Fusing Vision and Contact-Rich Physics Improves Object Reconstruction Under Occlusion

Bibit Bianchini*[†], Minghan Zhu*[†], Mengti Sun[‡], Bowen Jiang[†], Camillo J. Taylor[†], Michael Posa[†]
*The first two authors contributed equally to this work.
[†]GRASP Lab, {bibit, minghz, bwjiang, cjtaylor, posa}@seas.upenn.edu, [‡]Amazon, mengtis@amazon.com

Abstract*—We introduce Vysics, a vision-and-physics framework for building an expressive geometry and dynamics model of a rigid body, using a seconds-long RGBD video and robot proprioception. While the computer vision community has built powerful visual 3D perception algorithms, cluttered environments can limit visibility of objects of interest. However, observed motion of partially occluded objects can imply physical interactions took place, such as robot or environment contacts. Inferred contacts supplement the visible geometry with "physible" geometry, which best explains the observed object motion through physics. Vysics uses a vision-based tracking and reconstruction method, BundleSDF, to estimate the trajectory and visible geometry from an RGBD video, and an odometry-based model learning method, Physics Learning Library (PLL), to infer the "physible" geometry from the trajectory through implicit contact dynamics optimization. The visible and "physible" geometries jointly optimize the object's signed distance function (SDF). Vysics does not require pretraining, nor tactile or force sensors. Compared to vision-only, Vysics yields object models with higher geometric accuracy and better dynamics prediction in experiments where the object interacts with the robot and environment under heavy occlusion. Project page: https://vysics-vision-and-physics.github.io/

I. Introduction

Robots will encounter a vast array of different objects in in-the-wild manipulation. While some might be recognized from an existing database, others will require physical interaction to be understood on the spot. Dexterous manipulation of these objects will benefit from the ability to rapidly identify properties: geometry is most critical, but inertial properties are also valuable for predicting motion, particularly under force.

Rapid modeling requires combining all available information in a unified fashion. This work presents Vysics, which leverages recent results from visual tracking and object reconstruction [41] combined with contact-implicit model learning [7, 34] via the shared connection of object geometry. Visual information is limited by occlusions, but contact, which typically occurs on occluded faces of objects, provides a secondary source of information: "physible" geometry. However, estimating geometry through contact-rich interactions is nontrivial [4, 33]. Our approach embraces the multi-modal nature of the dynamics [6, 34], starting by feeding it visually-estimated trajectories from RGBD data, then fusing visually-observed with physically-inferred geometry. Vysics automatically generates a Unified Robotics Description File (URDF) with learned geometry that matches or outperforms vision-based approaches, in addition to other critical simulation parameters like inertia, with only seconds of data.

