SUPPLEMENTARY MATERIAL

LISA: Learning Implicit Shape and Appearance of Hands

Enric Corona[1] Tomas Hodan[2] Minh Vo[2] Francesc Moreno-Noguer[1]
Chris Sweeney[2] Richard Newcombe[2] Lingni Ma[2]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain [2]Reality Labs, Meta

In this supplementary, we first detail the implementation. Then, we present additional evaluation results. Last, we provide a video to demonstrate LISA's hand modeling properties and a new application to hand tracking.

1. Implementation details

We implement LISA following the open-source implementation of NeRF [3] from PyTorch3D [7]. The code is further adapted to integrate the VolSDF [9] algorithm, in order to predict SDF instead of density, and we use an additional MLP to learn disentangled texture. For the articulated models (LISA, NASA [1] and NARF [6]) we adopt the same characteristics to provide fair comparisons. We add positional encoding for NASA and duplicate the original geometry parts with parallel color MLPs. We also add the skinning weight loss in NARF [6] to help the model take advantage of each part-MLP, even though it was not explicitly described in their paper.

We use positional encoding with 10 frequencies for the input 3D positions and 4 frequencies for the ray direction. During training, we sample 512 rays per image and 64 points per ray between 2 and 8 depth units, which covers the range of camera positions with respect to the training hands. Since the training data provide foregound masks and the images contain a big part of background, we sample rays only within the larger bounding box (20 pixels extra) that contains the hand. Training is done on images at a resolution of 667×1024. The network is trained for 20k epochs with initial learning rate 0.0005 with an exponential decay to 0.00005. We use Adam [2] optimizer with $\beta_1 = 0.9$ $\beta_2 = 0.999$. As mentioned in the main document, we use L1 rendering loss, which resulted in sharper edges and details. The weight for each loss is $\lambda_{col} = 50$, $\lambda_{Eik} = 1e^{-3}$, $\lambda_w = 0.1$, $\lambda_{reg} = 1e^{-5}$, $\lambda_{surf} = 10$ and $\lambda_N = 10$.

2. Additional evaluations

Monocular reconstruction. In Fig. 1 we show more results on InterHand2.6M [5], in addition to the content of Figure 4 in the main submission. Similarly, Fig. 2 presents further examples of in-the-wild inference on the FreiHand dataset [10], which supplements Figure 5 of the main submission. Both experiments are achieved with the same settings as discussed before. We show LISA can reconstruct diverse testing scenario for a large range of poses, shapes and appearance. To highlight the limitations of LISA we also include failure cases. We think the main cause of errors is the fact that the proposed losses are not always sufficient to constrain the optimization for some challenging single-view cases. We plan to further study the current limitations in future works by including a learnt data-driven prior that can also allow faster inference.

Evaluation with noisy joints. In our experiments, we assume that ground truth 2D joints are available. These joints are obtained from heavy multi-view dense reconstruction and annotations in data collection. In practice, the joints can be inferred by state-of-the-art algorithms, which will introduce a certain amount of error. We therefore conduct an experiment to analyze how noise impacts our reconstructions. To be independent of any backbone detection algorithm, we add different levels of noise to the ground truth joints and summarize the performance in Tab. 1. The results indicate that our reconstructions yield similar performances for noise ratios of up to $\sigma \leq 5$mm. This range is close to the performance of state-of-the-art algorithms on hand joints detection. Note that the ground truth joints used in our evaluation are reported to have an average error of 2.78mm according to [4].

Visualization of error heatmaps We include a visualization of the error heatmaps, in the task of 3D hand reconstruction from point clouds, in Fig. 3. This shows the error per-point for a single example, and illustrates the fact that LISA provides the lowest quantitative error amongst baselines. Most notably, even though MANO provides realistic registrations, the use of implicit models leads to a significantly lower error during registration.

