AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction

Zerui Chen, Yana Hasson, Cordelia Schmid and Ivan Laptev - Inria, ENS, PSL Research University

Motivation

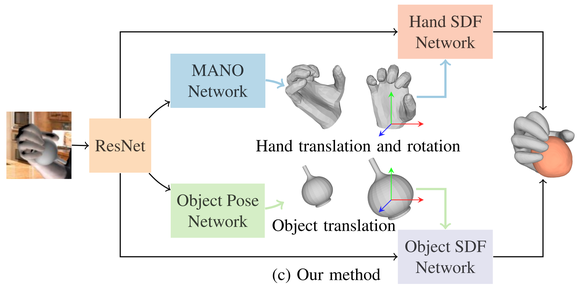
- Though the previous approach [1] embeds strong prior knowledge about the human hand by using the parametric model [3], it can only produce hand and object meshes of limited resolution.
- The recent approach [2] employs signed distance fields (SDFs) to reconstruct surfaces at high resolution but fails to leverage any prior knowledge about hands or objects.

Our contributions:

- We propose to combine the advantages of parametric mesh models and SDFs and present a joint learning framework for 3D reconstruction.
- To effectively embed prior knowledge into SDFs learning, we propose to disentangle the pose learning from the shape learning for this task.

(a) Hasson *et al.* [1]

(b) Karunratanakul *et al.* [2]



(c) Our method

Hand Pose Estimation

Our model can be generally split into two parts: the pose estimation part and the shape reconstruction part. The shape reconstruction part employs pose estimation results to normalize SDFs prediction into their canonical frames.

Hand Pose Estimation: we employ a parametric hand mesh model, MANO [3], to capture the kinematics for the human hand. In implementation, we integrate MANO as a differentiable layer into our model and use it to predict the hand vertices ($\vec{v}_h$), the hand joints ($\vec{j}_h$) and hand poses ($\vec{\theta}_h$). During training, we define the supervision on the joint locations ($L_{\vec{j}_h}$), the shape parameters ($L_{\vec{\beta}_h}$) and the predicted hand poses ($L_{\vec{\theta}_h}$):

$$L_{hand} = \lambda_{\vec{j}_h} L_{\vec{j}_h} + \lambda_{\vec{\beta}_h} L_{\vec{\beta}_h} + \lambda_{\vec{\theta}_h} L_{\vec{\theta}_h}, \tag{1}$$

where we set $\lambda_{\vec{j}_h}$, $\lambda_{\vec{\beta}_h}$ and $\lambda_{\vec{\theta}_h}$ to 5×10^{-1}, 5×10^{-7} and 5×10^{-5}, respectively.

Object Pose Estimation

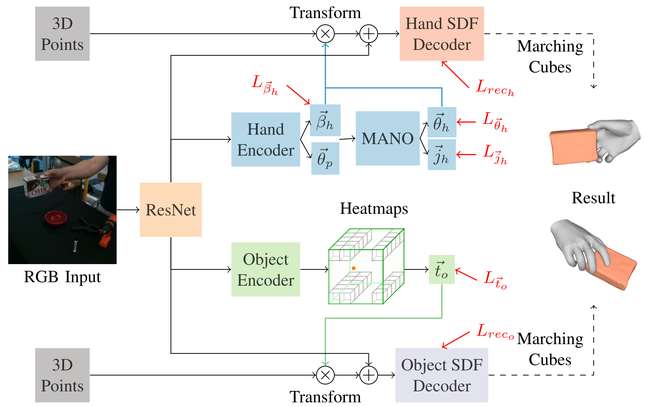
We set the origin of our coordinate system as the wrist joint defined in MANO. We employ the volumetric heatmap to predict per voxel likelihood for the object centroid and extract its 3D location from the heatmap differentiably. During training, we optimize network parameters by minimizing the L2 loss between the estimated 3D object translation $\vec{t}_o$ and its corresponding ground truth. The resulting loss L_{obj} is:

$$L_{obj} = \lambda_{\vec{t}_o} L_{\vec{t}_o}, \tag{2}$$

where we empirically set $\lambda_{\vec{t}_o}$ to 5×10^{-1}.

Hand-object Shape Reconstruction

Our approach makes an attempt to disentangle the shape learning and the pose learning.



