper due to the page limit.

3 1 Details of Dataset

4 We mainly use two datasets for experiments: InterHand2.6M [S4] and Goliath [S5]. Considering
5 practical scenarios in the real world where the hand occupies less than 1% from the full body captured
6 image, we set the upscaling factor to $\times 16$ in our main experiments. Furthermore, Sec. 3.6 describes
7 that our method achieves plausible results in any arbitrary upscaling factors.

8 1.1 InterHand2.6M

9 InterHand2.6M [S4] is constructed by capturing sequential frames from multi-view videos of single
10 hands and two-hand interactions. For each frame, we crop the hand region and resize it to specific
11 resolutions. To train GIIF, we use the 'train' split, with 'Capture0' \sim 'Capture22' as the training
12 dataset and the remaining captures ('Capture23' \sim 'Capture26') as the validation dataset.

13 For 3D hand mesh reconstruction, we primarily use a single hand from the 'test' split, which differs
14 from the GIIF training dataset, and incorporate identities from *Capture0* and *Capture1*. We use
15 a sequence from *ROM03_RT_No_Occlusion*, following prior works. For the 3D reconstruction
16 experiments, we sample 20 views with 20 frames. After training, we test on the 'cam400262' and
17 'cam400263' camera views with 398 frames. The results are reported as the mean value from both
18 camera perspectives.

19 1.2 Goliath

20 Goliath [S5] provides high-resolution images of a single hand, including OLAT (one-light-at-a-time)
21 captures and scanned meshes. Since our work does not strictly account for lighting conditions, we
22 sample fully illuminated images. As the Goliath dataset does not provide ground truth MANO [S1]
23 parameters, we fit the MANO mesh to the UHM [S8] template mesh using a joint loss for pose
24 and translation parameters and Chamfer distance loss for the shape parameter. Once the MANO
25 parameters are optimized, we use the 'test' split for 3D hand reconstruction of the subject 'AXE977'.
26 Frames are sampled uniformly at fixed intervals, and evaluation is performed by skipping frames.

27 2 Implementation Details

28 In this section, we provide more details on implementations of GIIF and prior work modifications.

29 2.1 Geometric-aware Implicit Image Function

30 Geometric-aware Implicit Image Function (GIIF) takes low-resolution images I_{lr} , and normal maps
31 N are extracted from the MANO [S1] template mesh. I_{lr} is encoded through the Residual Dense

Network encoder [S3], and N is encoded through the stacked hourglass encoder [S6]. After extracting each embedded feature, we concatenate the latent features along the channel-wise dimension. Following LIIF [S2], we employ cell decoding, local feature aggregation, and sample coordinates as queries in MLPs. After obtaining the predicted color for each query coordinate, we use adversarial learning with a discriminator. Our discriminator architecture is based on the UNet structure with spectral normalization.

We train GIIF for 20 epochs, after which we apply adversarial learning with our discriminator. The learning rate for GIIF is 0.0001, and for the discriminator, it is 0.001. Both are trained for 10 epochs.

2.2 Geometric-aware LIIF

Geometric-aware LIIF, denoted as LIIF* in this paper, geometric-aware LIIF is implemented similar to the GIIF. Compared to the GIIF, geometric-aware LIIF is not trained with adversarial learning.

2.3 Geometric-aware IDM

We modify IDM to create a geometric-aware version, denoted as IDM* in this paper. IDM [S7] is based on an encoder-decoder framework, where the decoder represents a continuous field and the encoder takes a normal map as a condition. Following conventional conditioning mechanisms, we incorporate a residual attention network and a style layer, enabling the encoder to output features that fuse both low-resolution (LR) image information and normal map details. The decoder then processes the concatenated feature representation to compute the target values based on query coordinates. This geometric-aware design enhances the model’s ability to preserve fine geometric details, resulting in sharper and more structured hand reconstructions.

