Hand Avatar:
Free-Pose Hand Animation and Rendering from Monocular Video

Xingyu Chen · Baoyuan Wang · Heung-Yeung Shum
Xiaobing.AI

Paper ID 3630
Presentation Date: June 21, 2023
Tag: WED-AM-044

Overview

Contribution

- We propose a HandAvatar framework, the first method for neural hand rendering with self-occluded illumination
- We develop MANO-HD and a local-pair occupancy field (PariOF) that fit hand geometry with personalized shape
- We propose a self-occlusion-aware shading field (SelF) that can render hand texture with faithful shadow patterns
- Our framework is end-to-end developed for free-pose realistic hand avatars. Extensive evaluations indicate our method outperforms prior arts by a large margin

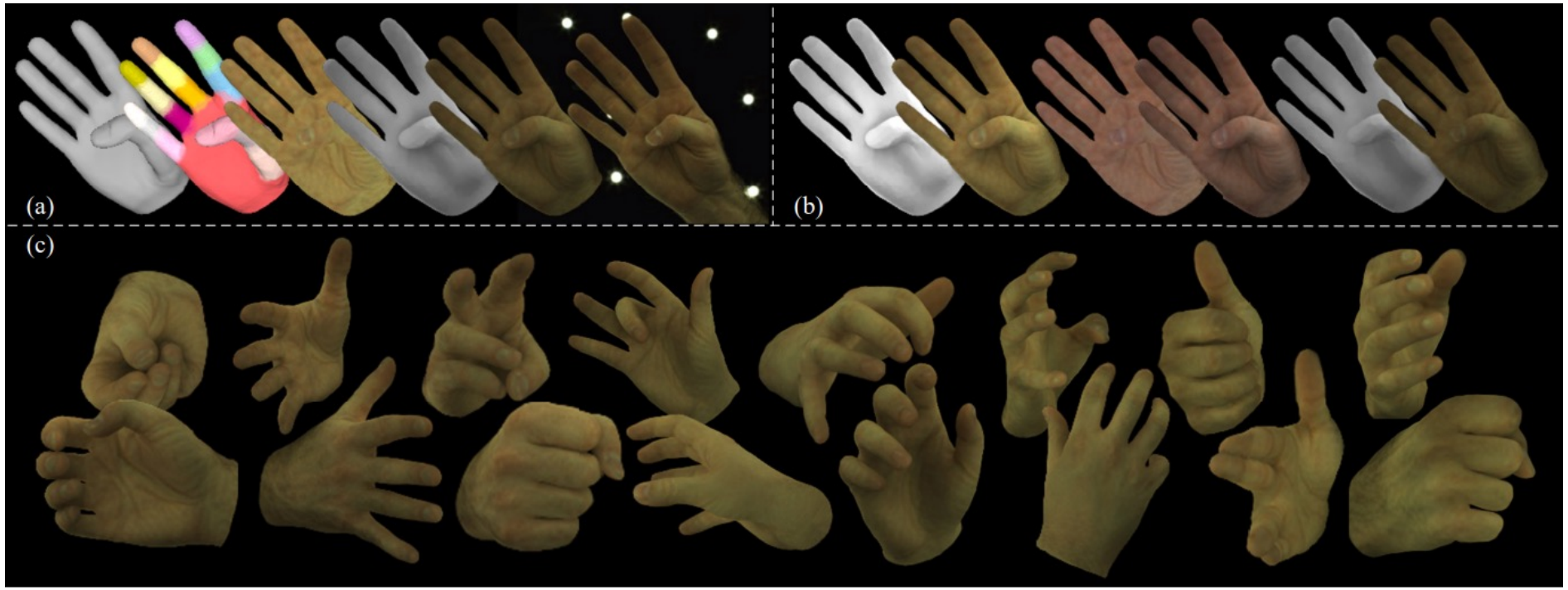

Figure 1. Demonstration of HandAvatar. (a) Personalized hand rendering. From left to right: hand mesh, compositional occupancy, albedo, illumination, shaded appearance, and ground truth; (b) three groups of texture editing in terms of lighting, albedo, and shadow (by altering the self-occlusion effect of thumb); (c) free-pose hand animation and rendering.