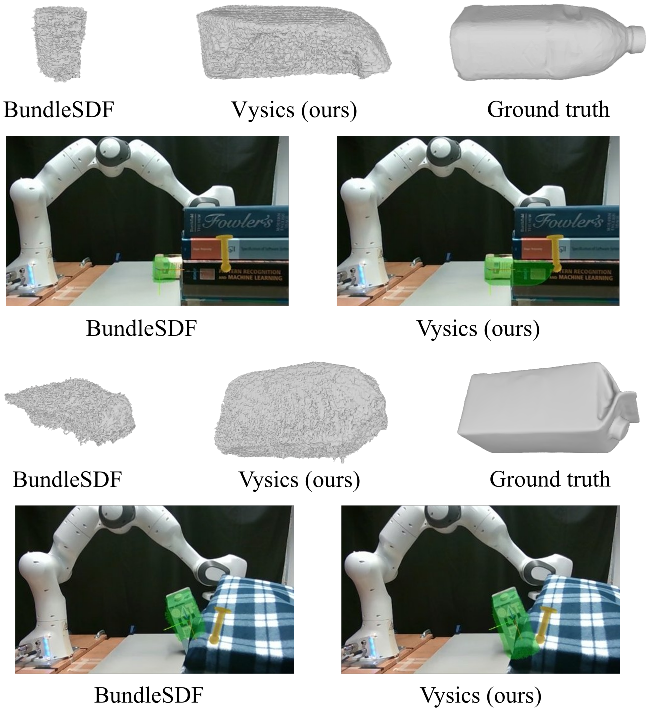


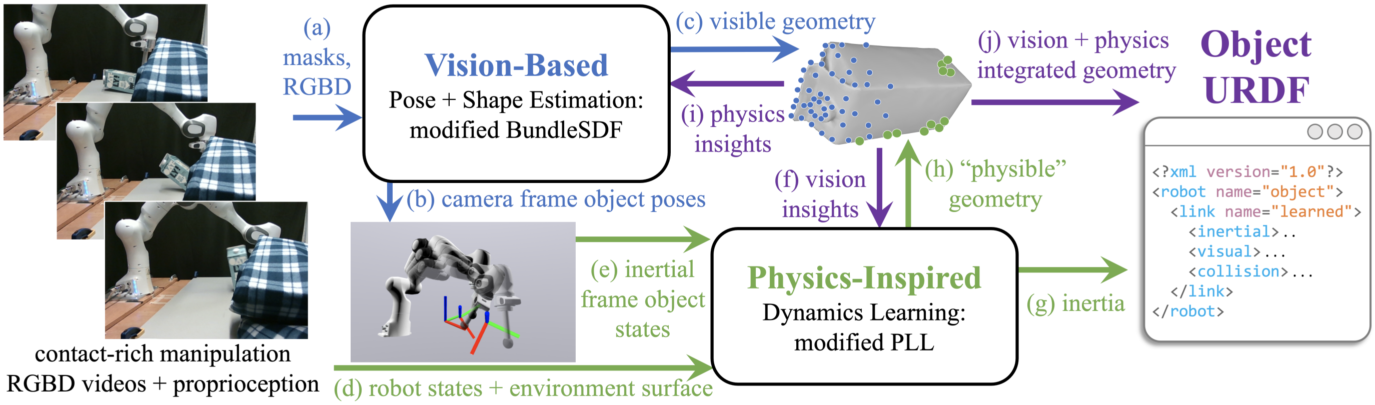
Fig. 1: Vision-based shape reconstruction (e.g. BundleSDF [41]) is limited by occlusion. Fusing vision and contact-rich physics, Vysics recovers occluded geometry through object interactions with the robot and environment. Shape (green) and end effector (yellow) projections show the interaction.

II. Related Work

Vysics is situated on rich histories of vision-based shape reconstruction, vision-based pose estimation, and trajectory-based dynamics model learning. Approaches combining vision and physics are newer and fewer but provide interesting alternatives with similar motivations. Here is a brief summary.

Vision-based geometry reconstruction and completion is a longstanding computer vision task. Classical reconstruction methods [8, 39, 29] leave occluded regions unresolved. Potentially offering a solution to fill in unseen gaps, learned completion methods take many forms, e.g. non-exhaustively [13, 30, 12, 45, 27, 44, 21]. We compare to several learned shape completion methods in our results, and Vysics consistently outperforms them by discerning physics as a secondary source of geometric information, instead of pretraining.

A common assumption in robotic manipulation is access to


(a) masks, RGBD
Vision-Based
Pose + Shape Estimation: modified BundleSDF
(c) visible geometry
(i) physics insights
(j) vision + physics integrated geometry
Object URDF
(b) camera frame object poses
(f) vision insights
(h) "physible" geometry
(e) inertial frame object states
Physics-Inspired
Dynamics Learning: modified PLL
(g) inertia
contact-rich manipulation
RGBD videos + proprioception
(d) robot states + environment surface

```
<?xml version="1.0"?>
<robot name="object">
  <link name="learned">
    <inertial>..
    <visual>...
    <collision>...
  </link>
</robot>
```

Fig. 2: Detailed Vysics diagram. Blue arrows denote vision-based information flow through BundleSDF [41], and green for PLL [7, 34]. Purple arrows indicate unifying connections to factor both vision and contact-rich physics into geometry learning.

vision-based object pose estimation. Many methods require the 3D model of the target object to facilitate the pose estimate [24, 23, 32], impractical in novel scenarios. Others do not require geometry models beforehand [42, 9, 38] but can be susceptible to long-term drift. Limiting to in-category objects can boost pose accuracy [26, 15, 10] but limits generalization. A new and exciting approach is to perform **simultaneous tracking and shape reconstruction** [47, 22, 41, 37], which has the benefit that maintaining a geometry estimate can improve novel object pose estimation and vice versa [36, 40].