2.4 Universal Hand Model

Since the Universal Hand Model (UHM) [S8] is designed for single-view reconstruction, it is not adaptable for multi-view reconstruction. Thus, we train UHM on a single view and evaluate it with trained viewpoints, which benefits the testing conditions. In total, 400 frames are used, with 200 frames for training and the remaining frames for evaluation, ensuring a balanced and evenly distributed dataset. Although UHM has easier testing conditions, our method still achieves better results, as shown in Tab. 3 in the main paper.

3 More Ablation Studies

3.1 Adaptive Fine-tuning

We present the results of inconsistencies at each step to demonstrate the effectiveness of adaptive fine-tuning in Fig. 1. The SR module is fine-tuned every 50 steps, starting from the midpoint of the total epochs. We observe that inconsistencies tend to decrease while enhancing performance within the time step t increases.

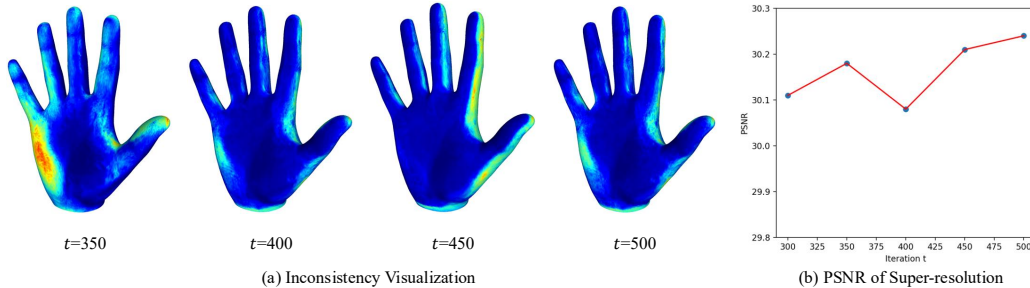


Figure 1: Qualitative results showing (a) inconsistencies in texture and (b) photometric performance variations over the time steps.

65 We further validate the effectiveness of the adaptive fine-tuning process in the 3D reconstruction
 66 process at 2. Applying adaptive fine-tuning enables the 3D reconstruction model to converge more
 stably by addressing inconsistency issues.

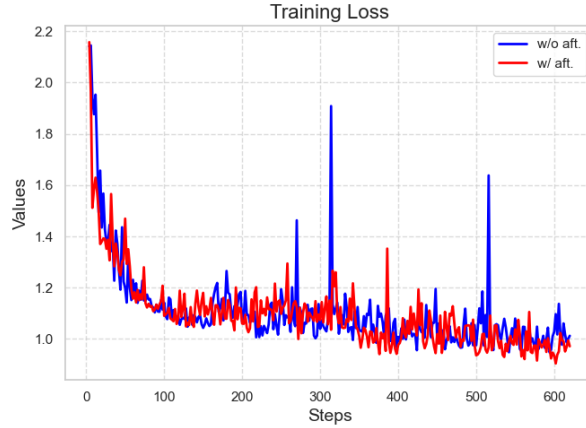


Figure 2: We plot 3D reconstruction training loss to show the effectiveness of adaptive fine-tuning. Mark "aft." stands for adaptive fine-tuning.

67 68 3.2 Effectiveness of Multi-View / Pose Consistency

69 We present the results of the effectiveness of multi-view consistency and pose consistencies. Tab. 1
 70 and Fig. 3 show the quantitative and qualitative results, respectively. Each multi-view consistency
 71 and multi-pose consistency helps to increase consistencies between images; notably, applying both
 72 consistency loss shows better results.

Table 1: Quantitative experiments of the effectiveness of multi-view consistency and multi-pose consistency.

