Instance-Agnostic Geometry and Contact Dynamics Learning

Mengti Sun*, Bowen Jiang*, Bibit Bianchini, Camillo Jose Taylor, Michael Posa

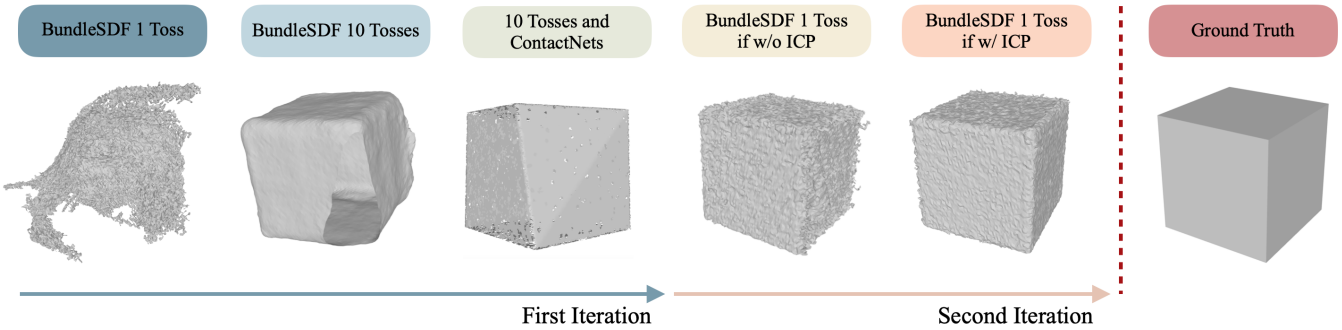

Fig. 1: Reconstructed geometries at different steps of the framework. The right-most geometry represents the final result, which significantly outperforms the initial reconstruction on one toss and even the reconstruction from ten tosses.

***Abstract*— This work presents an instance-agnostic learning framework that fuses vision with dynamics to simultaneously learn shape, pose trajectories, and physical properties via the use of geometry as a shared representation. Unlike many contact learning approaches that assume motion capture input and a known shape prior for the collision model, our proposed framework learns an object's geometric and dynamic properties from RGBD video, without requiring either category-level or instance-level shape priors. We integrate a vision system, BundleSDF, with a dynamics system, ContactNets, and propose a cyclic training pipeline to use the output from the dynamics module to refine the poses and the geometry from the vision module, using perspective reprojection. Experiments demonstrate our framework's ability to learn the geometry and dynamics of rigid and convex objects and improve upon the current tracking framework.**

I. INTRODUCTION

Robots are increasingly being employed in complex real-world scenarios where they interact with a variety of objects. Vision systems are often employed to perform basic object detection and segmentation tasks [1]–[5]. However successful manipulation often requires more detailed information about the object shape and relevant contact dynamics [6].

This work explores an approach that combines an object modeling system with a system for refining geometry and modeling contact dynamics from data. It expands the horizon of robotic applications and allows us to contemplate systems that can manipulate previously unseen objects.

In the pursuit of estimating dynamics, recent research emphasizes the modeling of irregularities in an observed object trajectory, where the most interesting manipulation or frictional contact behaviors usually happen [6]. ContactNets [6], as a state-of-the-art example, addresses this challenge from a motion capture perspective, obviating the need for force-sensing instruments. It optimizes inter-body distance functions and contact-frame Jacobians; however, it relies on having estimates for the object pose throughout the trajectory and a shape prior, such as a 3D mesh of the object, to learn a non-penetration loss function. Acquiring such a shape prior like a CAD model [7], [8], especially in open-world settings, can be labor intensive.

The BundleSDF [9] framework is a state of the art system that can be used to build a 3D model of a moving object while simultaneously estimating its pose in space [10]–[12]. This makes it an appropriate candidate to provide pose and shape priors to the ContactNets dynamics framework.

There are, however, some challenges associated with using BundleSDF in practice. When the trajectory is brief, pose estimation accuracy is high, but the distribution of viewpoints might be inadequate to achieve comprehensive geometry reconstruction. Besides, if the trajectory is lengthy, error accumulation compromises pose estimation and tracking accuracy, although the geometry reconstruction tends to be more precise. Motivated by this observation, our work is designed to learn a better shape representation, and then leverage it to improve pose estimation on shorter trajectories.