Provided trajectories, **trajectory-based dynamics model learning** methods reason about physics. Differentiable simulators [25, 20, 14] have pushed advancements though can struggle in contact-rich settings [4, 6]. While high-stiffness dynamics generally are a challenge for system identification [33], creative strategies can efficiently find inertial parameters [16], contact parameters [34], or both [7]. While Vysics is not the first to use **physics as a prior for vision-based shape reconstruction**, other works assume objects are statically stable [31, 2], require tactile measurements and pretraining [43], or do not learn anything beyond geometry [1, 35].

III. Approach

Fig. 2 illustrates Vysics from input RGBD videos and robot states (left) to URDF output (right). Its core components are BundleSDF [41] for vision-based tracking and shape reconstruction, and PLL [7, 34] for physics-inspired dynamics learning. BundleSDF and PLL both train on and generate results from only the measurements provided or inferred by their per-instance input data, suitable for immediate application when a robot encounters and needs to model an object. Beyond systems integration insights, our main contribution lies in how Vysics incorporates these two powerful tools together to supervise each other and output an object dynamics model, featuring geometry informed by both vision and contact.

Referring to the labeled arrows in Fig. 2, BundleSDF estimates the object trajectory (b) and initial shape estimate (c) from masked input RGBD images (a). The object trajectories are converted (e) to an inertial reference frame, where the table surface (d) identified in the depth images is on a known plane. From these, PLL detects "physible" geometry by inferring contact events in the observed dynamics, subject to supervision from BundleSDF (f) to encourage consistency with the visible geometry. Lastly, BundleSDF runs again, fusing both the visible (a) and "physible" data (i) into a geometry consistent with both. The final output of Vysics inherits the physics-supervised inertial parameters (g) and jointly-supervised geometry (j), exported as a URDF which can be simulated.

A. Supervising Contact-Based Geometry with Vision

PLL [7, 34] represents geometry as a deep support function (DSF) [17], an input-convex, homogeneous deep neural network [3]. A DSF takes as input a unit vector and outputs the scalar distance the geometry extends in that direction [5]. The gradient of a DSF with respect to its input is (almost always [17]) the 3D point on the object geometry that extends furthest in the queried input direction. Mathematically, for an object whose surface (or volume) is represented by the set $\mathcal{S}$, a DSF yields the following output and gradient,

$$\mathrm{DSF}(\hat{\mathbf{n}}) = \max_{\mathbf{s}_i \in \mathcal{S}} \mathbf{s}_i \cdot \hat{\mathbf{n}}, \;\; \nabla_{\hat{\mathbf{n}}} \mathrm{DSF}\left(\hat{\mathbf{n}}\right) = \arg\max_{\mathbf{s}_i \in \mathcal{S}} \mathbf{s}_i \cdot \hat{\mathbf{n}} =: \mathbf{s}. \tag{1}$$

Fig. 3 exemplifies a queried normal direction $\hat{\mathbf{n}}$, its corresponding support point $\mathbf{s}$, and their implications for an SDF.

