Betsu-Betsu: Multi-View Separable 3D Reconstruction of Two Interacting Objects

Supplementary Material

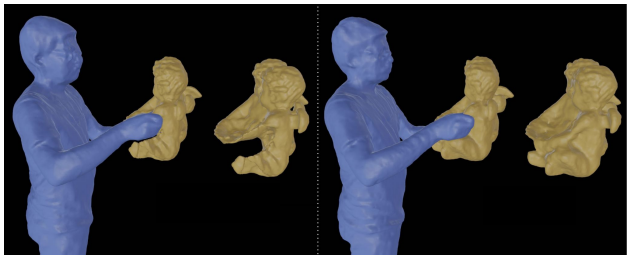

Figure 10. Naïvely reconstructing the human and the object SDFs using separate NeuS2 [40] reconstruction leads to extreme geometric artefacts due to occlusion.

In this document, we provide more information regarding our implementation in Sec. 7, more details about our dataset in Sec. 8, a possible extension to the method by incorporating template priors in Sec. 11.1 and discussion about improvements of our method over ObjectSDF++ and 'Segmented Neus2' in Sec. 11.2 and Sec. 11.3 respectively. We also present additional qualitative comparisons in Sec. 13 and Sec. 14.

7. Implementation Details

Scene Encoding: The sampled point positions are encoded using the hashgrid encoding, $h(\mathbf{x})$, with $L = 18$ levels, two features per level and use a hashmap of size 2^{19}. We set the base resolution to 16 and the highest resolution to 8192. Following [24], we maintain an occupancy grid of resolution 128 and skip the empty space while ray marching, whenever the opacity is below 10^{-4}. The view direction $\mathbf{v}$ is encoded as spherical harmonics up to degree 4. We also use per-image latent of size 8, to account for slight color variations in zoomed-in camera views.

MLPs: The MLPs for human SDF Φ_h, the object SDF Φ_o and the feature extractor consist of two layers with 64 neurons each. The colour MLP $\mathcal{C}_s$ is also a 2-layer MLP but with 128 neurons each.

Sampling: We sample rays for each image in two ways: (1) from pixels within the segmentation masks, and (2) randomly from any pixel in the image. The probability of sampling rays from the masks is progressively increased (as training progresses), from 0.1 to 0.8, linearly increasing from steps 0 to 5000. From steps 0 to 5000, we sample equally from both the human and the object, and after step 5000, we sample randomly from the whole foreground mask.

Training: We train our method for $10k$ steps for each scene, which takes ≈30–45 minutes on a single A40 GPU.

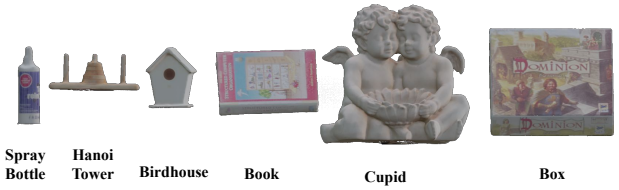

Figure 11. We capture human interactions with six objects of varying intricacy (book vs Hanoi tower) and scale (spray bottle vs sculpture). Yet, the overall scale of the objects remains comparably small

Changes to ObjectSDF++: We increase the total number of hashgrid levels, hashmap size, and resolution to match our implementation, as explained above. ObjectSDF++ also uses depth and normal supervision, since they show their method on indoor scenes. As we do not use either of them, for a fair comparison, we set the normal and depth loss weights to 0. Apart from these, we retain all the other hyperparameters as it is in their implementation. To complete training on one scene, ObjectSDF++ takes around 12–14 hours on a single A40 GPU.

Color MLP: Rather than using a single colour MLP, another option would be to use two separate colour MLPs for each object. But by doing so, each colour network has the freedom to learn the background (or the other object) as colour (black), instead of relying on opacity to give the accumulated colour as 0, in Eq. (7). Instead, using a single colour MLP ensures that for any position that is occupied by either of the objects, the colour network predicts the correct colour, but overall accumulation depends on the corresponding object opacity being one, and the other opacity being zero.