Our contribution can be summarized as follows:

1) A novel integration method that use RGBD videos to produce a refined geometry and dynamics model from a physical learning framework.
2) Design of a cyclic pipeline utilizing improved geometry to refine pose estimation on short trajectories, harnessing reprojection and iterative closest points.
3) Experiments on a cube toss dataset collected with a manipulator that demonstrates improved pose estimation and geometry reconstruction over SOTA baselines.

*Authors have equal contribution.

All authors are affiliated with the GRASP Lab at the School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, 19104. {mengti, bwjiang, bibit, cjtaylor, posa}@seas.upenn.edu

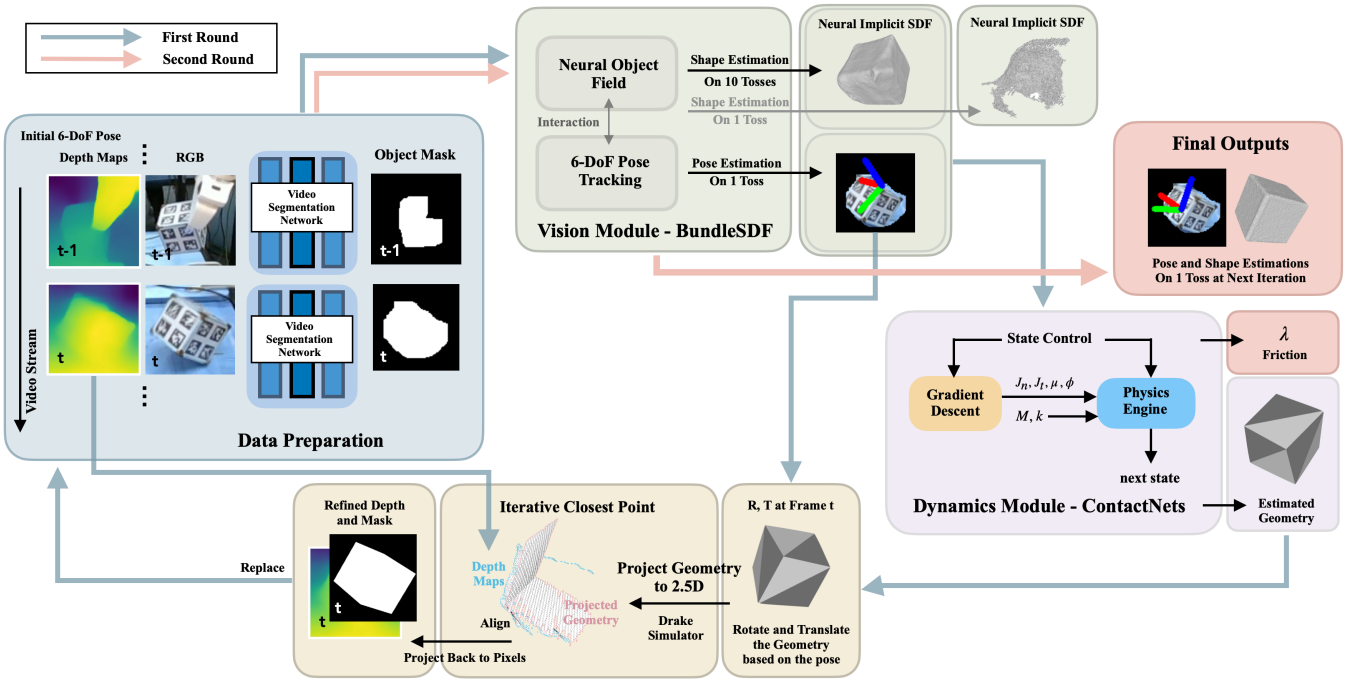

Fig. 2: Framework overview. In the first round, BundleSDF predicts poses and geometry, and ContactNets refines the geometry and predicts frictional forces. The refined geometry is transformed based on the estimated pose and projected into the frame of the RGBD camera as a 2.5D point cloud. ICP is used to align this predicted depth and the raw depth in the dataset. The aligned geometry is projected to the pixel domain as a refined mask and depth. In the next round, BundleSDF can predict better poses and geometry from the refined dataset.