View Cons.	Pose Cons.	PSNR (SR)	PSNR	LPIPS	P2P (<i>mm</i>)
✓		29.96	29.17	0.0404	3.09
		29.99	27.08	0.0541	3.23
✓	✓	30.01	28.05	0.0452	2.25
	✓	30.06	29.88	0.0362	2.16

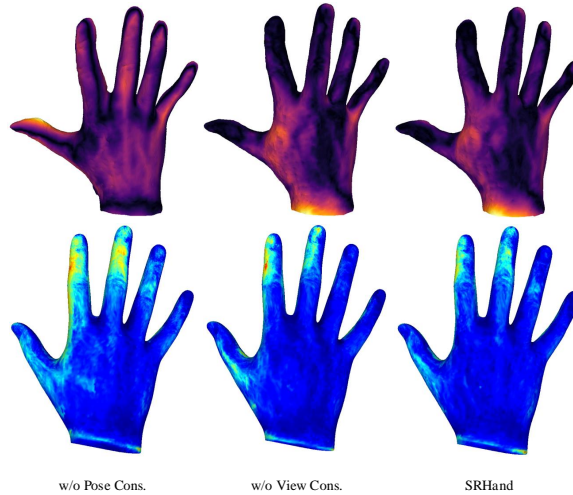


Figure 3: Qualitative results of multi-view, multi-pose ablation study.

73 3.3 Statistical Significance

74 We report the evaluation metrics averaged over 10 trials with different random seeds. The random
 75 seeds effect on the initial weight of MLP network in SRHand. As shown in Table 2, we report
 76 the mean, p-value, and 1-sigma (standard deviation) error bars of 3D hand reconstruction from SR
 77 images experiment (InterHand2.6M [S4]) across the runs. To validate the reliability of the statistical
 78 reporting, we conducted a Shapiro-Wilk normality test for each metric. The results ($p > 0.05$ for all
 79 metrics) confirm that the distribution of errors can be considered approximately normal, justifying
 80 the use of 1-sigma error bars under the assumption of normality.

Table 2: Quantitative evaluation results over 10 random seeds. We report each metric with the mean \pm standard deviation (1σ). Except for 'PSNR' in super-resolution, all metrics passed the Shapiro-Wilk normality test ($p > 0.05$). We report the PSNR of super-resolution as the median \pm interquartile range (IQR).

Category	Metric	Mean	1-sigma Error Bar	p-value
Super-Resolution	PSNR	30.07	± 0.055	0.0425
	LPIPS	0.0302	± 0.0002	0.6495
3D Reconstruction	PSNR	29.58	± 0.85	0.1319
	LPIPS	0.0382	± 0.0051	0.0911

81 3.4 Non-GT MANO Parameters

82 Although using ground-truth (GT) template parameters is a widely adopted strategy in 3D hand and
 83 body avatar reconstruction [S13, S14, S15, S9, S17, S18], and our method follows the same strategy,
 84 we additionally provide results where noise is added to the GT parameters. Tab. 3 shows that MANO
 85 parameter errors do not have a critical effect on super-resolution performance, while they affect the
 86 quality of 3D hand reconstruction.

Table 3: Quantitative results of showing the effectiveness of GT parameters.

	MPJPE (<i>mm</i>)	Super-Resolution		3D Reconstruction	
		PSNR	LPIPS	PSNR	LPIPS
Non-using GT Param.	9.54	30.11	0.0306	27.18	0.0635
Using GT Param.	-	30.06	0.0302	29.88	0.0362

87 3.5 Additional Baseline Experiments

88 We present additional experiments in Tab. 4 for HARP [S14] and LiveHand [S16]. HARP reconstructs
 89 a 3D hand using mesh subdivision with 3,093 vertices, which is not sufficient to represent a detailed
 90 hand geometric shape (significantly less than ours: 49,281 vertices). Also, assuming a one-point
 91 light condition differs from the benchmark settings used in our experiments, and yields geometric
 92 artifacts. LiveHand employs a low-resolution NeRF combined with a super-resolution network for
 93 real-time hand rendering. While efficient, LR NeRF and CNN-based SR module likely loses the
 94 3D geometric information. Both methods encounter difficulties in reconstructing 3D hands from
 95 inconsistent super-resolved images and lack performance compared to ours.

96 3.6 Arbitrary Scale Experiment

97 Due to implicit-based image representation, our method can reconstruct a 3D hand with arbitrary
 98 resolution. We report the evaluation metrics in arbitrary upscaling factors along $\times 5.3$, $\times 8$, $\times 10.6$,
 99 and $\times 16$. Tab. 5 and Fig. 4 shows the arbitrary scale super-resolution and 3D reconstruction results.
 100 We report PSNR and LPIPS metrics for each super-resolved image and 3D reconstructed hand. The
 101 results indicate that lowering the upscaling factor correspondingly increases the performance.