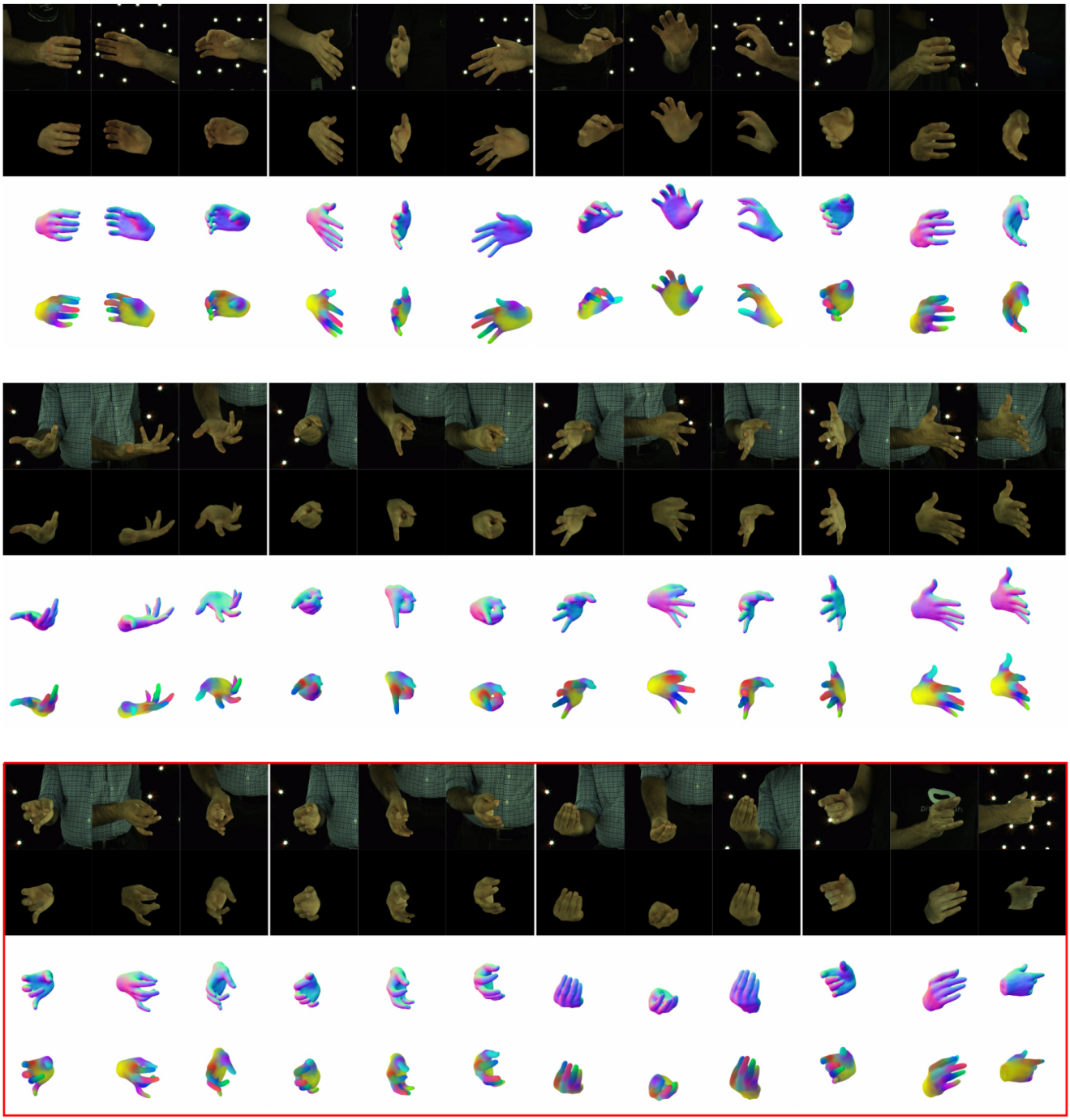

Figure 1. **Additional monocular reconstruction results for InterHand2.6M [5] produced by LISA-full**. In total 12 new cases are shown. For each scene, the four rows from top to bottom are: i) ground truth RGB views; ii) LISA's reconstruction rendered with appearance; iii) LISA's reconstruction rendered; and iv) LISA's reconstruction rendered with colored skinning weights. Note that all scenes are from the *test set*, with new users unseen during training. We only use the first camera for reconstruction, whereas the remaining cameras are shown for reference. The last 4 scenes marked in red are failure cases, where hand poses are not inferred correctly, despite the silhouette fitting well. This is due to the rendering loss from a monocular camera failing to constrain the optimization for these challenging cases. This problem can be mitigated by adding more views to reconstruction and by adopting temporary regulation in tracking.

Noise	GT Joint error	1 view				2 views				4 views			
		PSNR	V2V	V2S	Joint error	PSNR	V2V	V2S	Joint error	PSNR	V2V	V2S	Joint error
$\sigma = 0$	0 mm	25.43	3.84	3.68	8.63	29.40	3.70	3.56	6.81	29.69	3.53	3.38	6.17
$\sigma = 2mm$	$3.1mm$	25.18	3.99	3.84	8.65	29.27	3.71	3.56	7.01	29.50	3.71	3.55	6.21
$\sigma = 5mm$	$7.7mm$	24.86	4.18	4.04	8.43	28.93	4.06	3.92	7.40	28.91	3.99	3.86	6.65
$\sigma = 10mm$	$15.3mm$	24.20	5.50	5.39	12.19	28.41	4.84	4.71	8.62	28.38	4.44	4.32	7.16
$\sigma = 20mm$	$30.6mm$	24.19	6.96	6.87	17.32	27.99	5.75	5.63	10.46	27.59	5.01	4.90	8.76

Table 1. **Reconstruction on DeepHandMesh [4] with noisy joints.** To study the impact of noisy joints that can be produced by any detection backbone algorithm, we add Gaussian noise to the ground truth 3D joints and conduct reconstruction using the projected noisy 2D key points. The results suggest that our algorithm yields similar performances with low deviations of up to $\sigma \leq 5$mm.