II. RELATED WORK

Neural Radiance Fields NeRF [13], [14] have shown impressive results in constructing detailed 3D models of objects from a sparse set of views with zero 3D supervision. [15]–[17], [17] address the limitation of the original NeRF framework [13] by introducing the temporal dimension to capture moving objects. BundleSDF [9] utilizes an advanced NeRF model [13] by combining geometry reconstruction with pose estimation, efficient ray sampling with depth-guided truncation for faster convergence, and normal-guided implicit regularization for smooth surface extraction.

Contact Dynamics Learning In the realm of Real-to-Simulation [18]–[21], recent efforts focus on understanding and modeling the frictional contact dynamics of rigid bodies, estimating mass properties, inertia matrices, and the frictional forces in collisions. ContactNets [6] implicitly parameterizes the discontinuous contact behavior with a continuous inter-body signed distance function and contact-frame Jacobians. The system initializes the geometry from a shape prior, and optimizes it from trajectory measurements and a physics model. Relevant literature includes DANO [22] and GradSim [21]. This manuscript considers a new approach by integrating BundleSDF [9] to improve object modeling.

III. APPROACH

We depict the overview of our approach in Figure 2. The framework takes a collection of n RGBD videos of the same object manipulated by a robotic arm, denoted as $\mathcal{V} = \{v_1, v_2, ..., v_n\}$. Each video sequence $v_i \in \mathcal{V}$ represents a toss of the rigid body colliding with a rigid flat surface. The videos in $\mathcal{V}$ are consecutive in time, so they can be regarded as part of one longer video. Each frame in $v_i \in \mathcal{V}$ comprises an RGB image $\boldsymbol{I} \in \mathbb{R}^{3\times r\times c}$ and a depth map $\boldsymbol{D} \in \mathbb{R}^{r\times c}$, where r and c denote the number of rows and columns. A binary mask $\boldsymbol{M} \in \mathbb{R}^{r\times c}$ and a starting pose of the object to be tracked in the camera frame are only required for the very first frame in the sequence, although optionally, binary masks for more frames can be provided for better instance segmentation.

a) Semi-Supervised Instance Segmentation: Video object segmentation (VOS) [23]–[25] is an ideal approach to ensure the coherence of segmentation masks throughout a trajectory without labor-intensive annotations. Given the initial frame's segmentation mask of the object instance, we leverage the XMem [23] model to automatically generate segmentation masks for the target object in all subsequent frames. It employs multiple distinct yet interconnected feature memory stores initialized from the initial frame. For each following frame, it retrieves information from these memories to generate a mask. This innovative approach yields high-quality features while conserving memory usage, and is particularly well-suited for the extended video sequences in our experiment.

b) Initial Estimation on a Single Video Clip: This step runs BundleSDF on one video clip v_i as a baseline. It finds visual feature correspondence between consecutive frames and collects a memory pool of keyframes that will

participate in the pose graph optimization to ensure multi-view consistency and a Neural Object Field [9] for geometry reconstruction. It generates initial pose estimates $\boldsymbol{p}$ for all frames in v_i and a rough geometry reconstruction $\boldsymbol{O}_0$.

c) Geometry Reconstruction: This step reconstructs a holistic geometry of the target object by sending the entire repertoire of video sequences in $\mathcal{V}$ to BundleSDF [9]. Compared to an individual sequence $v_i \in \mathcal{V}$, a more extended sequence offers a wider range of viewpoints and contributes to more complete and smooth shape reconstructions. It also eliminates the occlusion problem in certain frames.

The Neural Object Field represents shapes as level sets of a deep neural network Ω that learns geometry functions

$$\Omega : \boldsymbol{x} \in \mathbb{R}^3 \mapsto s \in \mathbb{R}, \quad \left\{\boldsymbol{x} \in \mathbb{R}^3 \mid \Omega(\boldsymbol{x}) = 0\right\}, \tag{1}$$

where $\boldsymbol{x}$ is a 3D coordinate, s is a signed distance value, and the zero set of the geometry function is the implicit surface of the object instance that can be derived from the first-order derivative of the neural network.

