DyTact: Capturing Dynamic Contacts in Hand-Object Manipulation

Xiaoyan Cong[1] Angela Xing[1] Chandradeep Pokhariya[2] Rao Fu[1] Srinath Sridhar[1*]
[1]Brown University [2]IIT Delhi

Section provides ablation study on our time-dependent deformation refinement module and contact-guided adaptive density control strategy. Section 2 introduces the details of our capture procedure for capturing new sequences for *DyTact*-21 benchmark. Section 3 discusses how we visualize the contact map. Section 4 describes the implementation details of DyTact. Section 5 shows more qualitative comparisons on novel view synthesis. Section 6 discusses limitations and future work. We kindly refer readers to our supplementary video for more results.

1. Ablation Study

Time-dependent Deformation Refinement Module. Qualitative comparisons in Figure 1 show that the refinement module $\mathcal{R}_\theta$ plays an important role in alleviating blurry artifacts around contacting regions caused by time-dependent high-frequency deformations of hand skin. We also observe consistent quantitative results in Table 1 (II) that removing $\mathcal{R}_\theta$ degrades the overall performance.

Contact-guided Adaptive Density Control Strategy. Figure 2 shows that the accuracy and coverage of contacts estimation improves by a large margin after introducing the contact-guided adaptive density control strategy. Table. 1 (I) indicates that the contact-guided adaptive density control strategy helps accumulate more gradients among contacting regions and provide an effective inductive bias for the densification process during optimization, which is important to improve the accuracy of the occluded areas.

	mIoU$^\uparrow$	SSIM$^\uparrow$	PSNR$^\uparrow$	LPIPS$^\downarrow$	F1 score$^\uparrow$
I. w/o CG	0.216	0.975	31.82	0.021	0.352
II. w/o $\mathcal{R}_\theta$	0.225	0.971	31.79	0.021	0.375
III. Full	0.226	0.978	31.88	0.020	0.378

Table 1. **Ablation studies** for the refinement module $\mathcal{R}_\theta$ and the contact-guided adaptive density control strategy (CG). Both components improve reconstruction accuracy.

*Corresponding author

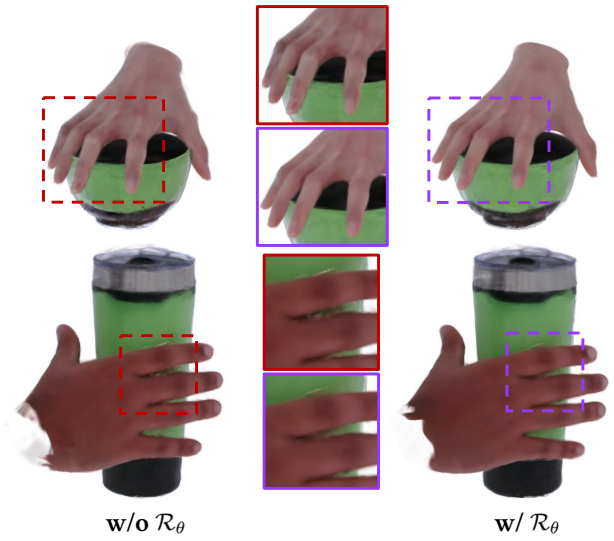

Figure 1. **Ablation studies for the refinement module** $\mathcal{R}_\theta$. $\mathcal{R}_\theta$ captures time-dependent deformation which reduces blurry artifacts around the contact region.

Figure 2. **Ablation studies on the Contact-Guided Adaptive Density Control Strategy.** This strategy effectively regulates the contact regions by allocating more isotropic gaussian surfels, yielding more accurate contact estimation.