Figure 2. **In-the-wild reconstructions generated by LISA-full.** We provide additional results on FreiHand [10] in addition to Figure 5 in the paper. Note that our model is not trained on this dataset but still can generate very plausible reconstructions even on significantly different lighting conditions. Each column from the first to the fourth row: input RGB image for single view reconstruction; the reconstructed mesh overlayed on the input; the reconstructed mesh in a novel view; and the reconstructed mesh rendered with texture in a novel view. The red box marks failure cases, where the reconstructed hand poses are wrong despite the silhouette fitting reasonably well.

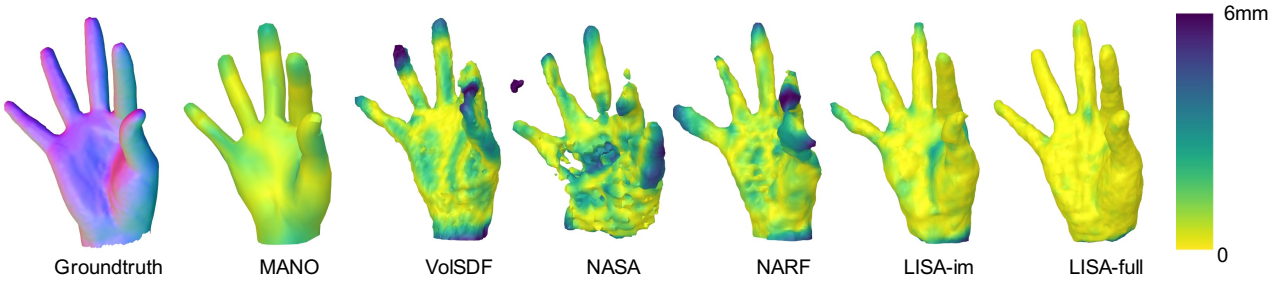

Figure 3. Visualization of error heatmaps from LISA and baselines when fitting 3D hand scans.

3. Video demonstrations

We encourage readers to view the video supplement, which summarizes this work and highlights two demos: (i) analysis of the generative properties of LISA hand modeling, and (ii) application of LISA to monocular hand reconstruction. Below we briefly discuss these two aspects.

Generative properties. The architecture of LISA is designed to automatically disentangle the embedding space of the hand shape, the hand pose and its appearance. To achieve this, we ensure each hand subject at training time is assigned the same shape and color embedding for different poses. In the video, we show hand animation by linearly interpolating a set of shape codes for a given hand with fixed shape and appearance. We also demonstrate pose transfer to different hand shapes, by applying a fix pose to a set of linearly interpolated shapes. Last, appearance is linearly interpolated for a fixed shape and pose.

Monocular hand tracking. The second part of the video presents two sequences of hand tracking with LISA. To this end, we assume the input is a monocular masked video, and 2D hand keypoints every six frames. For the first frame, we optimize the shape, pose and appearance for 4K iterations, following the same loss used in single-view hand reconstruction. Then, each successive frame is optimized for only 1200 iterations. In addition, the learning rate for the shape, the appearance and camera color correction are reduced to $1e^{-6}$ after the first frame. This setting in practice almost freezes these properties for all remaining frames, and only allows hand pose to adapt to new observations.

The demo sequences (captured at 30 FPS) are from the test set of InterHand2.6M [5] on two different subjects. The video shows LISA's reconstruction rendered in the input view as well as two novel views. For comparison, we also show the reconstruction from MANO [8], which we render with the same pose parameters. We remark that some frames show visible projection errors when rendered from novel viewpoints, while they fit well into the input views. Considering monocular dynamic reconstruction is a challenging ill-posed inverse graphical problem, but we believe our method lays a promising basis for future developments.

References

[1] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–628. Springer, 2020. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 1

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1

[4] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 440–455. Springer, 2020. 1, 3

[5] Gyeongsik Moon, Shoou-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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 4

[6] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[7] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv*, 2020. 1

[8] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 4

[9] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *ArXiv*, abs/2106.12052, 2021. 1

[10] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 1, 3