MOHO: Learning Single-view Hand-held Object Reconstruction with Multi-view Occlusion-Aware Supervision

[1]Chenyangguang (Cyrus) Zhang* [1]Guanlong Jiao* [2]Yan Di* [1]Gu Wang [1]Ziqin Huang
[1]Ruida Zhang [3]Fabian Manhardt [1]Bowen Fu [2,3]Federico Tombari [1]Xiangyang Ji
[1]Tsinghua University [2]Technical University of Munich [3]Google

CVPR JUNE 17-21, 2024 SEATTLE, WA

Task Definition

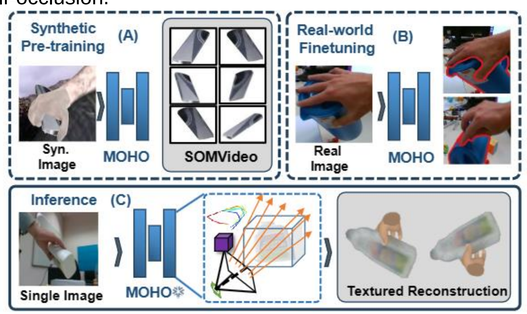
[Hand-held Object Reconstruction]

- Given a single RGB image, DDF-HO predicts a 3D model for the object grasped by the hand. It is an essential technique with many practical applications, e.g. robotics, augmented and virtual reality, medical imaging.

Motivation

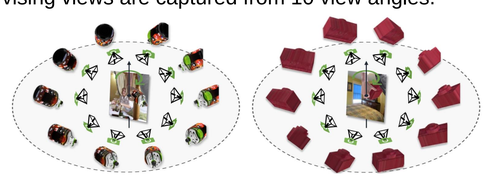
- **[Current Setting]** Current single-view hand-held object reconstruction methods typically need 3D ground-truth models for supervision, which are hard to collect in real world.

- **[Our method]** We present **MOHO**, to exploit **M**ulti-view **O**cclusion-aware supervision from hand-object videos for **H**and-held **O**bject reconstruction from a single image, tackling two predominant challenges in such setting: hand-induced occlusion and object's self-occlusion.



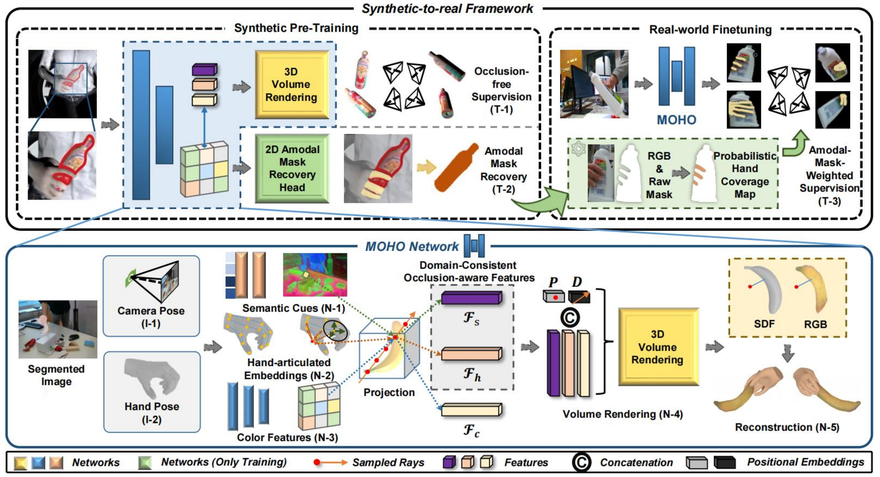
SOMVideo Dataset

SOMVideo is synthesized with 141,550 scenes with the same 2,772 objects as the ObMan dataset. Each corresponding occlusion-free supervising views are captured from 10 view angles.



Method

MOHO Pipeline



Domain-consistent Occlusion-aware Features

Semantic cues by DINO and hand-articulated geometric embeddings: for more stable knowledge transferring in the whole synthetic-to-real pipeline.