2. Capture Procedure

Specifically, our capture procedure consists of three steps:

1. The object is wrapped in disposable shrink wrap, vacuum-sealed to ensure close surface conformity, and coated with wet paint designed to leave residue upon contact;
2. The subject wears a pair of disposable, tight-fitting, transparent gloves and performs the manipulation task with the painted object;
3. Following the interaction, paint residue is transferred to the gloves, providing clear visual evidence of hand-object contact. This residue serves as a physically

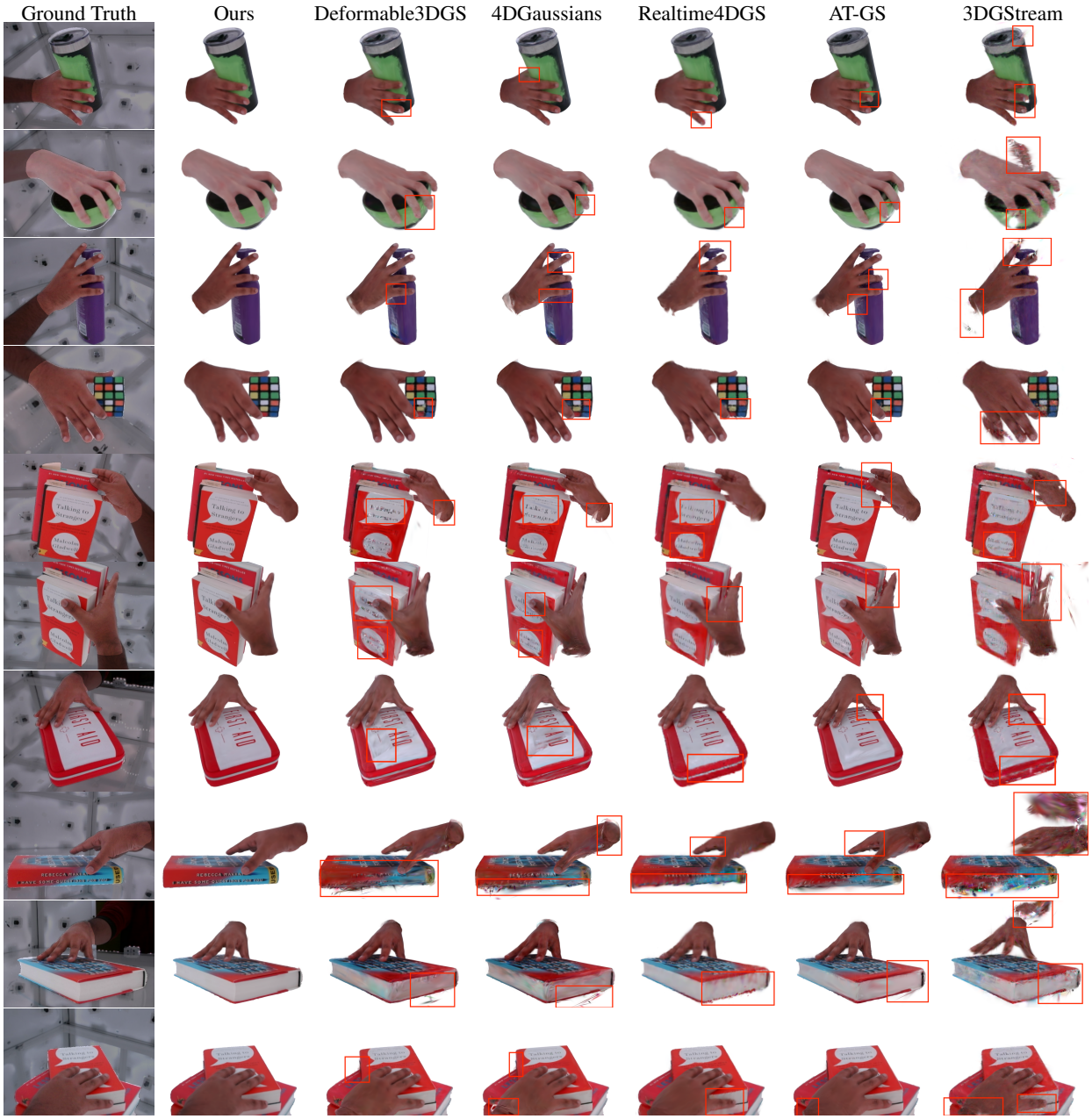

Figure 3. **Qualitative Comparisons on Novel View Synthesis.** Our method produces superior reconstruction quality with sharper novel view synthesis renderings, comparing with Deformable3DGS [10], 4DGaussians [8], Realtime4DGS [9], AT-GS [1], and 3DGStream [7]. DyTact delivers more fine-grained details, particularly in occluded regions and around edges, whereas baseline methods exhibit artifacts and blurriness. Please zoom in for better views.

grounded proxy for accumulated contact areas.