Segmentation masks: We obtain segmentation masks for the human-object, human-human and object-object (WildRGBD) datasets using a pipeline of GroundingDINO [19] and Segment-Anything [15] implemented in [21]. We further add a CLIP [29] similarity based filtering, when multiple masks are predicted. Since hand-object (AffordPose [12]) is synthetic dataset, we get the ground-truth segmentation masks while rendering the meshes.

8. Human-Object dataset

Our new human-object dataset consists of 3 different people each with 6 objects shown in Fig. 11. Each scene consists of a maximum of 120 views (some views might be removed because of bad segmentation) with many scenes containing

views zoomed into the occupied area. More specifically, for subject 0, all scenes except 'Cupid' have only normal views, and for subject 0 'Cupid' scene, as well as all scenes of subjects 1 and 2, have 19 zoomed-in views. Irrespective of the zoom, all images have been cropped to a resolution of 1200×1600 px.

9. Limitations

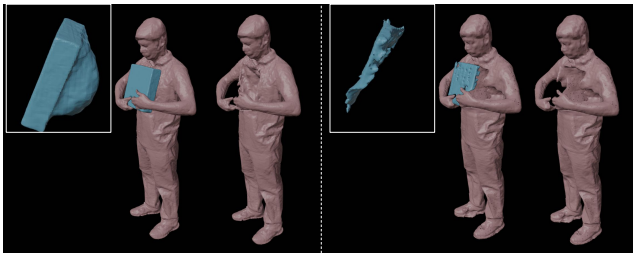

Figure 12. Failure case: under heavy occlusion, our method generates extended, yet separate geometries. Segmented NeuS2, on the other hand, reconstructs the scene with holes.

Our method is designed to separately reconstruct two interacting objects. While this is largely addressed by encoding both geometries in a shared hash grid and ensuring the opacities are disjoint, some artefacts remain. Consider the case in Fig. 12: The object's face towards the body is occluded beyond observation. Hence, the method does not have sufficient prior to disambiguate the human-object boundaries. Yet, it ensures that the boundaries are separate. A potential solution to this problem would be incremental training, as shown in Fig. 13, or fine-tuning the joint SDF with a pre-scanned template providing a useful geometric prior, assuming it is available. Similar refinement can be done for the human; see the discussion in Sec. 11.1 Another limitation is in cases of thin gaps between different structures, as in the hand fingers of the *Bag* scene shown in Fig. 16. Sometimes, we observe undesired *bridges* between such thin gaps due to the nature of ray sampling during optimisation.

10. Consecutive Frame-by-Frame Reconstruction

Our approach can also be applied on consecutive frames by initialising the parameters for the current frame from the previous frame. This also helps prevent certain defects, as the model has prior on the shapes observed before occlusions. One such example is presented in Fig. 13, which improves the failure case Fig. 12. Fig. 13 shows normal renderings for a few sampled time steps.

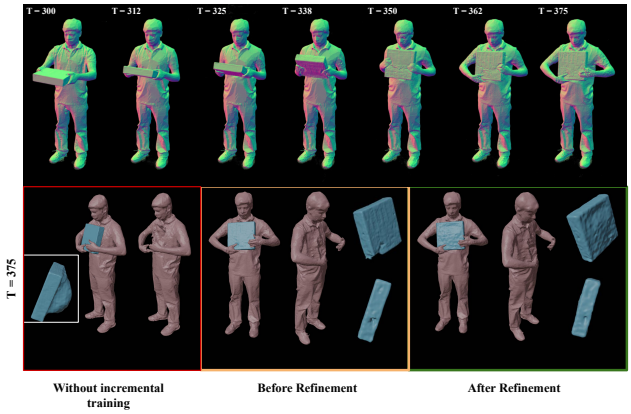

Figure 13. Improvements in the 3D object geometry using incremental training. We observe that the largest erroneous object deformations caused by heavy occlusions are mostly corrected using incremental training.

