MagicHOI: Leveraging 3D Priors for Accurate Hand-object Reconstruction from Short Monocular Video Clips

1. Method

1.1. Space alignment

As described in the main paper, the goal of space alignment is to optimize the rotation $\mathbf{R}_{\text{a}} \in SO(3)$ and translation $\mathbf{t}_{\text{a}} \in \mathbb{R}^3$ to register the object coordinate space with the novel view synthesis (NVS) model space using the following loss terms:

3D correspondence term: This term penalizes the distance between corresponding 3D points $\{\mathbf{P}^{\text{ref}}, \tilde{\mathbf{P}}^{\text{ref}}\}$ so that they coincide in the NVS model space:

$$\mathcal{L}_{3D} = \sum_{j=1}^{N} \rho\big(\|\mathbf{R}_{\text{a}}\mathbf{P}_j^{\text{ref}} + \mathbf{t}_{\text{a}} - \tilde{\mathbf{P}}_j^{\text{ref}}\|^2\big). \tag{1}$$

where ρ is the Huber loss, which lessens the impact of outliers.

Perspective-n-Point (PnP) term: This term minimizes the projection error so that the 2D projections of the transformed 3D points $\{\tilde{\mathbf{P}}^{\text{ref}}\}$ match their 2D correspondences $\{\mathbf{p}^{\text{ref}}\}$ in the NVS reference image:

$$\mathcal{L}_{2D} = \sum_{j=1}^{N} \rho\Big(\Big\|\pi\left(\mathbf{K}_i(\mathbf{R}_{\text{a}}\tilde{\mathbf{P}}_j^{\text{ref}} + \mathbf{t}_{\text{a}})\right) - \mathbf{p}_j^{\text{ref}}\Big\|^2\Big). \tag{2}$$

where $\pi(\mathbf{x}) = \left[\frac{x_1}{x_3}, \frac{x_2}{x_3}\right]^\top$ is the perspective projection.

1.2. Regularization of observed object regions

As described in the main paper, object regularization under limited observed views aims to optimize the object parameters ψ_o to recover object accurate shape and texture.

Similar to HOLD [2], we incorporate visual observations by adding an RGB loss, $\mathcal{L}_{\text{RGB}}$, defined as the L1 distance between each rendered pixel and its corresponding observed pixel. To encourage consistency with the object's segmentation mask, we enforce segmentation loss $\mathcal{L}_{\text{segm}}$, computed between the rendered mask and the ground-truth mask obtained from the video-segmentation model Cutie[1].

In contrast to HOLD [2], we regularize the geometry to enforce surface smoothness through the following constraint:

$$\mathcal{L}_{\text{smooth}} = \sum_i \|\mathbf{n}_i - \mathbf{n}_j\|^2 \tag{3}$$

where $\mathbf{n}_i$ and $\mathbf{n}_j$ are normal vectors at neighboring pixels i and j.

1.3. Object model training

We represent the object using an efficient hash grid with neural SDF rendering. The 3D implicit SDF representation is trained for 3000 iterations, taking approximately 25 minutes on an RTX 4090.

During the first 1,000 iterations, we apply only $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{segm}}$, and $\mathcal{L}_{\text{smooth}}$, supervising them solely with the observed images with a batch size of 1. From iteration 1,000 to 3,000, $\mathcal{L}_{\text{NVS}}$ takes effect, and $\mathcal{L}_{\text{RGB}}$, $\mathcal{L}_{\text{segm}}$, and $\mathcal{L}_{\text{smooth}}$ are supervised by both the observed and reference images with the batch size of 1. At iteration 2000, the visibility grid is determined. From iteration 2000, the weighting factor μ takes effect.

To enhance training efficiency and stability, we follow Magic3D [3] for computing the $\mathcal{L}_{\text{NVS}}$ loss on novel views. Training starts at a resolution of 64×64 pixels for the first 400 iterations with a batch size of 8, increases to 128×128 pixels from 400 to 700 iterations with a batch size of 4, and reaches 256×256 pixels from 1000 to 2000 iterations with a batch size of 2.

Novel views are sampled in spherical coordinates with an elevation range of $[-30°, 30°]$, an azimuth range of $[-180°, 180°]$, a fixed distance of 2 meters, and a vertical field of view of 41.5°. All pixels in each image are sampled at every iteration.

1.4. Hand-object alignment

As described in the main paper, we optimize the hand translation $\mathbf{t}_h$ and global scale s to achieve accurate hand-object alignment.

Similar to HOLD [2], we enforce consistency between the ground-truth (GT) and predicted hand-joint projections.