Table 4: Quantitative comparison of 3D reconstruction results. For fair comparison, our GIIF module was used for super-resolving images.

SR Module	3D Recon. Methods	PSNR	LPIPS	SSIM	P2P (mm)
GIIF	HARP [S14]	26.25	0.0872	0.8773	5.20
	LiveHand [S16]	26.91	0.0821	0.8805	–
	XHand [S9]	27.71	0.0507	0.8876	3.43
	Ours	29.88	0.0362	0.9206	2.16

Table 5: Quantitative results of arbitrary scale experiment. The two left columns show the PSNR and LPIPS results of super-resolved images and the right two columns present photometric metrics of a 3D reconstructed hand.

	Super-Resolution		3D Reconstruction	
	PSNR	LPIPS	PSNR	LPIPS
$\times 32$	28.82	0.0341	28.23	0.0409
$\times 16$	30.06	0.0302	29.88	0.0362
$\times 10.6$	30.52	0.0293	30.08	0.0343
$\times 8$	31.05	0.0262	30.23	0.0341
$\times 5.3$	31.42	0.0274	30.43	0.0334

3.7 Occlusion Ratio Experiment

We also present complex occlusion experiments. Hand reconstruction is well-known to be challenging due to frequent self-occlusions and highly articulated poses. Our multi-view video setup is designed to alleviate these issues, yet detailed geometric reconstruction remains difficult in occluded regions. To simulate such challenging conditions, we reduce the number of training cameras to one or two and sample pose-varying frames. The occlusion rate is quantified through the percentage of invisible vertices of the hand in the camera viewing space. As expected, the reconstruction accuracy decreases as the occlusion rate increases, shown in the Tab. 6; however, not drastically.

Table 6: Quantitative results under diverse occlusion rates. The occlusion rate is measured as the percentage of invisible vertices in the camera viewing space.

Occlusion Rate (Train View #)	PSNR	LPIPS	SSIM
24.7% (1)	25.54	0.0704	0.8636
19.8% (1)	26.56	0.0597	0.8874
9.43% (2)	28.05	0.0510	0.8986
0%	29.88	0.0362	0.9206

3.8 Training and Inference Time

We present the training and inference time specifications of SRHand. In the case of image super-resolution, GIIF processes each image in 3.5 ms, requiring a total of 17 seconds to upsample all training images. In contrast, a diffusion-based method [S7] takes about 50 seconds per image, amounting to 5.5 hours in total. Tab. 7 provides a detailed comparison of the training and inference times of our full model. SRHand achieves shorter training times and faster inference speeds, reaching up to 56 FPS, enabling real-time rendering.

4 More Qualitative Results

GIIF. First, we present additional qualitative results of GIIF compared with prior works in Fig. 7. Previous methods suffer from blurry textures and fail to represent geometric hand shapes accurately. Among prior methods, LIIF [S2] achieves the closest representation of the hand; however, it still struggles with blurry textures and unrealistic hand geometry, especially when fingers overlap. IDM [S7] effectively captures hand-like texture appearances but produces hand geometry that deviates significantly from common hand shapes.

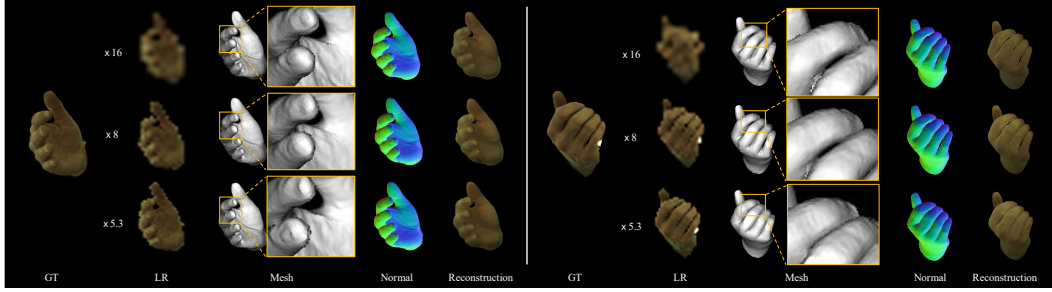


Figure 4: Qualitative results of arbitrary upscaling factor experiment. Lowering the upscaling factor shows a more accurate and detailed representation of geometry and appearance.

