Humans as Light Bulbs: 3D Human Reconstruction from Thermal Reflection

Supplementary Materials

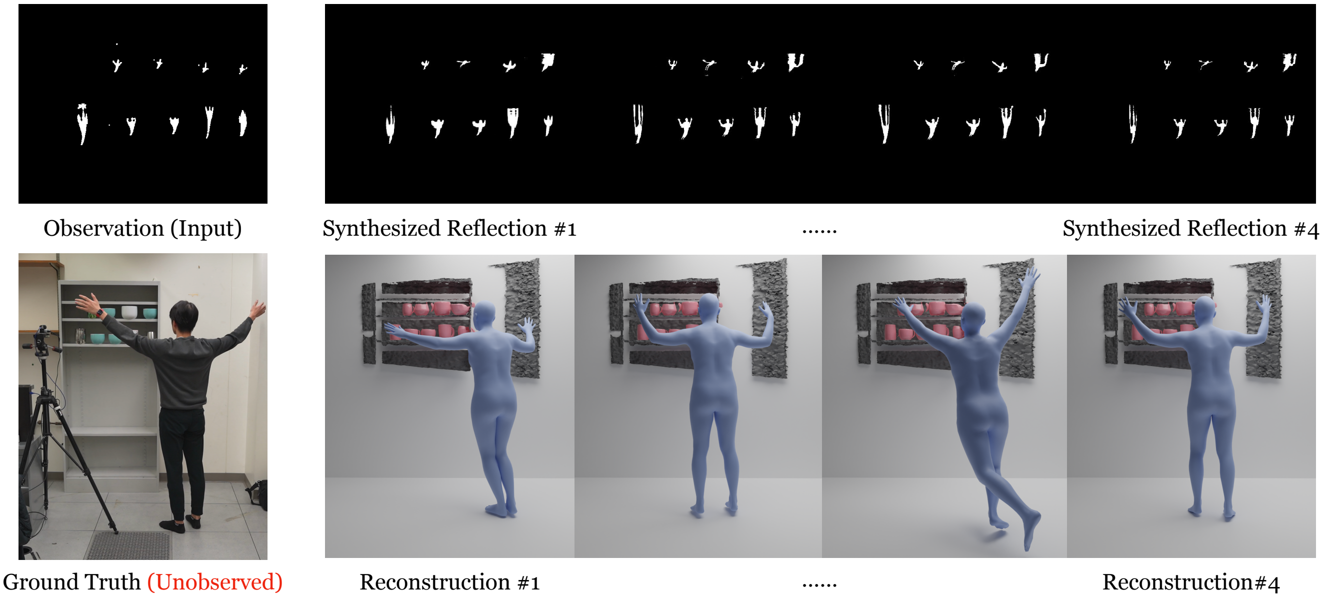
A. Diversity of Reconstruction

Because we adopted an analysis-by-synthesis approach by optimization in the latent space of generative models, our system can generate multiple explanations of the observation by randomly sampling from the generative latent space. In figure 1, we show an example of our system's ability to generate multiple reasonable reconstructions to match the observation – a thermal image. Note that across 4 different runs, our system converges to a similar solution at the end of the optimization, while the reconstruction contains diverse human poses and locations. This indicates the under-constrained nature of the task we are solving.

B. Emissivity of Surfaces

Thermal reflexivity of a surface is a function of the emissivity of the material, which spans 0 to 1. A surface with low emissivity (e.g. polished aluminum) behaves like a mirror under LWIR, in which case almost all of the thermal radiations captured by an LWIR camera are due to reflection. On the opposite, a surface with high emissivity (e.g. human skin) behaves like a black body, where most of the thermal radiations captured by an LWIR camera are emitted from the surface. An emissivity table of common materials can be found in [3].

In this paper, we use bowls and mugs made of ceramics (emissivity $\approx$ 0.85) and stainless steel (emissivity $\approx$ 0.7), all purchased through a mainstream online store. Due to lower emissivity, we observe that the thermal reflection is much stronger with stainless steel than ceramics.


Observation (Input)

Synthesized Reflection #1 ...... Synthesized Reflection #4

Ground Truth (Unobserved)

Reconstruction #1 ...... Reconstruction#4

Figure 1. Diversity of Reconstruction. We repeat the same optimization process multiple times with different randomly sampled latent variables as initialization. While the synthesized reflection image is close to the observation, indicating the optimization is successful, the reconstructed human exhibits diverse poses. This is consistent with the under-constrained nature of the problem we are studying.

C. Equipment

In all experiments, we use a FLIR Boson+ 640 thermal camera core for data collection, which is a high sensitivity, uncooled camera with a spectral range of 8μm - 14μm. The thermal sensitivity / NETD of the camera is 20mK. The resolution is 640x512, with a 50° horizontal field of view. More information about this camera can be found in [1].

The RGBD depth camera we used is an Intel RealSense D415, which uses active stereo for depth estimation. We applied standard post-processing procedures provided by the SDK software of the camera. More information about the camera can be found in [2].

D. Algorithm Implementation Details

D.1. Object Reconstruction

We used pretrained DeepSDF auto-decoder models publicly released by [6]. Optimization of objective function defined in Eq. 7 is performed in the latent space of the auto-decoder with a dimension of 64. For each object present in the scene, we optimize its embedding $\mathbf{z}_{obj}$ with an AdamW optimizer [8] (learning rate = 0.0005), its translation $\mathbf{T}_{obj}$ with an Adam optimizer [7] (learning rate = 0.005), its rotation ϕ_{obj} represented as a quaternions, with an Adam optimizer (learning rate = 0.05), and its scale s_{obj} with an Adam optimizer (learning rate = 0.01).

D.2. Human Reconstruction

For simplicity and ray-tracing efficiency, we used SMPL provided in the SMPL-X repository with $\approx$ 4000 fewer triangles and vertices [10]. Optimization is performed in the input space of the VPoser provided in SMPL-X, with a dimension of 32. We optimize the rotation ϕ_h, translation $\mathbf{T}_h$, pose vector $\mathbf{z}_h$ jointly with an Adam optimizer with a learning rate of 0.1. We perform 1200 steps of optimization on an NVIDIA RTX A6000 GPU with gradient updates every 8 steps (gradient accumulation).

D.3. Differentiable Rendering of Reflections

To calculate the intersection point between a ray and the surface of an SDF, we perform 3 steps of sphere tracing [5]. We calculate the distance matrix $\mathcal{D}$ – distance between each incoming ray $\mathbf{r}_i$ and triangle t_j, using the PyTorch library [9]. A ray-triangle intersection matrix Λ is calculated with the Trimesh library [4]. We sample 400 pixels (rays) from each object present in the scene for optimization.

References

[1] Boson®+, https://www.flir.com/products/boson-plus/?model=22640A050&amp;vertical=lwir&amp;segment=oem. 2

[2] Depth camera d415, https://www.intelrealsense.com/depth-camera-d415/. 2

[3] Emissivity table for infrared thermometer readings, https://ennologic.com/wp-content/uploads/2018/07/Ultimate-Emissivity-Table.pdf. 1

[4] Dawson-Haggerty et al. trimesh. 2

[5] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996. 2

[6] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. *arXiv preprint arXiv:2207.13691*, 2022. 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2

[10] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2