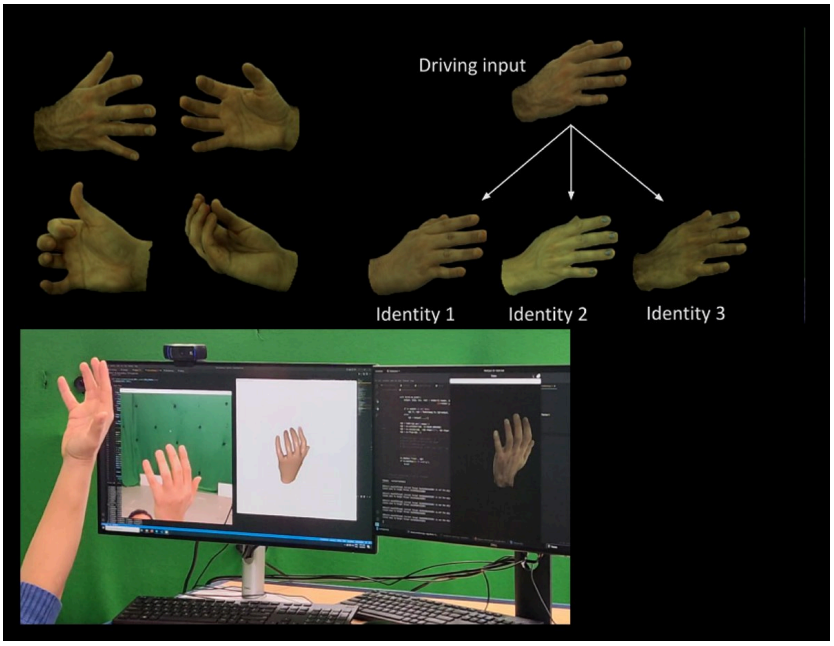
Real-Time, Photorealistic Hands for Neural Environments

Originally published at https://arquivo.pt/wayback/20240123235118oe_/https://blog.metaphysic.ai/real-time-photorealistic-hands-for-neural-environments/

February 16, 2023

New research from the Max Planck Institute, together with Saarland University in Saarbrücken, Germany, offers a method for interpreting a user's hand movements in real time, using Neural Radiance Fields (NeRF) and a range of other technologies.



Real-time reconstruction of a user's hands, captured from a webcam and reconstructed. Source: https://arxiv.org/pdf/2302.07672.pdf

The work is intended to advance the state of the art in the depiction of hands in video-games, VR/AR environments, videoconferencing, and other scenarios where low or zero-latency neural hand representation has proved a challenge to date.

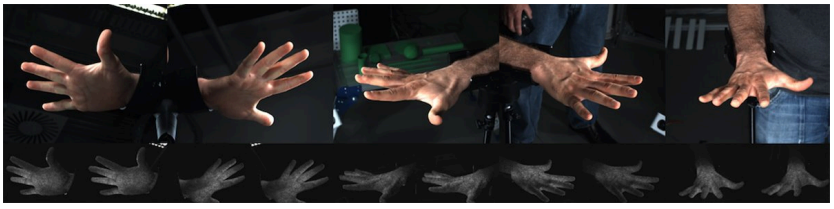
The method, called *LiveHand*, outperforms the nearest similar systems, and is the first of its kind to be able to recreate a user's hands accurately and in real time, achieving a latency of 30ms and an effective frame rate of 33fps.

Though a live demo is referenced in the paper, it has not been made available yet, though the researchers promise that they will release the code in time.

The new paper is titled *LiveHand: Real-time and Photorealistic Neural Hand Rendering*, and comes from six Max Planck researchers, two affiliated with Saarland.

Approach

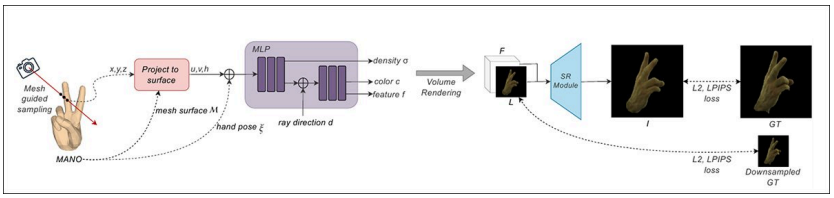
LiveHand is built partly from the MANO project, a prior Max Planck collaboration, with Body Labs Inc., which aimed to incorporate hands in the context of total-body reconstruction – a challenge that's only come into focus in recent times.



Collecting data for the 3DmdHand system, for the prior project Mano. Source: https://arxiv.org/pdf/2201.02610.pdf

The MANO model, now nearly six years old (despite the 2022 date on the latest iteration of the paper), is used as a 'coarse proxy' within LiveHand. In this respect, MANO performs the function of providing parameters for a posed mesh (i.e., a neural representation of a hand in a particular pose), using Linear Blend Skinning (LBS) weights and a canonical hand mesh.