The optimization considers the uncertain free space, empty space, and near-surface space [9] to handle imperfections in segmentation masks and depth maps. The Eikonal regularization $L \propto (\|\nabla\Omega(\boldsymbol{x})\|_2 - 1)^2$ [26] is used to train the near-surface signed distance field. The framework outputs a geometry $\boldsymbol{O}$ describing the object instance in the 3D space.

d) Contact Dynamics Learning: We now focus on a single video $v_i \in \mathcal{V}$. We let the ContactNets framework [27] take the 6-DoF pose estimations $\boldsymbol{p}_0$ in one video v_i and reconstructed shape prior $\boldsymbol{O}$ from all videos as the inputs.

ContactNets learns functional approximations of the inter-body distance $\Phi_j(\boldsymbol{q})$ with a geometric prior $\boldsymbol{O}$, the contact Jacobian $\boldsymbol{J}_j(\boldsymbol{q})$, and the friction coefficient μ_j using state transitions $\mathcal{D} = (\boldsymbol{x}_j, \boldsymbol{u}_j, \boldsymbol{x}'_j)_{j\in 1,\dots,D}$, where $\boldsymbol{x}_j = [\boldsymbol{q}_j; \boldsymbol{v}_j]$ is the state of j-th contact frame, $\boldsymbol{q}_j$ contains the robot configuration and pose, $\boldsymbol{v}_j$ denotes the velocity, and $\boldsymbol{x}'_j$ is the descrete-time dynamics of the system [27]. We follow the assumption in [27] that the inertia quantities and non-contact impulses are known. The training loss function on the contact impulses $\boldsymbol{\lambda}_j$ and current state transition $\mathcal{L}(\boldsymbol{x}_j, \boldsymbol{u}_j, \boldsymbol{x}'_j, \boldsymbol{\lambda}_j)$ is established on prediction quality, contact activation, non-penetration criterion, and maximal dissipation [27], and the problem is formulated as a tractable and convex program solved by gradient descent.

The learned inter-body distance function $\Phi_j(\boldsymbol{q})$ outputs a 3D geometry $\tilde{\boldsymbol{O}}$ of the object instance, refined from the shape prior $\boldsymbol{O}$ with richer information from contact dynamics.

e) Cyclic Pipeline for Pose and Geometry Refinement: We demonstrate a novel approach that interlaces the training of the BundleSDF and ContactNets frameworks in a cyclic fashion. Leveraging the geometry offered by ContactNets from full videos and contact dynamics, we devise a re-projection and iterative closest point (ICP) algorithm to allow BundleSDF to effectively refine its geometry and pose estimations in the next iteration. The framework can also be applied to a future video $v_{i'>n}$ without relearning $\tilde{\boldsymbol{O}}$.

Utilizing the previous pose estimate $\boldsymbol{p}_0$ for each frame and the current point cloud $\tilde{\boldsymbol{O}}$, we can rotate and translate each point $\boldsymbol{w} \in \mathbb{R}^3$ in the 3D point cloud $\tilde{\boldsymbol{O}}$ to align it with the object instance in each frame of the video v_i, respectively.

$$\tilde{\boldsymbol{O}}' = \{R \cdot \boldsymbol{w} + T \mid \forall \boldsymbol{w} \in \tilde{\boldsymbol{O}}\}, \tag{2}$$

where $R \in \mathbb{R}^{3\times3}$ and $T \in \mathbb{R}^3$ are the rotation and translation matrices obtained from $\boldsymbol{p}_0$, and $\cdot$ denotes matrix multiplication. We project the transformed 3D point cloud $\tilde{\boldsymbol{O}}'$ to the image pixel domain using the Drake simulator [28] to generate a new set of depth maps and segmentation masks.

$$\boldsymbol{D}' = \{K \cdot \boldsymbol{w} \mid \forall \boldsymbol{w} \in \tilde{\boldsymbol{O}}'\} \tag{3}$$

$$\boldsymbol{M}'[u,v] = \begin{cases} 1 & \text{if } \boldsymbol{D}'[u,v] \neq \text{background} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $K \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix.