Table 7: Training and inference time comparison.

	Training Time (h)	Inference Time (ms / FPS)
UHM [S8]	5.5	139.5 / 7.2
SRHand	4.2	17.8 / 56

124 In both methods, incorporating a normal map as a condition improves the representation of hand
 125 geometry and textures. While both Geometric-aware LIIF and Geometric-aware IDM better capture
 126 hand structure and surface details, Geometric-aware LIIF still suffers from blurry textures and
 127 missing details, whereas Geometric-aware IDM exhibits noticeable differences from the referenced
 128 low-resolution image.

129 **Geometric Enhancements.** We present more qualitative results of geometric enhancements in Fig.
 130 5. We observe that our proposed method captures more details, such as nails and wrinkles, compared
 131 to XHand [S9].

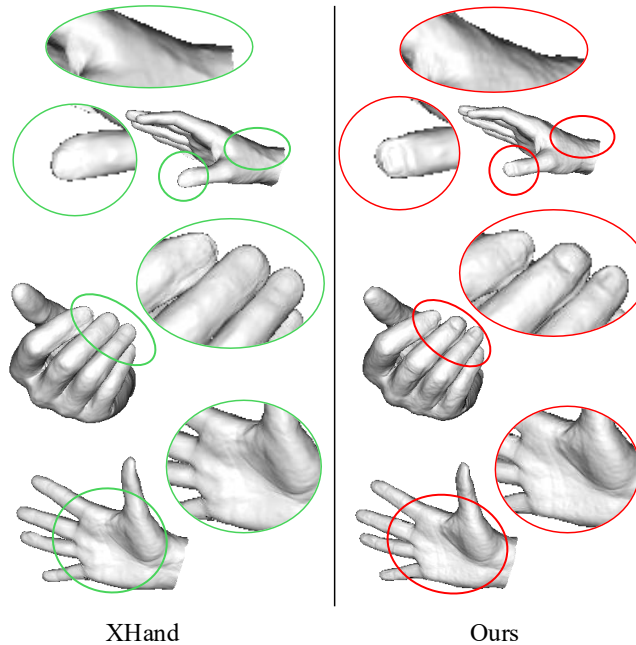


Figure 5: More qualitative results of showing geometric enhancements.

SRHand. We also present additional qualitative results of SRHand in Fig. 8. Starting from low-resolution images, our method reconstructs realistic and high-fidelity 3D hands. Furthermore, by predicting the albedo color of the hand, our method enables relightable scenarios, unlike previous work [S8], which is relighted from the captured scene environment.

5 Limitations and Future Works.

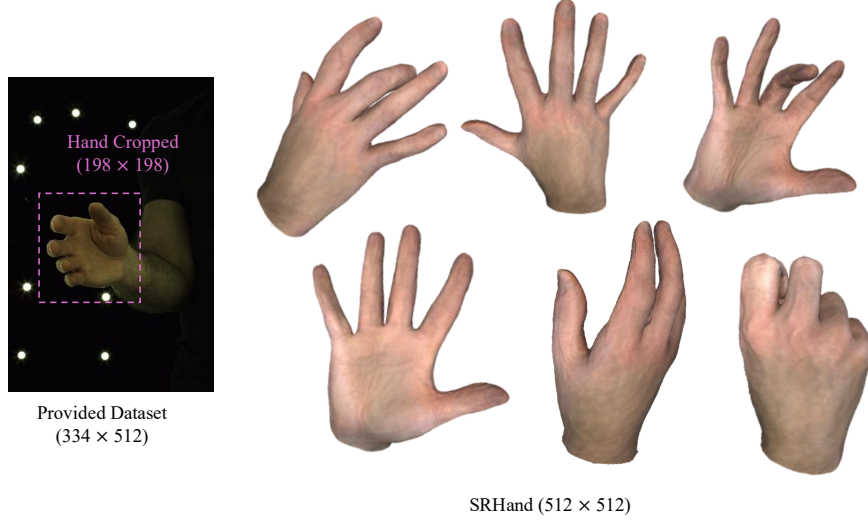


Figure 6: Qualitative results showing the generalization capability of our method to out-of-distribution scales, beyond the original dataset resolution.