In contrast to HOLD [2] which treats every fingertip as a potential hand–object contact point, we keep only high-confidence visible contact vertices, excluding

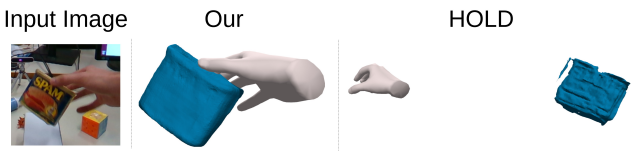

Figure 1. **HOLD[2] large deviation between hand and object after hand-object pose refinement**. The lack of reconstruction in occluded object regions leads to unreliable contact, resulting in significant hand-object deviation.

those occluded by either the object or the hand itself. Hand–object alignment is then encouraged by enforcing contact between $\mathcal{V}_\text{h}$ and $\mathcal{V}_\text{o}$, yielding physically plausible interactions. The loss term is defined as:

$$\mathcal{L}_\text{contact} = \sum_i \left\| \mathcal{V}_\text{h}^i - \mathcal{V}_\text{o}^i \right\|. \tag{4}$$

In addition, we enforce temporal smoothness of the hand vertices between consecutive frames with the smoothness constraint $\mathcal{L}_\text{vsmooth}$:

$$\mathcal{L}_\text{vsmooth} = \sum_{i=1}^{M} \left\| \mathbf{V}_t^i - \mathbf{V}_{t+1}^i \right\|_2^2 \tag{5}$$

where M is the number of hand vertices, $\mathbf{V}_t^i$ and $\mathbf{V}_{t+1}^i$ are the positions of the i-th vertex at frames t and $t+1$, respectively.

Meanwhile, we prevent hand–object interpenetration by introducing a penetration constraint $\mathcal{L}_\text{penetr}$:

$$\mathcal{L}_\text{penetr} = \frac{1}{N} \sum_{v \in \mathcal{H}} \max(0, -d(v)), \tag{6}$$

where $d(\cdot)$ is the query SDF value from the object implicit function f_{ψ_o}, $\mathcal{H}$ represents the hand mesh, and v is an arbitrary vertex of the hand mesh.

2. Experiment Details

2.1. Short sequence selection

All short sequences are extracted from the long HO3D sequences used in HOLD. Each long sequence is first divided into clips of 30 consecutive frames. We then run HLoc [5, 6] on every clip to estimate the camera poses; some frames, however, yield invalid poses because of low texture. We keep the first clip that contains all 30 valid poses. If no clip satisfies this requirement, we retain the clip with the largest number of valid poses and we simply discard the invalid frames. The resulting set of selected sequences is summarized in Table 1.

Object	Sequence name	Frames
bleach	ABF12	180 ∼ 209
bleach	ABF14	180 ∼ 209
potted meat	GPMF12	90 ∼ 119
potted meat	GPMF14	90 ∼ 119
cracker box	MC1	0 ∼ 29
cracker box	MC4	0 ∼ 29
power drill	MDF12	60 ∼ 89
power drill	MDF14	300 ∼ 329
sugar box	ShSu10	30 ∼ 59
sugar box	ShSu12	30 ∼ 59
mustard	SM2	90 ∼ 112
mustard	SM4	00 ∼ 26
mug	SMu1	00 ∼ 29
mug	SMu40	14 ∼ 28

Table 1. HO3D sequences for our method and baseline methods

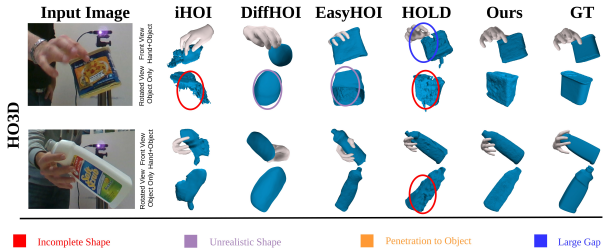

Figure 2. **Qualitative comparison with SOTA**. Reconstruction results from HO3D, comparing our method with SOTA baselines in both hand-object front view and object only rotated view.

2.2. Large hand–object deviation in HOLD

On short video clips, the pose-refinement stage of HOLD often produces large hand–object misalignment, as indicated by large CD_h in Table 1 of the main paper, which can cause the optimization to diverge, as illustrated in Figure 1. These failures stem from spurious contact-loss signals triggered by artifacts and noise on the unreconstructed backside of the object. By explicitly reconstructing the backside geometry, our method suppresses these errors and yields markedly more stable and accurate hand-pose estimates.

2.3. Qualitative comparison with SOTA methods

Additional examples are shown in Figure 2 to compare our results with recent SOTA approaches. Our method reliably completes unseen object regions and reconstructs complex shapes, even under severe hand-induced or self-occlusion, producing detailed geometry and a realistic hand–object spatial relationships. By contrast, HOLD [2] fails to recover the missing object parts and leaves a noticeable gap between the hand and the object; iHOI [9] struggles to reconstruct the full geometry; EasyHOI [4] hallucinates implausible shapes and exhibits hand–object penetration; and

Figure 3. **Reconstruction results of Trellis**. Reconstruction results from Trellis from front view and back view with the same input image as our method.

DiffHOI [10] typically recovers only overly simple geometry.

3. Discussion

Similar to HOLD [2], our method relies on accurate object pose initialization from HLoc [5, 6]; however, this initialization often fails for textureless or thin objects.

Although our method already surpasses the current SOTA reconstruction quality, integrating more advanced NVS models could further boost performance. As illustrated by the EasyHOI results, the SOTA image-to-3D method InstantMesh [8] still produces noticeable distortions. We therefore evaluated another SOTA image-to-3D approach, Trellis [7], using the same reference image as our method. As shown in Figure 3, even with this latest NVS model, some objects remain heavily distorted in the unobserved regions, and their surfaces tend to appear overly thick.

References

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *arXiv*, 2023. 1

[2] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 1, 2, 3

[3] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1

[4] Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild, 2024. 2

[5] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[6] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[7] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 3

[8] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3

[9] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3895–3905, 2022. 2

[10] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *International Conference on Computer Vision (ICCV)*, 2023. 3