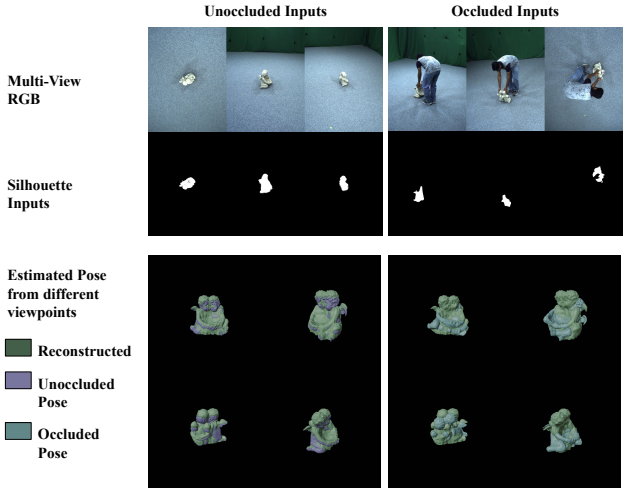

Figure 14. Comparison of occluded and unoccluded 6DOF fitting a template using Silhouette loss (initialised with known position). The orientation of the fitted template is compared against the reconstructed mesh.

11. Additional Discussion

11.1. On Object Templates

In this work, we assume that an object template is not available. While having a template, arguably, would make the task easier in an alternative setting, it would also substantially limit the method's applicability and extendability to scenarios with arbitrary objects. It would also necessitate the additional step of template acquisition, which can be infeasible in many downstream applications. Moreover, articulated objects like laptops would require a different approach to 3D reconstruction, even when the template in a canonical pose is available; similar observations apply to humans. Hence, we focus on modelling two-object interac-

tions at a fundamental level which can, if required, be extended if the template is available. We next briefly discuss several considerations in this regard.

Suppose an object template is available. How could our approach be extended or adjusted to account for this prior knowledge? A naive way would be to fit the object template to the image observations using 6DoF optimisation. This is, however, suboptimal since (1) the colour rendering loss cannot be used because the lighting conditions at the time of template acquisition and 6DoF optimisation would be likely different; (2) for the same reason as mentioned above (i.e. since the template appearance is likely to differ during the template acquisition and the main scene capture steps), the globally optimal 6DoF template pose could be inaccurate; and (3) the silhouette-based optimisation would also struggle as the objects are under severe occlusion, and the segmented silhouettes are not reliable as shown in Fig 14.

A better alternative would be to fit the template pose coarsely to the scene and use the posed template to sample the points for volume rendering. This is akin to the *canonicalised* representation in several 3D human and non-rigid reconstruction works [18, 27, 28, 37]. While this would improve the convergence speed, such an approach would still benefit from the shared representation and alpha-blending loss proposed in this work.

Another potential alternative would be instead to fit the object template using the object's reconstructed surface SDF. The optimisation would be performed in two alternating steps until convergence, i.e. (step 1) using the object SDF to update the template's pose and (step 2) using the optimised template pose to refine the joint scene geometry with the help of our method to alleviate penetration artefacts. Indeed, we observe that this joint optimisation improves scene reconstruction, especially in the case of human-object interaction. As shown in Sec. 11.1, the hand geometry benefits from template-guided optimisation using this iterative refinement policy.