As we can see in the schema below, from the paper, MANO's mesh-guided sampling is then projected onto a surface and mapped to a multi-layer perceptron (MLP) layer, where the ray direction is evaluated, before being passed on to the volume rendering stage.



The workflow for LiveHand.

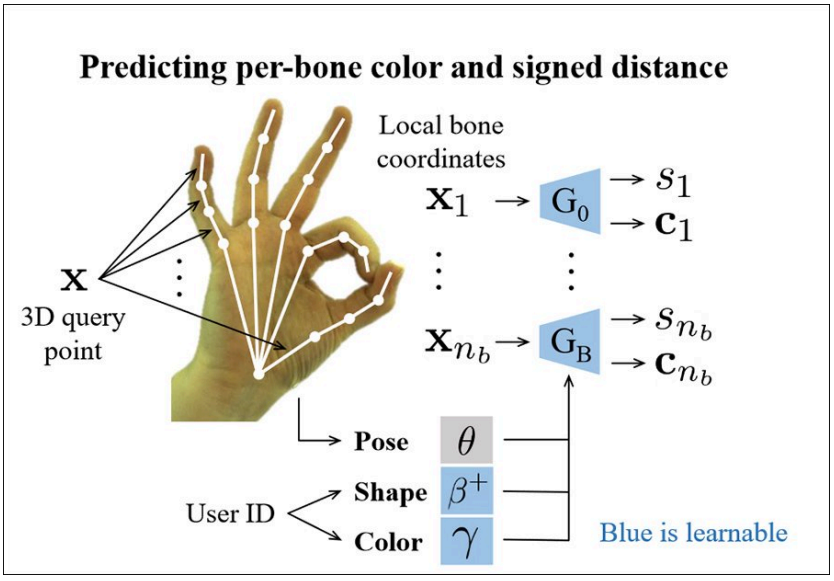
Since the NeRF component in the workflow can only model static representations, the generated radiance field is artificially extended to accommodate deformations (i.e., to allow the 'canonical' hand to be re-shaped to represent movement and new poses).

Sticking to Canon

Canonicalization is key to the fast performance of LiveHand; in neural image synthesis, a canonical reference is a 'default' or base position for the geometry, similar to Da Vinci's Vitruvian Man – a kind of teleological ideal from which diversions and distortions can be added to represent poses.

Thus, in the left part of the schema image above, the canonical pose is distorted to the extent that the outermost two fingers and the thumb are extended and moved to form a hand gesture, whereas the immobile central palm and the two main fingers remain in 'canonical' (i.e., default) disposition.

Prior approaches have used per-bone canonicalization, where the coordinates of the scene in world space are transposed into the local coordinate systems of each individual bone (a global>local coordinate mapping that will be familiar to CGI practitioners), and the implicit fields learned from the resulting transposition.



Per-bone coordinate mapping in LISA. Source: https://arxiv.org/pdf/2204.01695.pdf

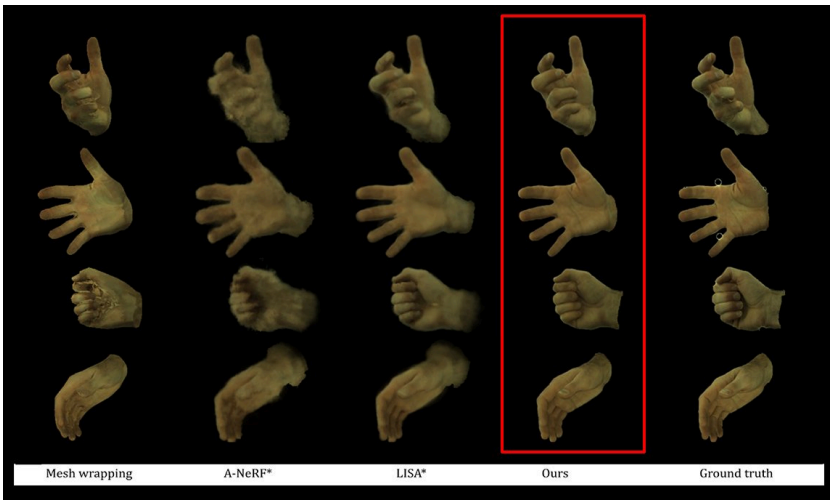
However, this method, used in the prior LISA hand reconstruction system, does not operate quickly enough at inference time. Therefore the researchers adopt a mesh-based approach first used in the NeRF-based project Neural Actor, and obtain a canonical representation using the texture space of a mesh surface.