Synthetic-to-real Training

$$\mathcal{L} = \mathcal{L}_{color} + \lambda_1 \mathcal{L}_{eik} + \lambda_2 \mathcal{L}_{mask} + \lambda_3 \mathcal{L}_{n_{ori}} + \lambda_4 \mathcal{L}_{n_{smo}}$$

$$\mathcal{L}_{amw} = BCE(\hat{M}_I^{ho} \oplus M_I, \hat{O}_I) \quad \mathcal{L}_{mask} = BCE(M_I^{co}, \hat{O}_I)$$

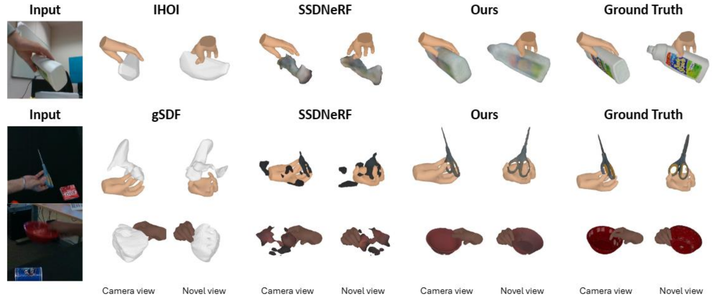
$$\mathcal{L}_{color} = \frac{1}{m}\sum_i |\hat{c}_i - c_i|$$

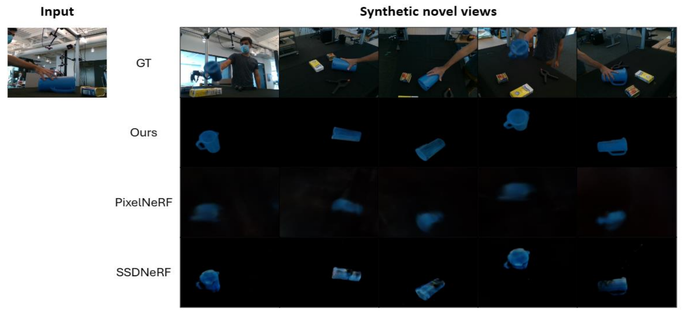
$$\mathcal{L}_{eik} = \frac{1}{nm}\sum_{i,j}(||\nabla \psi_S(P_{i,j})||_2 - 1)^2$$

$$\mathcal{L}_{n_{ori}} = \frac{1}{m}\sum_i (min(0, -\hat{n}_i \cdot D_i))^2 \qquad \mathcal{L}_{n_{smo}} = \frac{1}{K}\sum_k (\hat{n}_k - \overline{\hat{n}_k})^2$$

Experiments

Main Results





Method	F-5 ↑	F-10 ↑	CD ↓
HO [23]	0.11	0.22	4.19
GF [28]	0.12	0.24	4.96
IHOI [63]	0.28	**0.50**	1.53
PixelNeRF [65]	0.17	0.32	6.91
SSDNeRF [9]	0.25	0.40	2.60
Ours	**0.31**	**0.50**	**0.91**

Table 1. Geometric results on HO3D [22] compared with 3D supervised methods (top) and 2D supervised methods (bottom).

Method	F-5 ↑	F-10 ↑	CD ↓
HO [23]	0.38	0.64	0.42
GF [28]	0.39	0.66	0.45
AlignSDF [12]	0.41	0.68	0.39
gSDF [13]	0.44	0.71	0.34
PixelNeRF [65]	0.25	0.46	0.94
SSDNeRF [9]	0.27	0.49	0.58
Ours w/o SYN	0.52	0.74	0.18
Ours	**0.60**	**0.81**	**0.15**

Table 2. Geometric results on DexYCB [8] compared with 3D supervised methods (top) and 2D supervised methods (bottom).

Code & Data

Paper

Cyrus Webpage

Cyrus WeChat