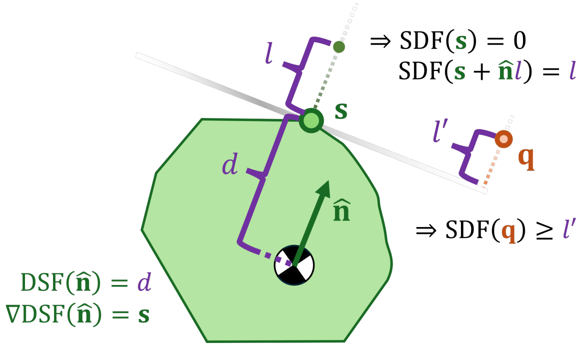

l
$\Rightarrow \mathrm{SDF}(\mathbf{s}) = 0$
$\mathrm{SDF}(\mathbf{s} + \hat{\mathbf{n}}l) = l$
$\mathbf{s}$
l'
$\mathbf{q}$
d
$\hat{\mathbf{n}}$
$\Rightarrow \mathrm{SDF}(\mathbf{q}) \geq l'$
$\mathrm{DSF}(\hat{\mathbf{n}}) = d$
$\nabla\mathrm{DSF}(\hat{\mathbf{n}}) = \mathbf{s}$

Fig. 3: A 2D depiction of the physical meaning of a DSF (1) and its SDF implications. Green points have exact SDF values and are subject to (3), and the $\mathbf{q}$ example's signed distance can be lower-bounded by the supporting hyperplane as in (4).

The visible geometry $\mathcal{V}$ can supervise PLL's DSF. For each vision-estimated surface point $\mathbf{s}^v \in \mathcal{V}$, we wish to penalize the distance from the nearest physics-estimated surface point $\mathbf{s}^p$ predicted by the DSF. The exact closest $\mathbf{s}^p$ from the DSF is not straightforward to obtain, so we approximate it by sampling many querying vectors $\hat{\mathbf{n}}^p$ and selecting the one with $\nabla\text{DSF}(\hat{\mathbf{n}}^{p'}) = \mathbf{s}^p$ closest to $\mathbf{s}^v$. The approximation is up to the angular resolution of the unit vector samples. We use the following as the vision-based supervision during PLL training:

$$\mathcal{L}_{\text{bsdf}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{s}^v \in \mathcal{V}} \left\| \nabla\text{DSF}(\hat{\mathbf{n}}^{p'}) - \mathbf{s}^v \right\|. \tag{2}$$

B. Supervising Vision-Based Geometry with Contact

BundleSDF [41] represents geometry as a signed distance function (SDF), where any 3D point $\mathbf{p}$ relative to a geometry can be queried, and the scalar signed distance d away to the nearest surface point is returned: $\text{SDF}(\mathbf{p}) = d$. Vysics incorporates physics into the SDF regression via contact-based loss terms. PLL's DSF yields a set of ∇DSF input/output pairs, $\{(\hat{\mathbf{n}}^p, \mathbf{s}^p)\}$, of which we retain only those with PLL-hypothesized contact force above a threshold. The filtered set, $\mathcal{P}$, is the "physible" geometry PLL outputs. Given one pair $(\hat{\mathbf{n}}^p, \mathbf{s}^p)$, any point on the ray $\vec{r}$ from $\mathbf{s}^p$ pointing in direction $\hat{\mathbf{n}}^p$, lying a distance $l \in [0, \infty)$ from $\mathbf{s}$, has a signed distance of l (see Fig. 3). Extending the possible range for l to go negative, i.e. $l \in [-\epsilon, \infty)$, means points with $l < 0$ are not guaranteed to be correct but encourage SDF zero-crossings, effecting change at the learned surface. Our **support point loss** encourages SDF consistency around "physible" points:

$$\mathcal{L}_{\text{sp}} = \left(\text{SDF}(\mathbf{s}^p + l\hat{\mathbf{n}}^p) - l\right)^2 \tag{3}$$

On this ray, (3) imposes strong supervision on the SDF network, though only local to areas near PLL-inferred contacts. However, as depicted in Fig. 3, *any* point $\mathbf{q}$ can have its signed distance minimum-bounded based on a support direction/point pair $(\hat{\mathbf{n}}, \mathbf{s})$. Consider a pair $(\hat{\mathbf{n}}^p, \mathbf{s}^p) \in \mathcal{P}$: its supporting hyperplane implies that the signed distance at $\mathbf{s}^{v'}$ can be lower-bounded by the distance from $\mathbf{s}^v$ to the supporting hyperplane. Thus, we introduce a **hyperplane-constrained loss** valid for any $\mathbf{s}^{v'} \in \mathbb{R}^3$,

$$\mathcal{L}_{\text{hc}} = \min\left(0,\ \text{SDF}(\mathbf{s}^{v'}) - (\mathbf{s}^{v'} - \mathbf{s}^p) \cdot \hat{\mathbf{n}}^p\right)^2. \tag{4}$$

While $\mathbf{s}^{v'}$ may be sampled arbitrarily, we sample them around a cylindrical neighborhood of the support points in practice.