We have demonstrated that SRHand performs well in the hand reconstruction domain, which is known for its high variance in pose articulations and view changes. All experiments are conducted on hand-focused datasets rather than a full-body captured dataset for fair comparison with prior works. Future work will be to extend our method to adapt to a full-body captured dataset. It remains to solve the occluded area caused by the body and to predict the accurate hand shape. Meanwhile, SRHand does demonstrate its generalization to out-of-distribution scales (see **Generalization Capability**). Our method can reconstruct high-fidelity 3D hands from 4K (4096×4096) images, given sufficient GPU memory. Further applying SRHand to deformable objects and full-body human reconstruction remains a promising direction for future work, to capture more detailed and diverse 3D shapes.

Generalization Capability. We further demonstrate that our method enables hand reconstruction to out-of-distribution scales (*e.g.* 512×512) that far exceed the resolution of given images from the datasets as shown in Fig. 6. This supports practical deployment conditions, highlighting the robustness of our approach to unseen scales.

Social Impact. Hand geometry and motion patterns inherently contain biometric cues, which could pose potential privacy concerns. Our model is trained on the InterHand2.6M [S4] dataset, which was collected for public research purposes and does not include identity labels. The GIIF module is designed to learn a generalized hand prior conditioned on normal maps, rather than encoding identity-specific representations, and the 3D reconstruction pipeline does not capture personal identifiers in the absence of such information in the input images. This design substantially reduces direct privacy risks. Nevertheless, geometric characteristics such as finger proportions and hand shape may still implicitly serve as biometric signals. To mitigate these risks, our study employs only publicly available datasets, avoids training with identity annotations, and emphasizes identity-agnostic objectives in both the GIIF and reconstruction stages.

