AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction (Supplementary Material)

Zerui Chen[1], Yana Hasson[1,2], Cordelia Schmid[1], and Ivan Laptev[1]

[1] Inria, École normale supérieure, CNRS, PSL Research Univ., 75005 Paris, France
[2] Now at Deepmind
firstname.lastname@inria.fr
https://zerchen.github.io/projects/alignsdf.html

This supplementary material provides additional details for our experimental settings as well as qualitative results of our method. We first present details for our network architecture in Section 1. Section 2 then provides additional implementation details for our training and evaluation procedures. Finally, we present and discuss additional qualitative results in Section 3 and demonstrate video examples in *results_video.mp4*.

1 Network Architecture

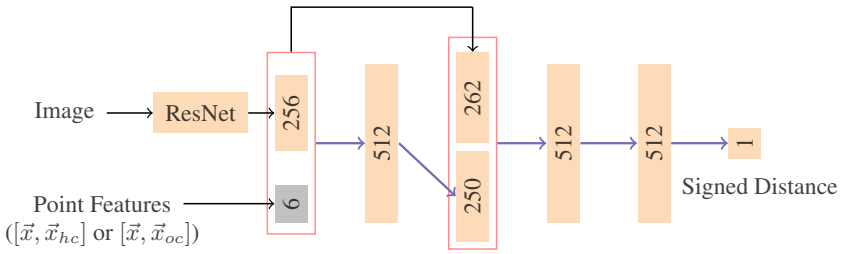


Fig. 1: Network architecture used for our hand and object SDF decoders. Following [6], we also use five fully connected layers (marked in purple) for the SDF decoder. The number in the box denotes the dimension of features. $\vec{x}$ denotes the original 3D coordinate. $\vec{x}_{hc}$ and $\vec{x}_{oc}$ denote the transformed 3D coordinate in the hand and object canonical coordinate system, respectively.

Following previous works [3, 6], we use ResNet-18 [4] as our backbone network. To achieve a fair comparison with the previous method [6], as shown in Figure 1, we also use five fully connected layers to estimate the signed distance from the query point to the hand surface or the object surface. The SDF decoder takes the 256-dimensional image features and 6-dimensional point features as inputs. The image features are extracted from the ResNet-18 backbone. Following Equation 3 and Equation 5 in our paper, we transform the original 3D point $\vec{x}$ into its counterpart $\vec{x}_{hc}$ in the hand canonical coordinate system or its counterpart $\vec{x}_{oc}$ in the object canonical coordinate system.