11.2. On ObjectSDF++

ObjectSDF++ is the closest work to our proposed method. There are, however, two key differences that allow us to outperform ObjectSDF++ across multiple evaluation settings. First, ObjectSDF++ uses a single(shared) MLP for all SDF outputs, whereas we model each SDF with a separate MLP. This introduces a tradeoff – better reconstruction and separation quality at the cost of the ability to model multiple-objects. This is also confirmed by the ablations presented in the main draft. It is noteworthy that our solution can also be extended to multiple-objects by having multiple MLPs, in-theory. This would require alpha-regularization on all combinations and is a direction for future exploration. Second, ObjectSDF++ proposes a ReLU-based regularizer which, we hypothesize, is harder to optimize. This is evident in Fig. 5 and Fig. 7 (Main) wherein some ObjectSDF++ reconstructions have deeper deformations near contact regions than ours. Finally, ObjectSDF++ is additionally trained using depth and normal maps whereas we do not need such supervision.

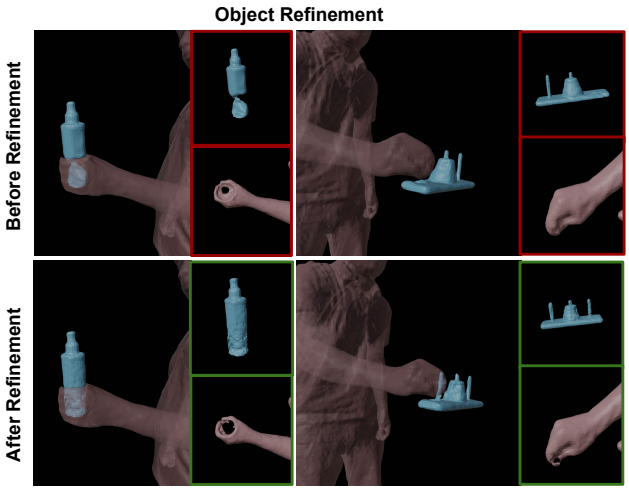

Figure 15. Illustrations of the object's geometry before and after the template-guided refinement. Notice that the peg of the tower, missing in the first state, re-emerges after jointly optimizing with the template. Interestingly, jointly optimizing with the template also improves hand reconstruction, as can be seen in the case of spray holding.

11.3. Reason for better reconstruction compared to Segmented Neus2

NeuS(2) is sensitive to occlusions in the input. Occlusion in one view implies all rays along the entire path hit blank space, whereas other views indicate the same space is non-empty. This mismatch leads to incorrect optimisation in the form of artefacts and holes. On the other hand, by jointly encoding and rendering both geometries through a shared hashgrid, we can maintain multi-view consistency, since the presence of one object explains the absence of the other object from a particular viewpoint. This is further improved by introducing alpha-regularization, which prevents penetrations.

12. Experiments (continued)

12.1. Hand-Object

We show the qualitative comparison for hand-object scene geometry reconstruction in Fig. 16.

12.2. Human-Human Appearance Evaluation

We also show qualitative comparison for novel-view synthesis on human-human interaction scenes in Fig. 17 and quantitative comparison in Sec. 12.2.