Here we rely on the initial pose estimates $\boldsymbol{p}_0$ derived from BundleSDF to transform the geometry $\tilde{\boldsymbol{O}}$, assuming that these poses are the pseudo ground-truth. Although the geometry $\tilde{\boldsymbol{O}}$ is better than the geometry from running BundleSDF on a single video clip, there could be a rotational and translational disparity between the modified depth maps $\boldsymbol{D}'$ obtained from $\tilde{\boldsymbol{O}}'$ and the original depth maps $\boldsymbol{D}$. Therefore, we implement an iterative closest point (ICP) algorithm to align these depth maps and bridge such disparities.

We project these two depth maps $\boldsymbol{D}$ and $\boldsymbol{D}'$ into the RGBD camera view (also referred to as the 2.5D view to denote partial 3D information from depth maps) to create the source point cloud $\boldsymbol{O}_s$ and target point cloud $\boldsymbol{O}_t$. The ICP algorithm aims to find the corresponding points between two point clouds and a rotation R and translation T that minimizes their sum of squared distances.

$$R, T = \arg\min_{R,T} \sum \|\boldsymbol{w}_t - R \cdot \boldsymbol{w}_s - T\|_2^2, \ \forall \boldsymbol{w}_s \in \boldsymbol{O}_s, \boldsymbol{w}_t \in \boldsymbol{O}_t \tag{5}$$

The optimized R and T resulting from the ICP alignment process are employed to transform the source point cloud $\boldsymbol{O}_s$ in the 3D space. Subsequently, the transformed point cloud $\boldsymbol{O}'_s$ is projected back to the pixel space, yielding refined depth maps $\boldsymbol{D}''$ and segmentation masks $\boldsymbol{M}''$, similar to the steps outlined in Equations 2-4. At this moment, the refined set of depth maps $\boldsymbol{D}''$, segmentation masks $\boldsymbol{M}''$, as well as the original RGB images and starting pose from the initial frame constitute a new dataset. This newly formed dataset is then subjected to another iteration of the BundleSDF framework. This iterative procedure results in further enhanced pose estimates and geometry reconstructions.

IV. EXPERIMENTS

We test our framework on a robot cube toss dataset we collected: a rigid cube affixed with AprilTags is tossed by a Franka Panda 7-DoF robot arm onto a flat rigid surface while data is captured by a 30Hz IntelRealSense D455 RGBD camera. The cube is tracked by three 30Hz PointGrey cameras using TagSLAM [29] to establish ground-truth poses.

Numerical results and ablation studies are shown in Table I. We evaluate pose estimation and geometry reconstruction separately as in [9]: For 6-DoF object pose, we compute

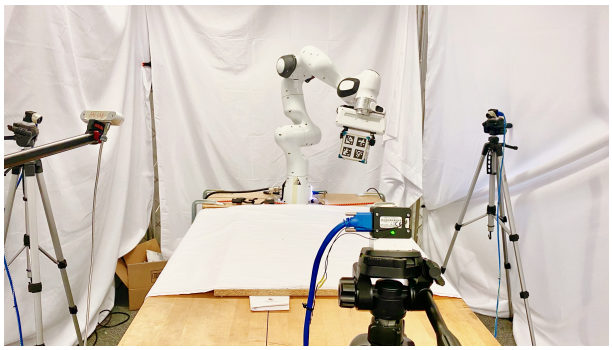

Fig. 3: Experimental setup used in the data collection process showcasing the robotic arm grasping the cube for tossing.

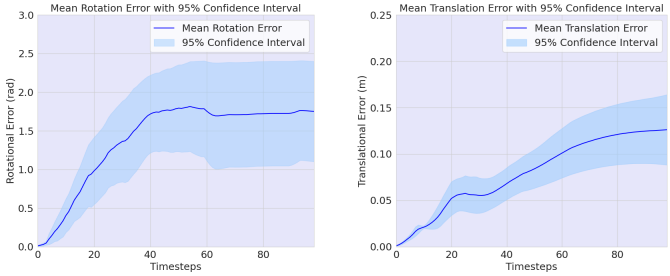

Fig. 4: Estimated rollouts from input initial condition and ContactNets model trained with a dataset of 9 tosses. The rotation and translation errors are compared with the input trajectory from BundleSDF output assuming as ground truth.