By estimating the hand pose, we could obtain the global rotation ($\vec{\theta}_{hr}$) and its rotation center ($\vec{t}_h$) defined by MANO. Using the estimated $\vec{\theta}_{hr}$ and $\vec{t}_h$, we transform any sampled 3D point $\vec{x}$ to the canonical hand pose (*i.e.*, the global rotation equals to zero):

$$\vec{x}_{hc} = \exp(\vec{\theta}_{hr})^{-1}(\vec{x} - \vec{t}_h) + \vec{t}_h, \tag{3}$$

where $\exp(\cdot)$ denotes the transformation from the axis-angle representation to the rotation matrix using the *Rodrigues formula*. Then, we concatenate $\vec{x}$ and $\vec{x}_{hc}$ to predict its signed distance to the hand. Similarly, by estimating the object pose, we obtain the object translation $\vec{t}_o$ and transform $\vec{x}$ to the canonical object pose:

$$\vec{x}_{oc} = \vec{x} - \vec{t}_o. \tag{4}$$

Then, we concatenate $\vec{x}$ and $\vec{x}_{oc}$ and feed them to the object SDF decoder and predict its signed distance to the object.

Evaluations

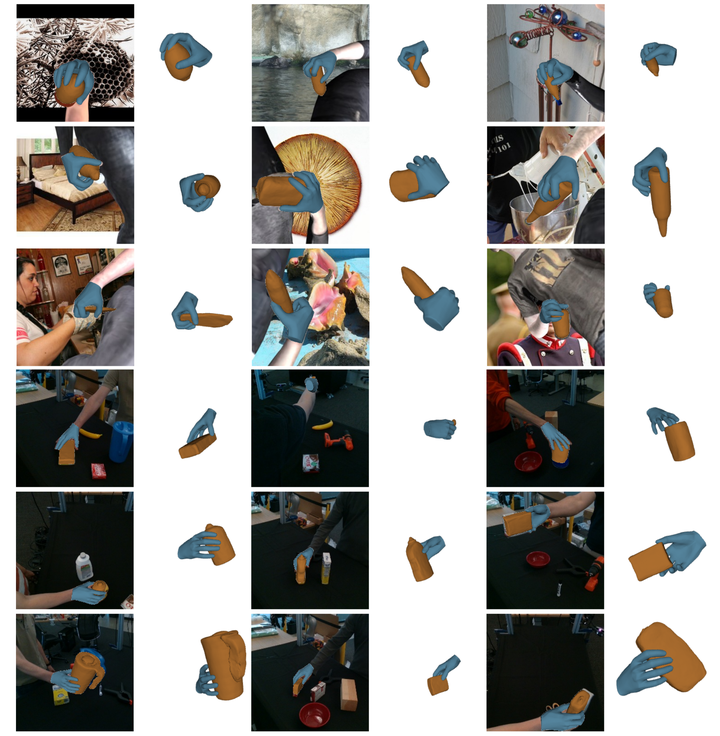
Our approach achieves state-of-the-art hand-object shape reconstruction performance on Obman and DexYCB benchmarks.

Comparison with previous state-of-the-art methods on ObMan.

Methods	H_{se} ↓	H_{ve} ↓	O_{se} ↓	H_{je} ↓	O_{te} ↓	C_r	P_d	I_v
Hasson *et al.* [1]	0.415	0.383	3.60	**1.13**	-	94.8%	1.20	6.25
Karunratanakul *et al.* [2]-1De	0.261	0.246	6.80	-	-	5.63%	0.00	0.00
Karunratanakul *et al.* [2]-2De	0.237	-	5.70	-	-	69.6%	0.23	0.20
Ours	**0.136**	**0.121**	**3.38**	1.27	**3.29**	95.5%	0.66	2.81

Comparison with previous state-of-the-art methods on DexYCB.

Method	H_{se} ↓	H_{ve} ↓	O_{se} ↓	H_{je} ↓	O_{te} ↓	C_r	P_d	I_v
Hasson *et al.* [1]	0.785	0.594	4.4	2.0	-	95.8%	1.32	7.67
Karunratanakul *et al.* [2]	0.741	0.532	5.8	-	-	96.7%	0.83	1.34
Ours	**0.523**	**0.375**	**3.5**	**1.9**	**2.7**	96.1%	0.71	3.45



References

[1] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[2] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang. Grasping Field: Learning implicit representations for human grasps. In *3DV*, 2020.

[3] J. Romero, D. Tzionas, and M. J. Black. Embodied Hands: Modeling and capturing hands and bodies together. *TOG*, 2017.