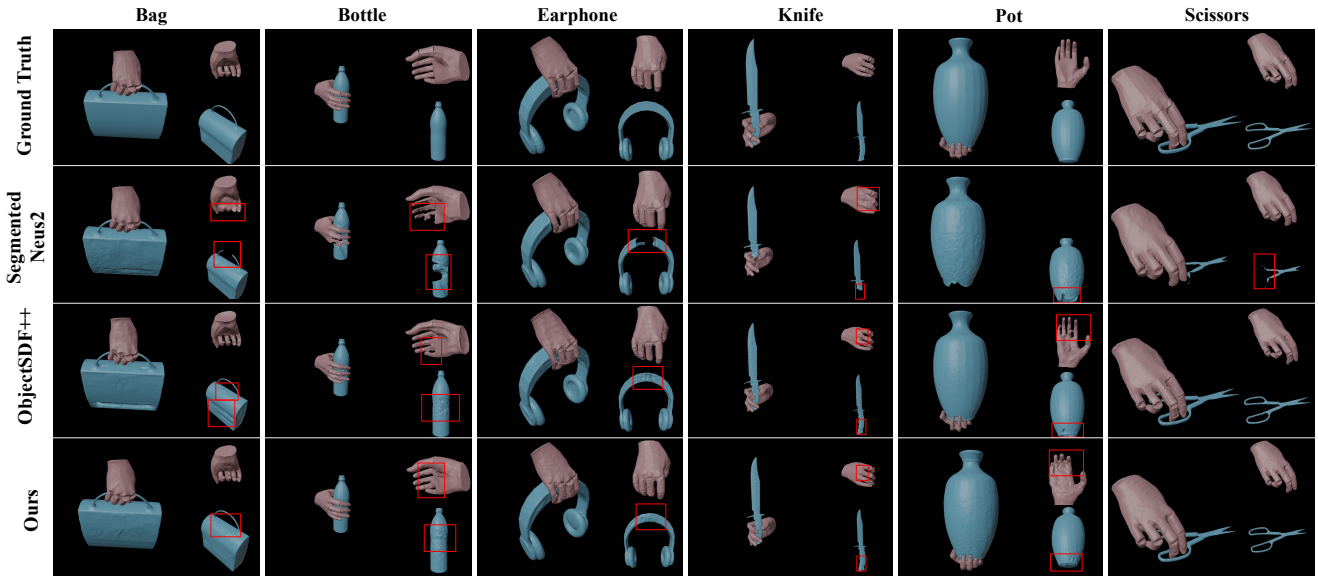

Figure 16. Qualitative comparison on the AffordPose dataset. The regions highlighted in red indicate apparent differences in the reconstructions. In the case of Segmented Neus2, for the pot scene, reconstruction of the hand fails (hence not shown). **Best viewed when zoomed**

Seq ID	PSNR↑		SSIM↑		LPIPS↓	
	Object SDF++	Ours	Object SDF++	Ours	Object SDF++	Ours
1	25.84	**30.50**	0.92	**0.94**	0.14	**0.13**
2	29.12	**31.90**	**0.95**	**0.95**	**0.12**	0.13
3	24.20	**28.66**	0.91	**0.92**	0.18	**0.17**

Table 6. Quantitative comparison of our method with the ObjectSDF++ on the novel-view synthesis task of the Human-Human scenes.

12.3. NeuralDome results

We show a qualitative comparison on a human-table interaction scene from NeuralDome [49] dataset. While NeuralDome uses pre-scanned template of the object, along with markers, our method can obtain a similar reconstruction quality using only multi-view images.

13. Comparison against VolSDF

Recall that we show results with "Segmented NeuS2" by training two different NeuS2 models for the human and object. This makes the geometric reconstruction of the object agnostic of the presence of a human and vice-versa. In order to confirm that the failure of reconstruction for the "Segmented Neus2" is not just because of NeuS [39] formulation, we also use VolSDF [46], and train it in isolation for human and object, by providing the respective masks. We show the results on the two biggest objects in our evaluation dataset, i.e. Box and Cupid statue, in Fig. 19. While human reconstruction works rather well, the box is not reconstructed and the cupid is reconstructed poorly. This demonstrates, yet again, that the baseline approach of two isolated reconstructions is suboptimal and that sharing the scene parameters is crucial to separable reconstruction.

14. Comparison against ObjectSDF

We also compare against ObjectSDF [42] (which was the predecessor to ObjectSDF++) for a few scenes and show the qualitative results in Fig. 20. Note that ObjectSDF inaccurately assigns large parts of the book to the human (in red). The actual book (in blue) is poorly reconstructed.