The setup can be easily reset between sequences by replacing the shrink wrap and gloves.

3. Contact Map Visualization

To estimate the contact maps, DyTact first estimates contacting Gaussian surfels and allocate green color to them. For DyTact and MANUS [5], we utilize the differentiable rasterizer for Gaussian Splatting [4] to render contact maps. For MANO [6] and HARP [3], we utilize Blender's emission renderer to render the contact maps. For fair comparisons, we increase the resolution of MANO and HARP vertices from 778 to 49,000 vertices by subdividing the meshes before estimating contacts.

4. Implementation Details

We train for $60,000$ iterations using the Adam optimizer [2]. During training, we optimize the parameters of Gaussian surfels, object poses, and MANO hand parameters. The learning rate for the positions of Gaussian surfels starts at 0.008 and decays exponentially to 0.000008 by the final iteration. The learning rate for surfel scaling is set to 0.05, while the remaining primitive parameters use the same learning rates as those in 3D-GS [4]. In addition to Gaussian surfel parameters, we finetune the MANO parameters and object tracking poses. For the MANO parameters, the learning rates are set to 0.001 for pose and shape, and 0.0001 for relative rotation and translation. For object pose parameters, the learning rate starts at 0.005 and decays exponentially to 0.00005 by the final iteration. We enable the contact-guided adaptive density control every 200 iterations, starting from iteration 1,000 and continuing until the end of training. Additionally, we reset the opacities of Gaussian surfels every 6,000 iterations to ensure stable updates. All the experiments are conducted on a single RTX A6000 GPU.

5. Results on Dynamic Reconstruction

In Figure 3, we show more qualitative comparisons of dynamic reconstruction with five baselines Deformable3DGS [10], 4DGaussians [8], Realtime4DGS [9], AT-GS [1], and 3DGStream [7]. DyTact exhibits a more 3D consistent and detailed reconstruction of dynamic manipulation scenes, especially around the contact regions.

6. Limitations & Future Work

While this paper primarily focuses on accurate dynamic contact estimation, we acknowledge that the full complexity of everyday hand and object dynamics vastly exceeds the scope of our current investigation. Our work is capable of modeling two hands manipulating rigid objects, thereby deferring the challenges posed by articulated or more general object types to future research. Furthermore, since the evaluation of DyTact relied on an indoor multi-view capture system, exploring dynamic contact modeling for bi-manual manipulation in outdoor environments or under sparse-view conditions presents a key avenue for future work. We also see potential for developing more comprehensive evaluation metrics for dynamic contacts in future works. Finally, the wet-paint technique employed in this study captures only ground-truth accumulated contacts. A crucial direction for future research is therefore the development of new strategies to accurately and efficiently capture ground-truth dynamic (or instantaneous) contacts in a scalable and cost-effective manner.

References

[1] Decai Chen, Brianne Oberson, Ingo Feldmann, Oliver Schreer, Anna Hilsmann, and Peter Eisert. Adaptive and temporally consistent gaussian surfels for multi-view dynamic reconstruction. *arXiv preprint arXiv:2411.06602*, 2024. 2, 3

[2] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 3

[3] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12802–12813, 2023. 2

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[5] Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless grasp capture using articulated 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2197–2208, 2024. 2

[6] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2

[7] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20675–20685, 2024. 2, 3

[8] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2, 3

[9] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2, 3

[10] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20341, 2024. 2, 3