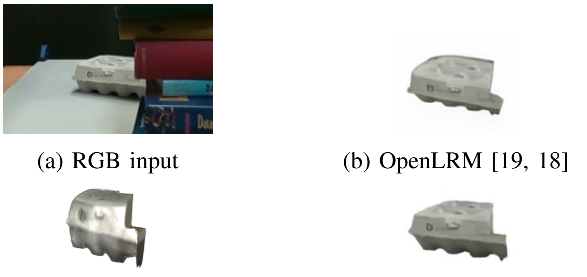
In comparison to the dense visible points, "physible" points are sparser and can be located far away from the set of visible points. With the assumption that the robot is interacting with a single object at a time, we add a bias **convexity loss** term to encourage the estimated shape to be convex when no observed RGBD data signals otherwise. This helps the sparse contact points attach to the visible shape in the SDF regression.

IV. Experiments and Results

We consider a new dataset of 30Hz RealSense D455 RGBD videos of and joint states from a teleoperated Franka Emika Panda arm with a spherical end effector interacting with one of seven everyday objects repeatedly on a flat table surface. There are substantial visual occlusions preventing the camera from directly seeing much of the object geometry. Ground truth object meshes are used only for evaluation. The end-effector pose commands by the teleoperator are also included for the dynamics prediction evaluation.

Method	bakingbox	bottle	egg	milk	oatly	styrofoam	toblerone	all
3DSGrasp [30]	3.83	2.80	3.78	3.15	2.51	2.66	2.77	3.06
IPoD [46]	3.25	1.80	2.16	2.37	2.73	1.93	1.97	2.47
V-PRISM [44]	3.52	2.47	2.31	3.33	2.30	2.54	2.48	2.80
OctMAE [21]	3.11	2.22	1.52	2.93	2.13	2.00	2.36	2.45
BundleSDF [41]	3.84	2.65	3.70	3.17	2.45	2.55	2.44	2.98
Vysics (ours)	**1.83**	**1.36**	**1.05**	**1.53**	**1.25**	**1.45**	**1.02**	**1.45**

TABLE I: Average chamfer distance (unit: cm) of shape completion baselines compared with BundleSDF and our method.



(a) RGB input

(b) OpenLRM [19, 18]

(c) One-2-3-45++ [28]

(d) TriplaneGaussian [48]

Fig. 4: A qualitative example of generative single-view reconstruction on an occluded RGB image of the *egg* object.

The object masks are semi-automatically generated from manual masks on the first frame using XMem [11]. For every object, we collected multiple, roughly 10-second sessions of the robot arm interacting with the object with its spherical end effector, varying in starting configurations, occlusions, and interactions. The number of sessions per object vary, since we exclude any in which BundleSDF lost track of the object. We use PLL to learn geometry and inertia, fixing the pair-wise friction coefficients to a reasonable value for all experiments.

A. Geometry Results

Table I presents the quantitative results of shape completion models in chamfer distance, averaged per object and over all objects, compared with BundleSDF and Vysics. Under severe occlusion, while the shape completion models can achieve similar or slightly lower chamfer distance than pure vision-based reconstruction, BundleSDF, they fall behind Vysics by a large margin, showing that the data-driven completion models are not as successful as Vysics at filling in the missing pieces. Qualitative results of the single-view 3D generation models are shown in Fig. 4. We find that these generative models typically assume an unobstructed view of the object and do not generate a complete shape when given a partially occluded view. Therefore, these models are not evaluated quantitatively.