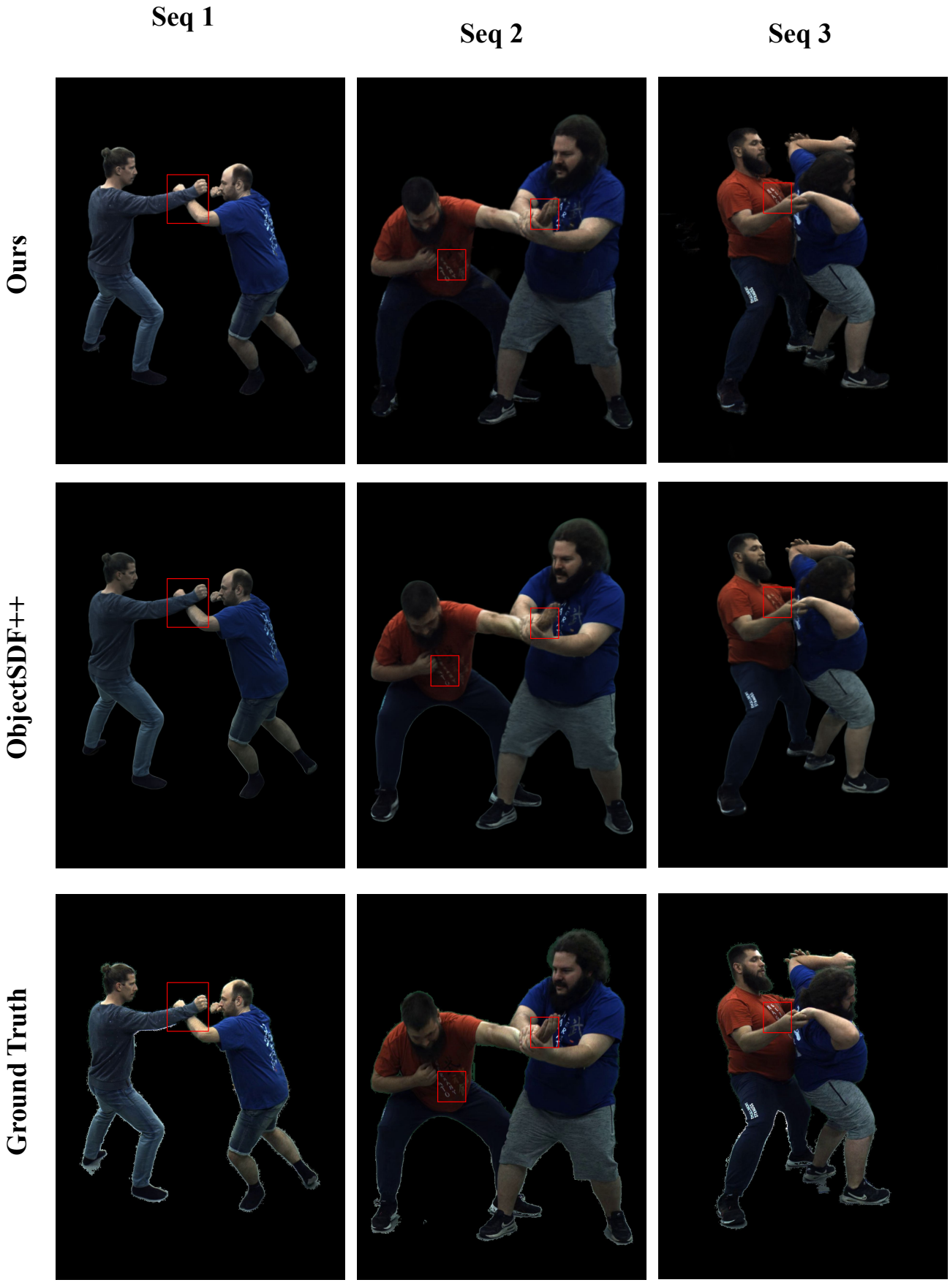

Figure 17. Qualitative comparison with ObjectSDF++ of human-human interaction novel-view synthesis

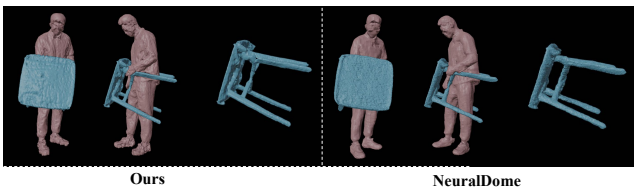

Figure 18. Reconstruction comparison with a scene from the NeuralDome dataset. NeuralDome provides a pre-scanned template of the object and a multi-view 3D reconstructed human.