Overall Framework

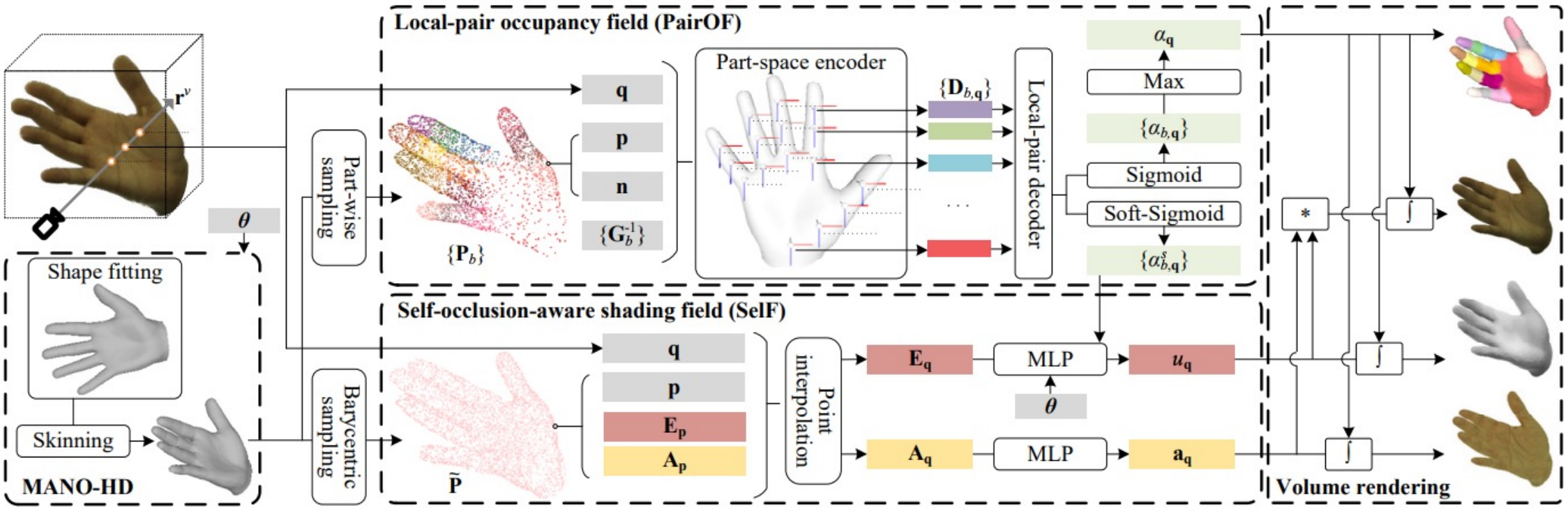

Figure 2. HandAvatar overview. Given hand pose, MANO-HD produces personalized mesh, while PariOF yields accordingly occupancy field. SelF estimates albedo and illumination fields under self-occlusion. Then, hand appearance is synthesized by volume rendering.

$\boldsymbol{\theta}$	hand pose parameter	$\boldsymbol{\beta}$	hand shape parameter
$\mathbf{P}$	point cloud	$\mathbf{r}$	ray direction
b	index of per-bone parts	$\mathbf{n}$	sampled point normal
$\mathbf{q}$	query point position	$\mathbf{p}$	sampled point position
$\mathbf{G}_b$	bone transformation matrix	$\mathbf{D}_b$	part geometry encoding
$\mathbf{E}$	positional encoding	$\mathbf{A}$	albedo encoding
u	illumination value	$\mathbf{a}$	albedo value in RGB
α	occupancy value	α^s	soft occupancy value

Method Details

MANO-HD

- A high-resolution hand mesh template with
  - 12,337 vertices
  - 24,608 faces

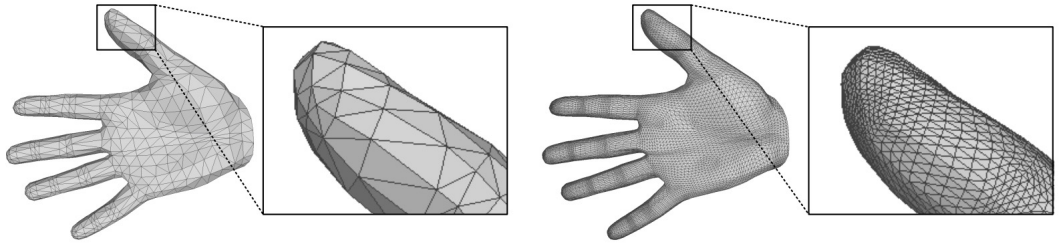

Figure 3. Templates of MANO (left) and MANO-HD (right).