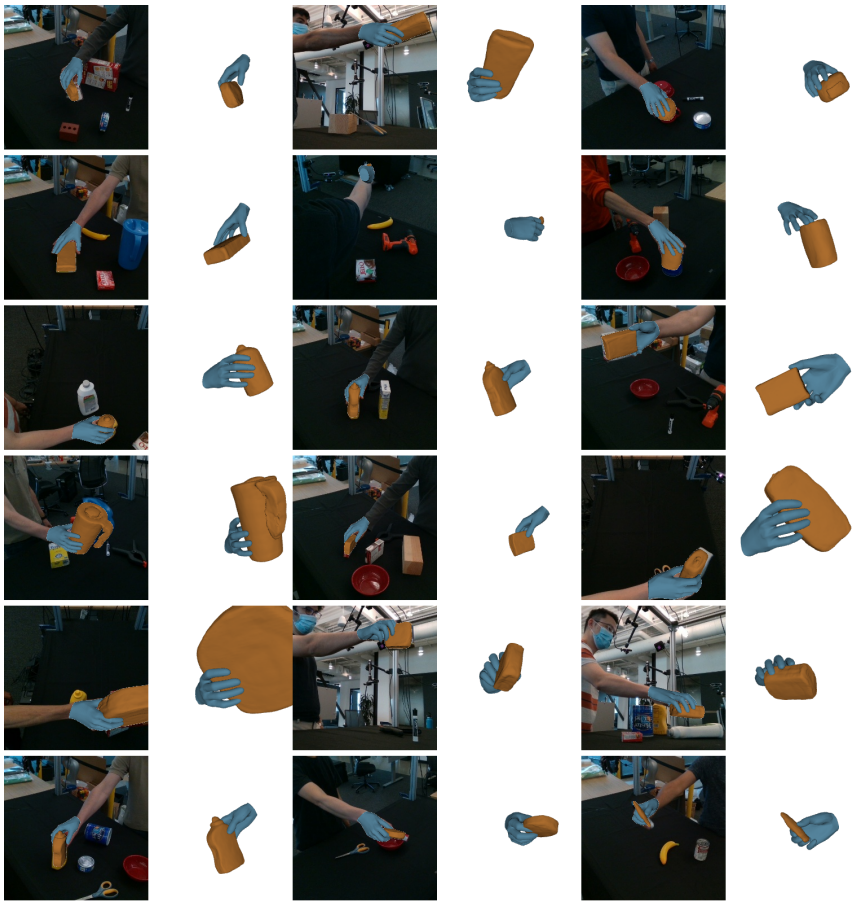


Fig. 2: Qualitative results of our method on the DexYCB [1] benchmark. Our method can also produce realistic 3D reconstruction results for real scenes.

Then, we construct point features by concatenating $\vec{x}$ and $\vec{x}_{hc}$ for the hand SDF decoder or by concatenating $\vec{x}$ and $\vec{x}_{oc}$ for the object SDF decoder.

2 Training and Evaluation

We train all of our models with the following data augmentation. We randomly rotate the input image and 3D points in the camera coordinate system. We empirically find that this data augmentation can boost the performance for 3D reconstruction. We randomly augment training samples via $[-45°, 45°]$ rotation for our experiments on ObMan [3] or $[-15°, 15°]$ rotation for our experiments on DexYCB [1].

We set the hand wrist joint defined by MANO [7] as the origin of our coordinate system. In training, we use a fixed scaling factor to scale all negative points (*i.e.,* points