We compare Vysics and BundleSDF in detail, as neither requires any pretraining. Fig. 5 shows the quantitative re-

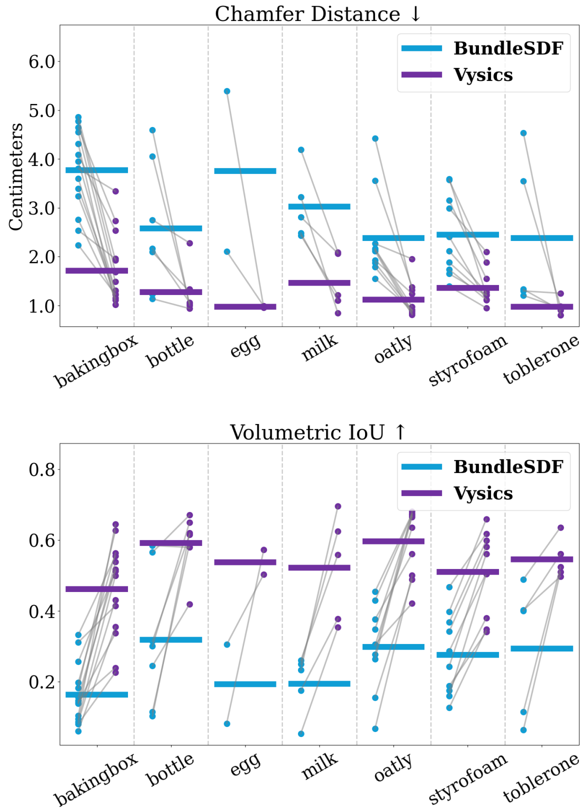


Fig. 5: The quantitative comparison of the geometric reconstruction accuracy. Each dot is one session. The results of the same session from different methods are connected by a gray line. ↑ means higher is better. ↓ means lower is better.

sults in terms of the surface-based metric, chamfer distance, and the volume-based metric, IoU. Vysics substantially and consistently improves the geometric accuracy in both metrics over BundleSDF. Qualitative comparisons are shown in Fig. 1. BundleSDF misses a significant portion of the objects in its geometry estimates, while our method recovers the occluded geometries so that the robot arm's interactions with the objects can explain the observed object trajectory.

B. Dynamics Prediction Results

We further use dynamics predictions to show that the geometry estimated by our method better explains the observed trajectory. Fig. 6 compares Vysics against the vision-only baseline, physics-only baseline, and a baseline featuring a simulation with the ground truth geometry. As expected, the ground truth geometry simulations maintained more accurate dynamics predictions for the longest. We point out that even this baseline is imperfect, despite using essentially perfect geometry, due to inaccurate modeling assumptions such as object rigidity and the divergent nature of the dynamics in many of our robot interactions. Vysics and PLL perform closely to this baseline, though Vysics is moderately worse in orientation divergence. While most of the dynamics performance by PLL is retained in Vysics, it is unsurprising to see a slight performance drop, given PLL optimizes only for physics accuracy while Vysics balances with visual objectives. The vision-only baseline is the least performant in both position and orientation rollout accuracy.