- Optimized skinning weight

Method	*Lap.*	*Cham.*	l_0 norm (%)
MANO	23.31	-	16.29
MANO-HD w/o $\mathbf{W}^{HD}$ opt.	1.923	7.014	17.97
MANO-HD w/o $\mathcal{L}_{l_0}$	1.576	7.170	44.15
MANO-HD	1.753	7.039	16.89

Table 5. Effects of the $\mathbf{W}^{HD}$ optimization (opt.).

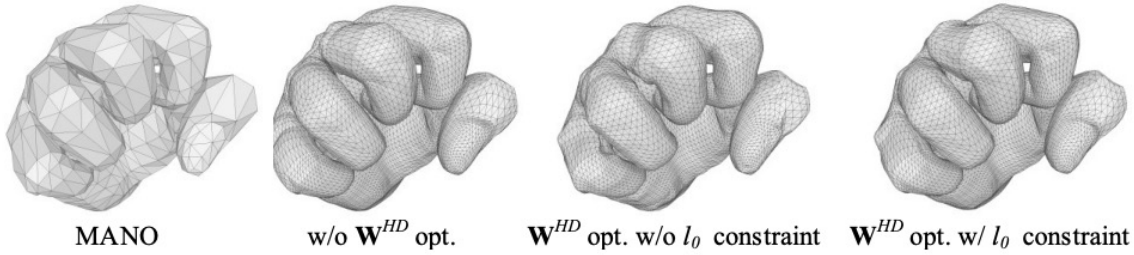

Figure 10. Hand mesh comparison under a large deformation.

- Shape Fitting:
  - An MLP to predict template displacement $\tilde{\mathbf{V}} = \bar{\mathbf{V}} + \mathcal{M}_{shape}([\mathcal{P}(\bar{\mathbf{V}}), \boldsymbol{\theta}])$
  - A silhouette loss to optimize MANO-HD towards personalized shape $\mathcal{L}_{shape} = 1 - \mathbf{IoU}(\mathcal{D}(\mathcal{S}(\tilde{\mathbf{V}})), \mathbf{S}^*)$

Local-Pair Occupancy Field

- Local pair: two parts that are physically connected
- Part-space encoder with a PointNet $\mathbf{D}_{b,\mathbf{q}} = [\mathcal{Q}_{part}([\hat{\mathbf{P}}_b, \hat{\mathbf{N}}_b]), \hat{\mathbf{q}}_b]$
- Local pair decoder with a PointNet $\alpha_{b,\mathbf{q}} = \sigma(\max\{\mathcal{Q}_{pair}(\{\mathbf{D}_{b,\mathbf{q}}, \mathbf{D}_{b',\mathbf{q}}\}) | b' \in \mathbb{P}(b)\})$
- A L2 loss to learn occupancy field $\mathcal{L}_{PairOF} = \frac{1}{N^t} \sum_{\mathbf{q}} (\alpha_{\mathbf{q}} - \alpha^*_{\mathbf{q}})^2$
- Across-part consistency is achieved

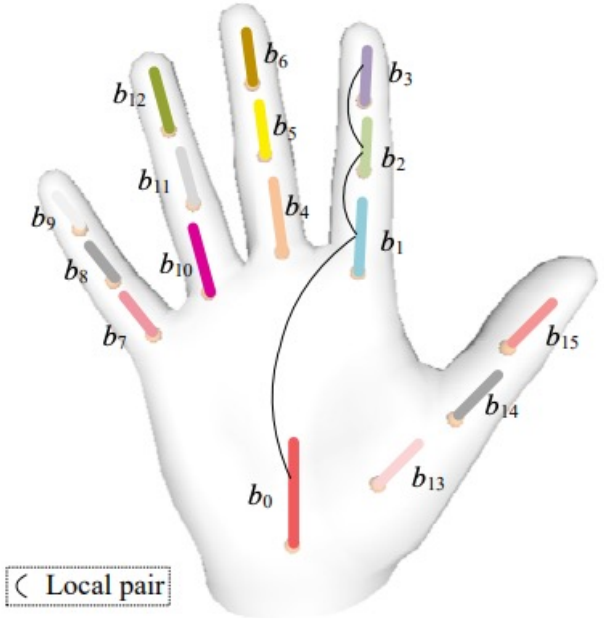

Figure 11. Kinematic tree and part indices of the hand. We also show local pairs along the kinematic chain of forefinger.