1) area under the curve (AUC) percentage of **ADD**. 2) AUC percentage of **ADD-S**.

$$ADD = \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{w} \in \mathcal{M}} \|(R \cdot \boldsymbol{w} + T) - (\tilde{R} \cdot \boldsymbol{w} + \tilde{T})\|_2 \quad (6)$$

$$ADD\text{-}S = \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{w}_1 \in \mathcal{M}} \min_{\boldsymbol{w}_2 \in \mathcal{M}} \|(R \cdot \boldsymbol{w}_1 + T) - (\tilde{R} \cdot \boldsymbol{w}_2 + \tilde{T})\|_2 \quad (7)$$

where the tilde denotes ground-truth R and T and $\mathcal{M}$ being object's ground-truth geometry. 3) success rate **SR** under 5°5cm metric [9], namely orientation error within 5° and translation error within 5 cm. For geometry reconstruction, we compute the Chamfer distance $\textbf{CD} = \frac{1}{2|\mathcal{M}_1|} \sum_{\boldsymbol{w}_1 \in \mathcal{M}_1} \min_{\boldsymbol{w}_2 \in \mathcal{M}_2} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + \frac{1}{2|\mathcal{M}_2|} \sum_{\boldsymbol{w}_2 \in \mathcal{M}_2} \min_{\boldsymbol{w}_1 \in \mathcal{M}_1} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$ between the generated and ground-truth meshes.

The last row of Table I indicates the results achieved by our framework, which exhibits a stronger performance than the baselines. We posit that the depth maps $\boldsymbol{D}''$ and masks $\boldsymbol{M}''$ derived from the refined geometry exhibit the capability to effectively recover occlusions and defective masks better than the original depth $\boldsymbol{D}$ and masks $\boldsymbol{M}$. Besides, the final reconstructed geometry, as shown in Figure 1, significantly outperforms the original BundleSDF, regardless of whether it is running on 1 toss or 10 tosses. It's intriguing that the entire system has no prior knowledge of the geometry, yet it still recovers the cube almost flawlessly.

In the ablation study of the cyclic training pipeline, having the cycle will decrease the CD score to 0.18cm with better pose estimation simultaneously. Besides, compared to BundleSDF on 10 tosses, the reconstructed geometry undergoes a refinement process by ContactNets with dynamic contact learning, which reduces CD from 1.75cm to 1.18cm.

When the ContactNets module is omitted, we directly reproject the geometry from running BundleSDF on 10 tosses to get $\boldsymbol{M}''$ and $\boldsymbol{D}''$, bypassing ContactNets. It shows that the process of learning contact dynamics exerts a positive influence on both pose estimations and geometry reconstruction by BundleSDF in the next iteration.

We also evaluate the ContactNets output by comparing the estimated rollouts with the ground-truth trajectories. The training dataset comprises a total of 9 trajectories, which are pose estimations from BundleSDF. As illustrated in Figure 4, we present the mean rotational error and translational error along with a 95% confidence interval across 10 trajectories and rollouts of 99 steps. This underscores ContactNets' ability to generate accurate trajectory predictions within the constraints of a relatively modest dataset.

TABLE I: Experiment Results and Ablation Studies

	Pose			Reconstruction
	ADD-S(%) ↑	ADD(%) ↑	SR(%) ↑	CD(cm) ↓ [a]
BundleSDF 1 toss [b]	65.02	58.08	40.17	34.20
BundleSDF 10 tosses [c]	64.89	32.42	32.10	1.75
Ours w/o cyclic pipeline [d]	65.02	58.08	40.17	1.18
Ours w/o ContactNets [e]	52.60	48.36	41.83	0.96
Ours w/o ICP [f]	65.43	60.45	48.20	0.74
Ours [g]	**69.50**	**68.40**	**49.86**	**0.18**

[a] ↑ means we prefer higher scores. ↓ means we prefer lower scores.
[b] Run BundleSDF on one video clip of 1 toss;
[c] Run BundleSDF on the full video of 10 tosses;
[d] Run ContactNets using poses from BundleSDF on 1 toss and geometry from BundleSDF on 10 tosses. Compared to the final result, it has no cyclic pipeline. Compared to the first baseline, it has an identical trajectory;
[e] Run BundleSDF on 10 tosses to get geometry, refine geometry via ICP, and run BundleSDF again on 1 toss using this geometry in a cyclic fashion. Compared to the final result, geometry is not refined by ContactNets;
[f] Run ContactNets using poses from BundleSDF on 1 toss and geometry on 10 tosses, and run BundleSDF again on 1 toss using geometry from ContactNets in a cyclic fashion. Compared to the final result, poses are not refined by ICP;
[g] Run ContactNets using poses from BundleSDF on 1 toss and geometry from BundleSDF on 10 tosses, align geometry from ContactNets via ICP, and run BundleSDF again on 1 toss using aligned geometry. Final result.