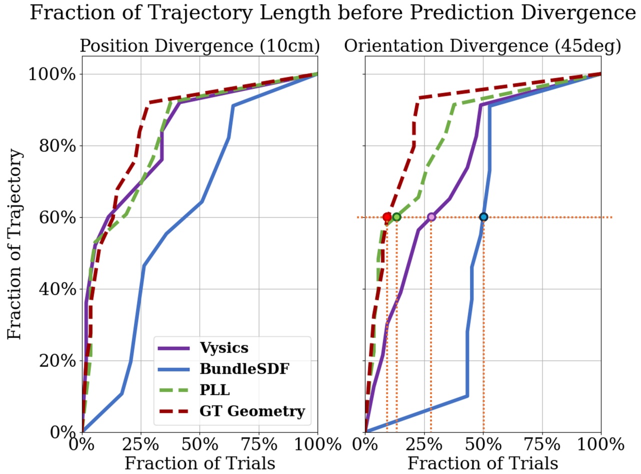


Fig. 6: For quantifying dynamics prediction performance, we compare how far into an open-loop rollout the predicted pose stays within 10cm of position error and within 45 degrees of rotational error from the BundleSDF tracked poses. We normalize the y-axis to the length of the trajectory. An example interpretation from the right plot (the orange dashed lines): 60% into the predicted trajectory, approximately half of the BundleSDF dynamics predictions diverged in orientation compared to 30% of the Vysics dynamics predictions and 10% of the ground truth geometry and PLL predictions.

V. Conclusion and Limitations

Vysics enables robots to construct high-fidelity dynamics models of novel objects, identified from vision and proprioception, in the face of contact-rich interactions and extremely little data. This is the first step toward unifying vision-based geometry estimation with contact dynamics. Future work might replace teleoperated data collection with autonomous, active exploration or the integration of these learned models with planning and control to accomplish some desired task.

A considerable limitation of our current implementation is that it often cannot recover from poor pose estimates. In our experience, this is accentuated by our occlusion-rife, contact-rich dataset. We find shorter video lengths usually result in more consistent pose tracking, but at odds with the benefits of more data for dynamics parameter regression. The geometry supervision from PLL during BundleSDF's second run might help it perform better pose estimation. In this case, we could cyclically repeat our BundleSDF-PLL process until both shape and trajectories converge. Future versions of Vysics may consider posing a single joint learning problem that performs pose estimation and dynamics model building simultaneously.

Acknowledgments

This work was supported by an NDSEG Fellowship, an NSF CAREER Award (grant FRR-2238480), and the RAI Institute.

REFERENCES

[1] Jad Abou-Chakra, Krishan Rana, Feras Dayoub, and Niko Suenderhauf. Physically embodied gaussian splatting: A realtime correctable world model for robotics. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=AEq0onGrN2.

[2] William Agnew, Christopher Xie, Aaron Walsman, Octavian Murad, Yubo Wang, Pedro Domingos, and Siddhartha Srinivasa. Amodal 3d reconstruction for robotic manipulation via stability and connectivity. In *Conference on Robot Learning*, pages 1498–1508. PMLR, 2021.

[3] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[4] Rika Antonova, Jingyun Yang, Krishna Murthy Jatavallabhula, and Jeannette Bohg. Rethinking optimization with differentiable simulation from a global perspective. In *Conference on Robot Learning*, pages 276–286. PMLR, 2023.

[5] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear programming: theory and algorithms*. John wiley & sons, 2013.

[6] Bibit Bianchini, Mathew Halm, Nikolai Matni, and Michael Posa. Generalization bounded implicit learning of nearly discontinuous functions. In *Learning for Dynamics and Control Conference*, pages 1112–1124. PMLR, 2022.

[7] Bibit Bianchini, Mathew Halm, and Michael Posa. Simultaneous learning of contact and continuous dynamics. In *Conference on Robot Learning*, pages 3966–3978. PMLR, 2023.

[8] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2560–2568, 2018.

[9] Pedro Castro and Tae-Kyun Kim. Posematcher: One-shot 6d object pose estimation by deep feature matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2148–2157, 2023.

[10] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 139–156. Springer, 2020.

[11] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[12] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4456–4465, 2023.

[13] Ruihang Chu, Enze Xie, Shentong Mo, Zhenguo Li, Matthias Nießner, Chi-Wing Fu, and Jiaya Jia. Diffcomplete: Diffusion-based generative 3d shape completion. *Advances in neural information processing systems*, 36: 75951–75966, 2023.

[14] Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J Zico Kolter. End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, 31:7178–7189, 2018.

[15] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022.

[16] Nima Fazeli, Samuel Zapolsky, Evan Drumwright, and Alberto Rodriguez. Learning data-efficient rigid-body contact models: Case study of planar impact. In *Conference on Robot Learning*, pages 388–397. PMLR, 2017.