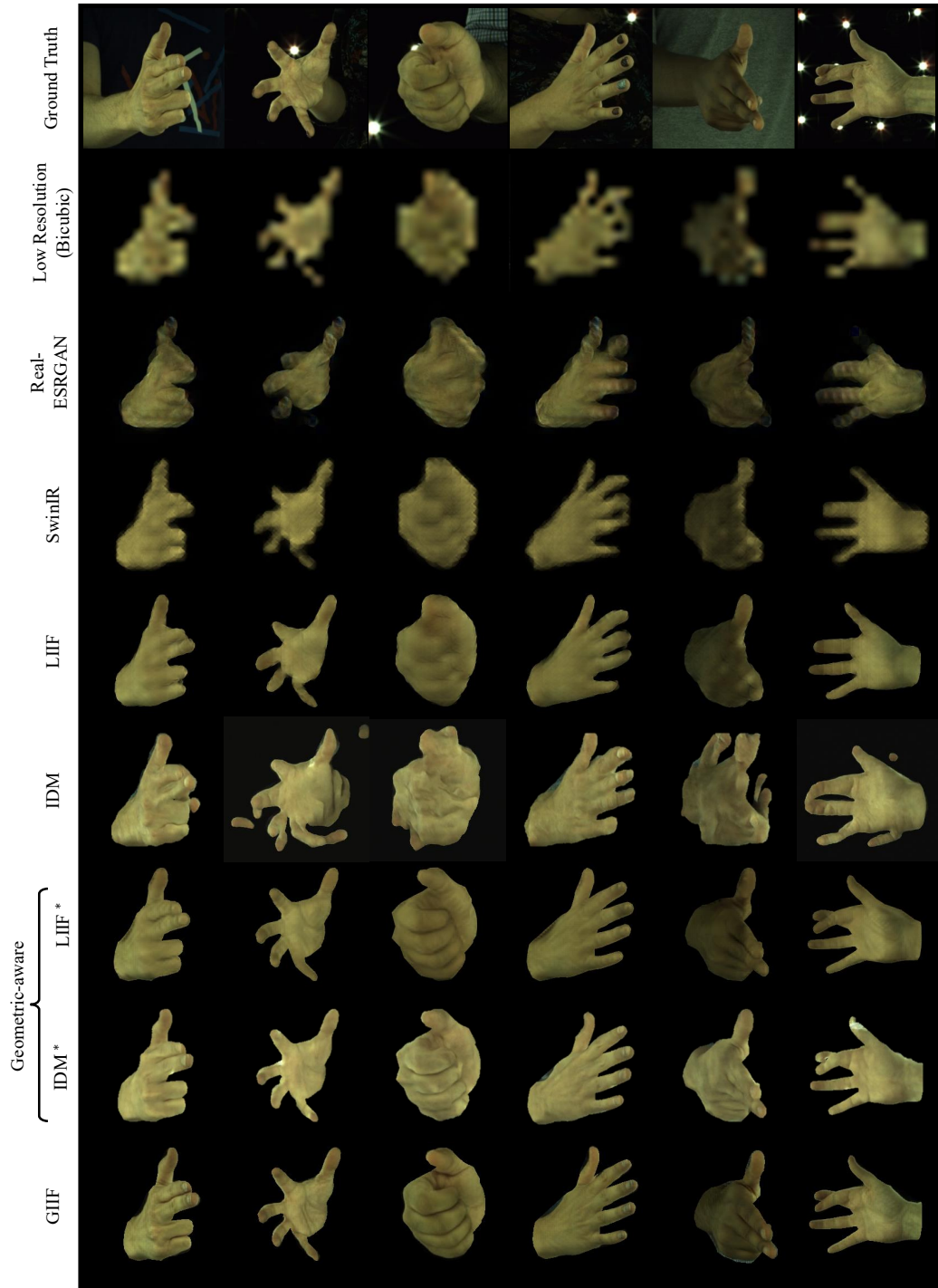


Figure 7: More qualitative comparisons of GIIF and prior works. Zoom in to check the details.