Figure 5. Effects of MANO-HD and PariOF. (a) COAP w/ MANO. (b) COAP w/ MANO-HD. (c) PariOF w/ MANO-HD.

Self-Occlusion-Aware Shading Field

- Volume rendering $\mathbf{X}_{\mathbf{r}^v} = \sum_{i=1}^{N^q}(\prod_{j=1}^{i-1}(1-\alpha_{\mathbf{q}_j}))\alpha_{\mathbf{q}_i}\mathbf{X}_{\mathbf{q}_i}$
  - **X** denotes albedo, illumination, or RGB values
- Albedo field
  - Anchors on MANO-HD surface with albedo encodings
  - Point interpolation and an MLP to predict albedo value

$$\mathbf{a_q} = \mathcal{M}_{albedo}(\mathbf{A_q})$$

- Directed soft occupancy to reveal self-occlusion relationship
  - Occupancy with soft sigmoid $\sigma^s(x) = \frac{1}{1+e^{-\tau x}}, \quad 0 < \tau < 1$
  - Accumulate maximal soft occupancy along rays

$$\alpha^s_{b,\mathbf{q},\mathbf{r}} = \max\{\alpha^s_{b,\mathbf{q}_i} | \mathbf{q}_i \leq \mathbf{q}\}$$

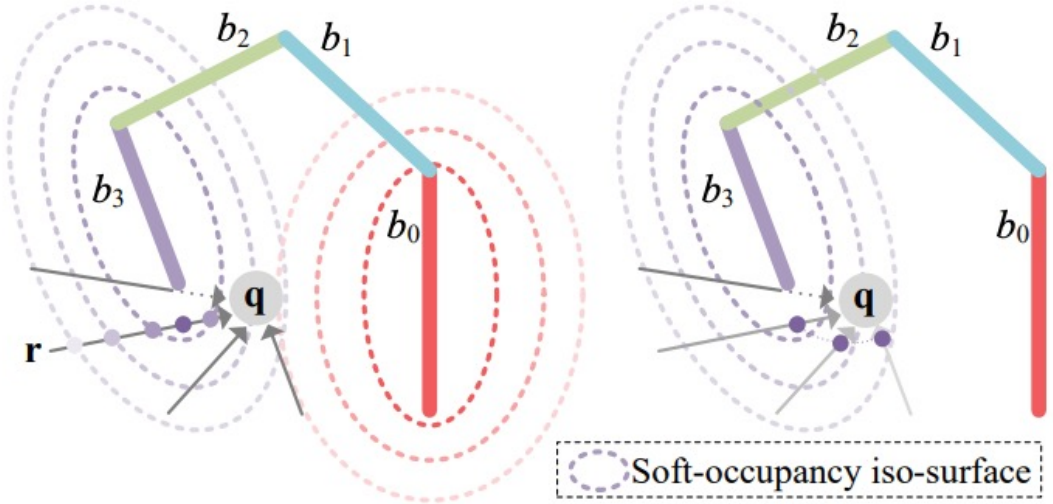

Figure 4. Illustration of self-occluded illumination along the kinematic tree of the forefinger. The deepest purple positions contribute to directed soft occupancy value.

- Illumination Field
  - A MLP to predict illumination value with pose, positional encodings, and directed soft occupancy

$$u_{\mathbf{q}} = \mathcal{M}_{illum}([\boldsymbol{\theta}, \mathbf{E_q}, [\alpha^s_{b,\mathbf{q},\mathbf{r}^v}]_{b=1}^B])$$

- LIPIPS loss and L1 loss to learn from RGB information $\mathcal{L}_{SelF} = \mathcal{L}_{LPIPS}(\mathbf{C}, \mathbf{C}^*) + ||\mathbf{C} - \mathbf{C}^*||_1$

Experiments

Numerical Metrics

Method	*test/Capture0*			*test/Capture1*			*val/Capture0*		
	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
SelfRecon [20]	0.1421	26.38	0.8786	0.1389	25.18	0.8758	0.1490	25.78	0.8687
HumanNeRF [63]	0.1145	27.64	0.8836	0.1177	26.31	0.8803	0.1192	27.80	0.8816
ours	**0.1035**	**28.23**	**0.8941**	**0.1076**	**26.56**	**0.8902**	**0.1062**	**28.04**	**0.8900**

Table 4. Rendering quality comparison among our HandAvatar and prior arts on the InterHand2.6M dataset.