[17] Mathew Halm. *Addressing Stiffness-Induced Challenges in Modeling and Identification for Rigid-Body Systems With Friction and Impacts*. PhD thesis, University of Pennsylvania, 2023.

[18] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023.

[19] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

[20] Taylor A Howell, Simon Le Cleac'h, J Zico Kolter, Mac Schwager, and Zachary Manchester. Dojo: A differentiable simulator for robotics. *arXiv preprint arXiv:2203.00806*, 2022.

[21] Shun Iwase, Katherine Liu, Vitor Guizilini, Adrien Gaidon, Kris Kitani, Rareş Ambruş, and Sergey Zakharov. Zero-shot multi-object scene completion. In *European Conference on Computer Vision*, pages 96–113. Springer, 2024.

[22] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. *International Conference on 3D Vision (3DV)*, 2024.

[23] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020.

[24] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef

Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

[25] Simon Le Cleac'h, Mac Schwager, Zachary Manchester, Vikas Sindhwani, Pete Florence, and Sumeet Singh. Single-level differentiable contact simulation. *IEEE Robotics and Automation Letters*, 2023.

[26] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1258–1264. IEEE, 2022.

[27] Stefan Lionar, Xiangyu Xu, Min Lin, and Gim Hee Lee. Nu-mcc: Multiview compressive coding with neighborhood decoder and repulsive udf. *Advances in Neural Information Processing Systems*, 36:63011–63022, 2023.

[28] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10083, 2024.

[29] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018.

[30] Seyed S Mohammadi, Nuno F Duarte, Dimitrios Dimou, Yiming Wang, Matteo Taiana, Pietro Morerio, Atabak Dehban, Plinio Moreno, Alexandre Bernardino, Alessio Del Bue, et al. 3dsgrasp: 3d shape-completion for robotic grasp. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3815–3822. IEEE, 2023.

[31] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. 2024.

[32] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

[33] Mihir Parmar, Mathew Halm, and Michael Posa. Fundamental challenges in deep learning for stiff contact dynamics. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5181–5188. IEEE, 2021.

[34] Samuel Pfrommer, Mathew Halm, and Michael Posa. ContactNets: Learning Discontinuous Contact Dynamics with Smooth, Implicit Representations. In *The Conference on Robot Learning (CoRL)*, 2020. URL https://proceedings.mlr.press/v155/pfrommer21a.html.

[35] Changkyu Song and Abdeslam Boularias. Inferring 3d shapes of unknown rigid objects in clutter through inverse physics reasoning. *IEEE Robotics and Automation Letters*, 4(2):201–208, 2018.

[36] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6855–6865, 2022.

[37] Edgar Sucar, Kentaro Wada, and Andrew Davison. Nodeslam: Neural object descriptors for multi-view shape reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 949–958. IEEE, 2020.

[38] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.

[39] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[40] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020.

[41] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023.

[42] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. *arXiv preprint arXiv:2312.08344*, 2023.

[43] Youngsun Wi, Andy Zeng, Pete Florence, and Nima Fazeli. Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects. *arXiv preprint arXiv:2210.03701*, 2022.

[44] Herbert Wright, Weiming Zhi, Matthew Johnson-Roberson, and Tucker Hermans. V-prism: Probabilistic mapping of unknown tabletop scenes. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1078–1085. IEEE, 2024.

[45] Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, and Georgia Gkioxari. Multiview compressive coding for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9065–9075, 2023.

[46] Yushuang Wu, Luyue Shi, Junhao Cai, Weihao Yuan, Lingteng Qiu, Zilong Dong, Liefeng Bo, Shuguang Cui, and Xiaoguang Han. Ipod: Implicit field learning with point diffusion for generalizable 3d object reconstruction from single rgb-d images. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20432–20442, 2024.

[47] Weiming Zhi, Haozhan Tang, Tianyi Zhang, and Matthew Johnson-Roberson. Simultaneous geometry and pose estimation of held objects via 3d foundation models. *IEEE Robotics and Automation Letters*, 2024.

[48] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10335, 2024.