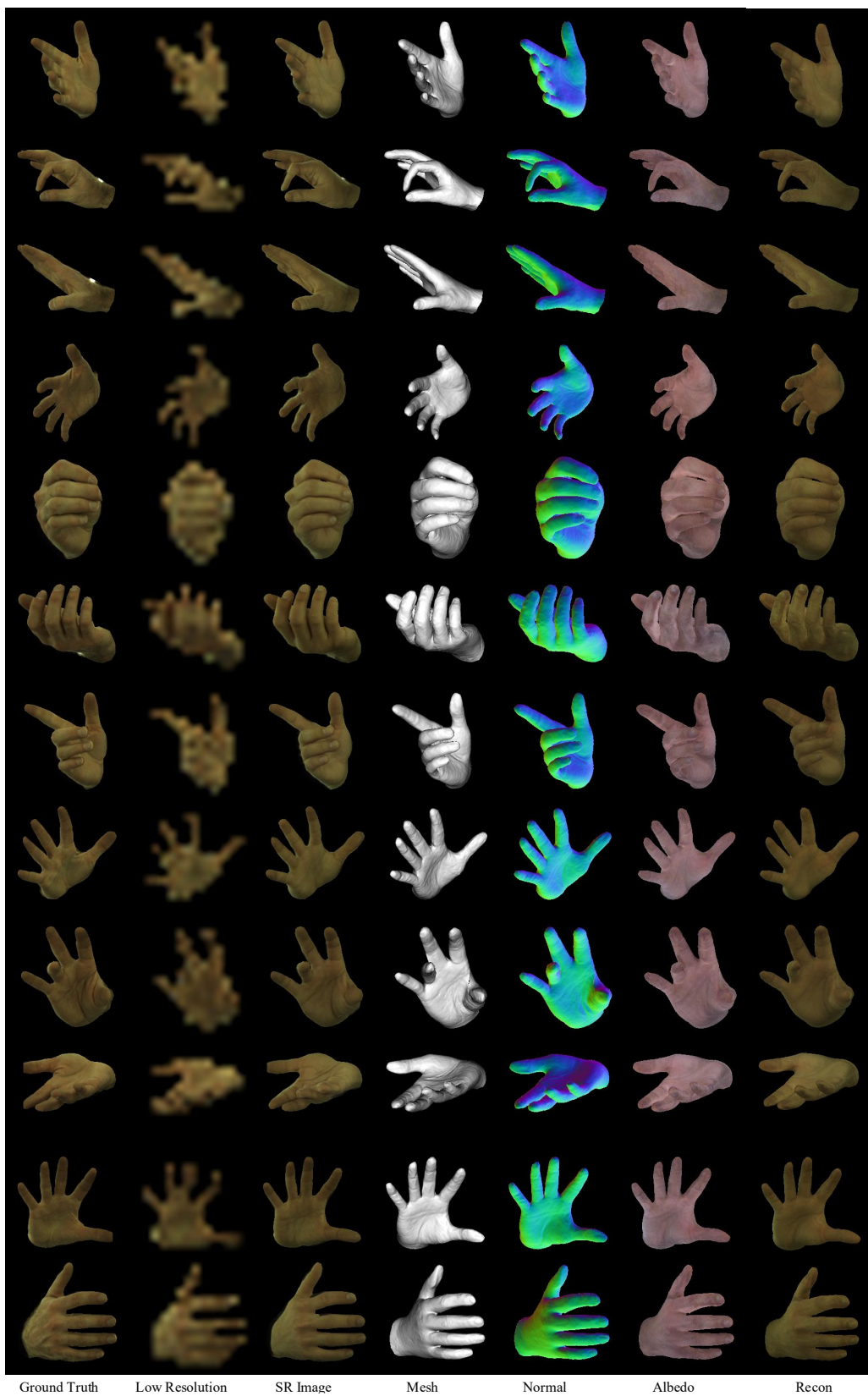


Figure 8: More qualitative results of SRHand. Zoom in to check the details.

References

- [S1] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017.
- [S2] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, 2021.
- [S3] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018.
- [S4] Gyeongsik Moon, Shouo-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020.
- [S5] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shouo-I Yu, Stuart Anderson, Michael Zollhofer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Hummerston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In *NeurIPS Track on Datasets and Benchmarks*, 2024.
- [S6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.
- [S7] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baочang Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, 2023.
- [S8] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. Authentic hand avatar from a phone scan via universal hand model. In *CVPR*, 2024.
- [S9] Qijun Gan, Zijie Zhou, and Jianke Zhu. Xhand: Real-time expressive hand avatar. *arXiv:2407.21002*, 2024.
- [S10] Yuchuan Tian, Hanting Chen, Chao Xu, Yunhe Wang. Image Processing GNN: Breaking Rigidity in Super-Resolution. In *CVPR*, 2024.
- [S11] Mary Aiyetigbo, Wanqi Yuan, Feng Luo, Nianyi Li. Implicit Neural Representation for Video and Image Super-Resolution, *arxiv:2503.04665*, 2025.
- [S12] Jinseok Kim and Tae-Kyun Kim. Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In *CVPR*, 2024.
- [S13] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *CVPR*, 2023.
- [S14] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *CVPR*, 2023.
- [S15] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [S16] Akshay Mundra, Mallikarjun B R, Jiayi Wang, Marc Habermann, Christian Theobalt, and Mohamed Elgharib. Livehand: Real-time and photorealistic neural hand rendering. In *ICCV*, October 2023.
- [S17] Xiaozheng Zheng, Chao Wen, Su Zhuo, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. Ohta: One-shot hand avatar via data-driven implicit priors. In *CVPR*, 2024.
- [S18] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024.
- [S19] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021.
- [S20] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. *arXiv preprint arXiv:2103.14006*, 2021.