[20] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In CVPR, 2022.

[63] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In CVPR, 2022.

Visualization Comparison

SelfRecon HumanNeRF ours GT

SelfRecon HumanNeRF ours GT

Ablation on SelF

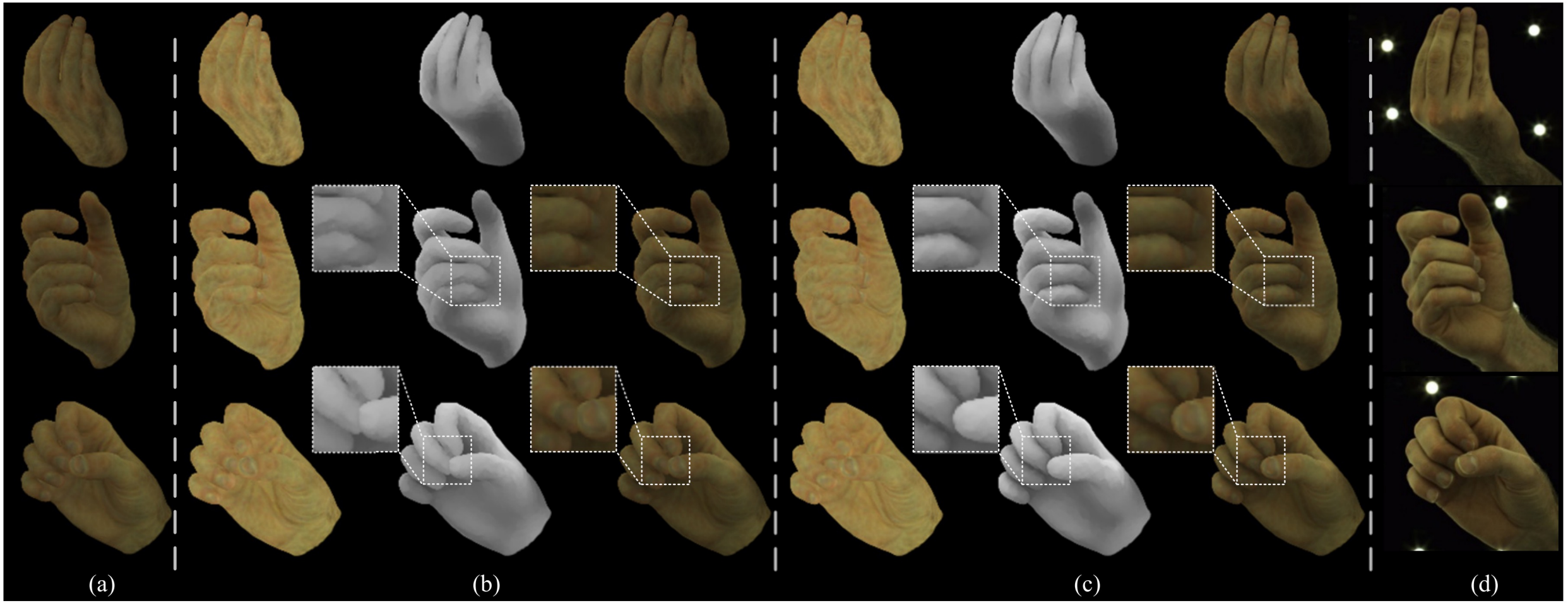

Figure 7. Effects of the disentangled albedo and illumination fields in SelF. (a) Coupled albedo and illumination. (b,c) Disentangled albedo and illumination; directed soft occupancy is not involved in (b); from left to right: albedo, illumination, shaded image. (d) Ground truth.

Thanks

seanchenxy.github. io/HandAvatarWeb