What this means is that the texture obtained from the source data (in the yet-to-be-revealed demo, that's the user's hands being transmitted through the webcam) is imposed on the nearest available points in the relatively crude MANO hand, to obtain its relative or canonical coordinates.

The authors state:

'This allows us to canonicalize the world coordinates to a representation that stays consistent with respect to hand surface irrespective of hand pose ξ, thus, preventing the dispersion of learned features in the input space.'

Despite this reliance on the coarse mesh provided by MANO, this method allows for the rendition of fine details that MANO's native workflow cannot accommodate, allowing LiveHand to outperform baselines for this task.



Indicated in the red-highlighted column, LiveHand's renderings are a qualitative improvement on prior approaches (see 'Experiments' below).

In the course of processing incoming frames, LiveHand evaluates the depth maps produced by MANO, thus removing the need for an additional MLP layer for this task, which would increase latency.

LiveHand would still, in ordinary circumstances, be unable to achieve real-time rendering under these conditions; therefore the authors added a super-resolution network (first presented by Stanford and NVIDIA in 2020) capable of upscaling the low-resolution output from the native architecture.

Hand Editing

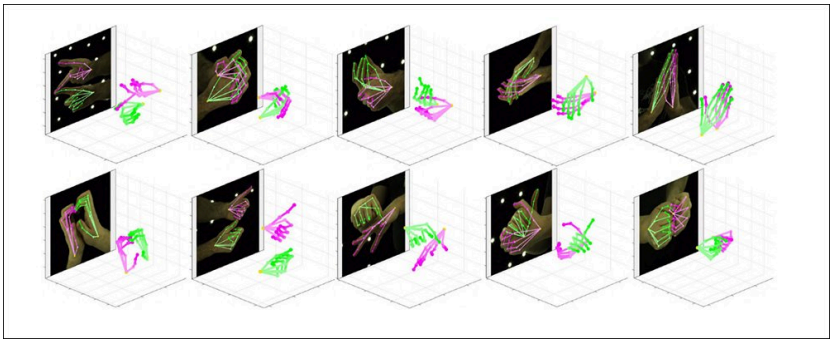
Additional to rendering, LiveHand also allows for re-modeling of hands, due to the way that the workflow integrates and evaluates the hand mesh during inference. Once the original hand parameters are available, the shape can be modified with any corresponding mesh.

The values normally passed to this process come from user-input (i.e., via the web-cam), but there is nothing to stop the method also being used to apply persistent deformations that can accompany the automated deformations, thereby offering a simplified modeling pipeline.

Hand geometry can be adapted at will, with no further retraining of the model necessary.

Experiments

To test the system, the researchers used the public release of the InterHand2.6m benchmark dataset, which contains multiple view images of various people performing a broad range of actions, in videos running at 5fps, and at 512x334px resolution.



Examples from the InterHand2.6m dataset. Source: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123650545.pdf

The right-hand sequences from four users featured in four subsets from the dataset were used in training, with the final 50 frames from each held back as a test set. For quantitative evaluation, four metrics were used: Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID).

Finding an apposite challenger was itself a challenge, since the LISA system, the nearest in terms of functionality to LiveHand, was trained on a version of InterHand2.6m that is not publicly available. Therefore the authors essentially recreated the original framework.

They also tested LiveHand against the body-modeling architecture of A-NeRF, a 2021 offering from the University of British Columbia and Facebook's Reality Labs

	PSNR ↑	SSIM ↑	LPIPS(x1000) ↓	FID ↓	FPS ↑
Mesh wrapping	28.28	0.94	49.44	298.28	**82.33**
A-NeRF* [Su et al. 2021b]	28.07	0.76	94.41	318.61	0.83
LISA* [Corona et al. 2022]	29.36	0.82	78.46	255.43	3.70
Ours	**32.04**	**0.95**	**25.73**	**197.39**	45.45

Results from qualitative testing. The results also include a strand for mesh-wrapping, wherein the texture was extracted from a base canonical hand pose and wrapped onto the target poses being tested.

Of the results, the authors comment:

'[Our] method outperforms all other neural implicit baselines while being much faster. These improvements in the metrics also translate to significant improvements in perceptual quality on the test set…'

The researchers attribute much of the improvement their system makes over prior works to the use of perceptual loss during the training of the network, versus the per-pixel loss utilized for A-NeRF and LISA.

They conclude:

'Note that modern graphics pipelines can achieve much higher frame rates for mesh rendering based on their implementation, and we only benchmark ours. However, by no means [can] such a simple rendering [method] achieve the complex appearance effects and photorealism [that] our method can.

'This demonstrates that our model can learn improvements upon what is possible using only the coarse geometric initialization.'