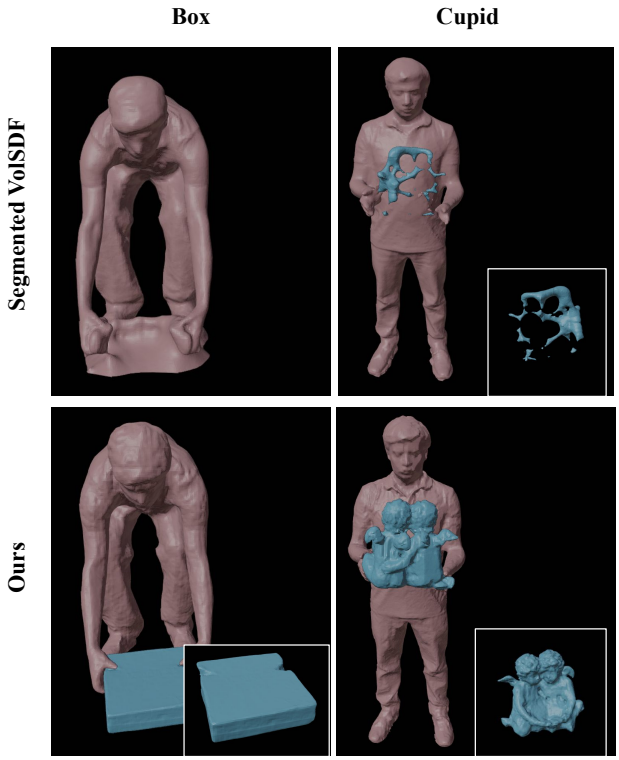

Figure 19. Qualitative comparison with Segmented VolSDF. We observe that similar to the Segmented NeuS2, the object is not reconstructed reasonably.

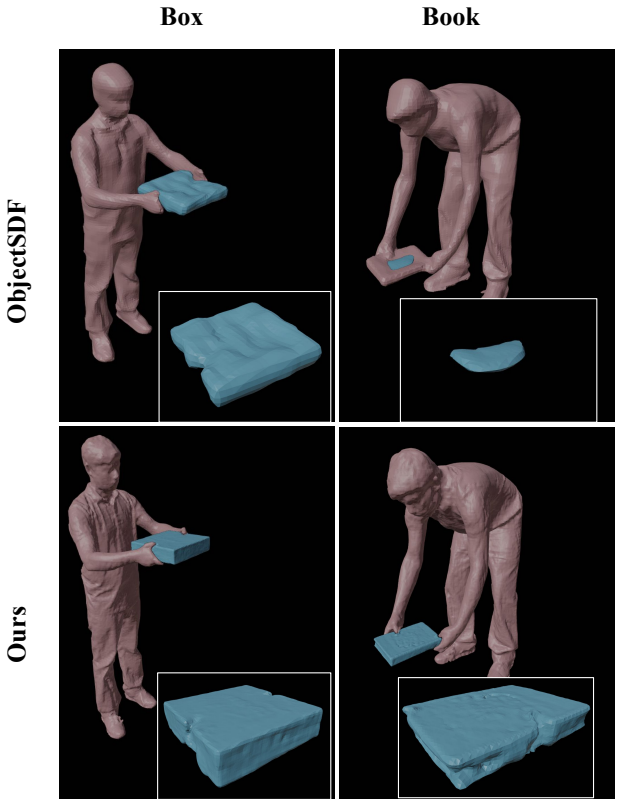

Figure 20. Qualitative comparison with ObjectSDF. We observe that our reconstruction results are much more detailed and well separated, whereas ObjectSDF produces incorrect geometry for the object, especially for the book.