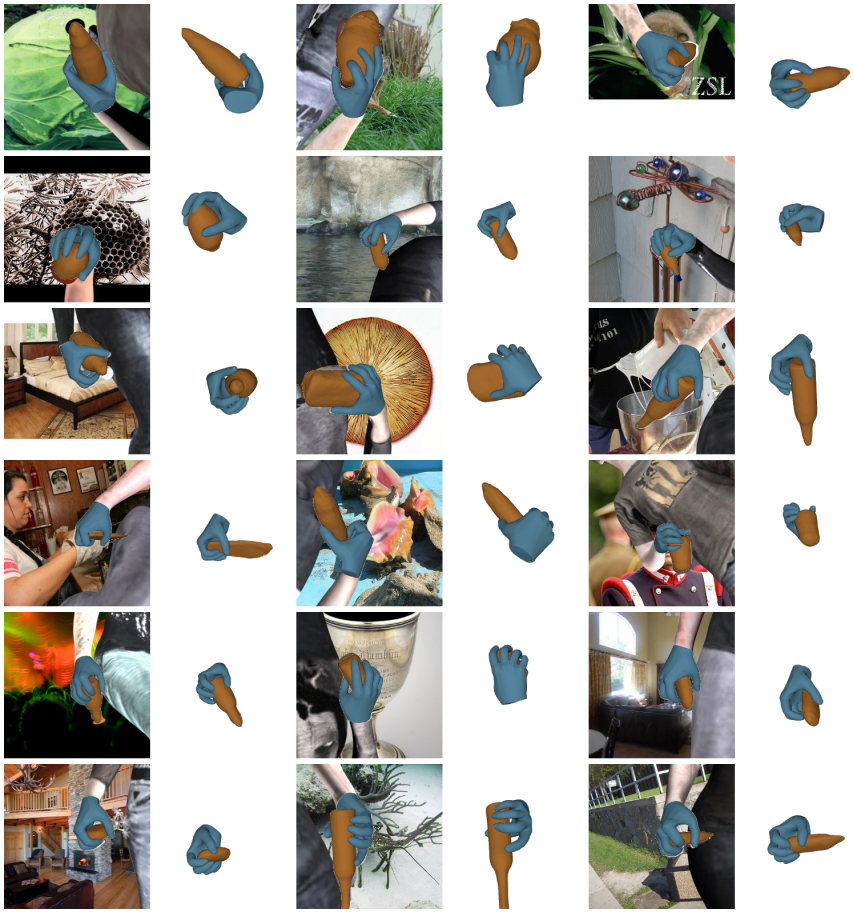


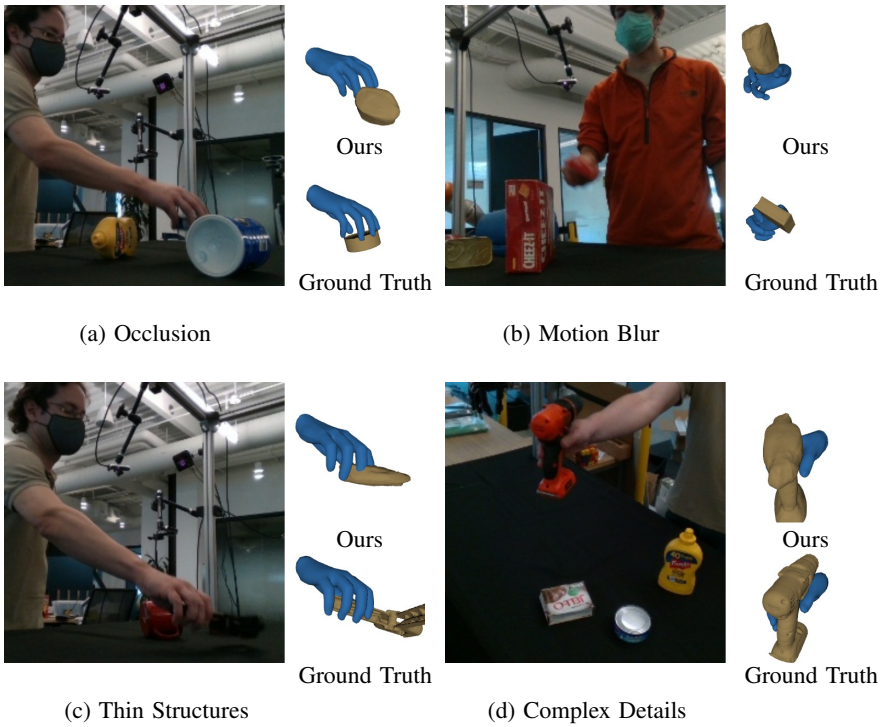
Fig. 3: Qualitative results of our method on the ObMan [3] benchmark. Our method can produce convincing 3D reconstruction results even in cluttered scenes.

that lie in the hand or object mesh) across the dataset within a unit cube. This results in a scaling factor of 7.02 and 6.21 on ObMan and DexYCB, respectively.

To measure the physical quality of our joint reconstruction, we report Contact Ratio (C_r), Penetration Depth (P_d) and Intersection Volume (I_v). We use the trimesh library [2] to detect whether there exists a collision between the hand mesh and the object mesh and compute the max penetration depth between two meshes. We follow the same process as [5, 6] to compute I_v.

3 Qualitative results

We present additional qualitative results on ObMan [3] in Figure 3 and DexYCB [1] in Figure 2. We also study failure cases on DexYCB in Figure 4. From Figure 3, we



(a) Occlusion (b) Motion Blur

(c) Thin Structures (d) Complex Details

Fig. 4: Failure cases of our method on the DexYCB [1] benchmark.

observe that our method can deal with a wide range of objects and recovers detailed interactions between the hand and the object. In Figure 2 we show qualitative results of our method for real images from the DexYCB benchmark. We can see that our method can reconstruct objects of different sizes and often achieve the excellent reconstruction of hands and objects.

While our method advances the state of the art accuracy by a significant margin, it still does not achieve satisfactory performance in some cases. In Figure 4 we show four typical failure cases on DexYCB. As shown in Figure 4(a), when the hand or the object is heavily occluded, our method sometimes cannot make robust predictions. In Figure 4(b), we show that motion blur in input images might also disturb 3D reconstruction results. As shown in Figure 4(c, d), the recovery of thin structures and objects with complex shapes remains challenging. To deal with these issues, future works could leverage the temporary information from videos to filter input noise and gather more details about 3D scenes.

References

[1] Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR (2021)
[2] Dawson-Haggerty et al.: trimesh, https://trimsh.org/
[3] Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
[5] Karunratanakul, K., Spurr, A., Fan, Z., Hilliges, O., Tang, S.: A skeleton-driven neural occupancy representation for articulated hands. In: 3DV (2021)
[6] Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 3DV (2020)
[7] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. TOG (2017)