V. DISCUSSION AND FUTURE WORK

Our study introduces a robust framework that performs instance-agnostic geometry and contact dynamics learning. The cornerstone of our approach lies in integrating two crucial components: the BundleSDF [9] pose estimation and geometry reconstruction framework, in conjunction with the ContactNets [6] contact learning framework. The approach resolves the lack of motion capture data and shape priors for contact dynamics learning. Experiments with ablation studies demonstrate the effectiveness of our system. Moving forward, we are extending our preliminary efforts toward continuing the cyclic pipeline to convergence, unifying loss functions for joint optimizations, and encompassing a wider variety of novel and nonconvex objects in the open world.

References

[1] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[3] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4233–4241.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[5] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[6] S. Pfrommer, M. Halm, and M. Posa, "Contactnets: Learning discontinuous contact dynamics with smooth, implicit representations," in *Conference on Robot Learning*. PMLR, 2021, pp. 2279–2291.

[7] Y. Labbé, J. Carpentier, M. Aubry, and J. 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*. Springer, 2020, pp. 574–591.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

[9] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. 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*, 2023, pp. 606–617.

[10] Y. Liu, Y. Wen, S. Peng, C. Lin, X. Long, T. Komura, and W. Wang, "Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images," in *European Conference on Computer Vision*. Springer, 2022, pp. 298–315.

[11] K. Park, A. Mousavian, Y. Xiang, and D. Fox, "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 710–10 719.

[12] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8067–8074.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[14] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," 2022.

[15] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[16] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6498–6508.

[17] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9421–9431.

[18] B. Shen, Z. Jiang, C. Choy, L. J. Guibas, S. Savarese, A. Anandkumar, and Y. Zhu, "Acid: Action-conditional implicit visual dynamics for deformable object manipulation," *arXiv preprint arXiv:2203.06856*, 2022.

[19] Q. Le Lidec, I. Kalevatykh, I. Laptev, C. Schmid, and J. Carpentier, "Differentiable simulation for physical system identification," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3413–3420, 2021.

[20] P. Sundaresan, R. Antonova, and J. Bohgl, "Diffcloud: Real-to-sim from point clouds with differentiable simulation and rendering of deformable objects," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 828–10 835.

[21] K. M. Jatavallabhula, M. Macklin, F. Golemo, V. Voleti, L. Petrini, M. Weiss, B. Considine, J. Parent-Lévesque, K. Xie, K. Erleben *et al.*, "gradsim: Differentiable simulation for system identification and visuomotor control," *arXiv preprint arXiv:2104.02646*, 2021.

[22] S. Le Cleac'h, H.-X. Yu, M. Guo, T. Howell, R. Gao, J. Wu, Z. Manchester, and M. Schwager, "Differentiable physics simulation of dynamics-augmented neural objects," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2780–2787, 2023.

[23] H. K. Cheng and A. G. Schwing, "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model," 2022.

[24] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9226–9235.

[25] K. Duarte, Y. S. Rawat, and M. Shah, "Capsulevos: Semi-supervised video object segmentation using capsule routing," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8480–8489.

[26] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," *arXiv preprint arXiv:2002.10099*, 2020.

[27] B. Bianchini and M. Posa, "Simultaneously learning contact and continuous dynamics," 2023.

[28] R. Tedrake and the Drake Development Team, "Drake: Model-based design and verification for robotics," 2019. [Online]. Available: https://drake.mit.edu

[29] B. Pfrommer and K. Daniilidis, "Tagslam: Robust SLAM with fiducial markers," *CoRR*, vol. abs/1910.00679, 2019. [Online]. Available: http://arxiv.org/abs/1910.00679