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

Suhas Gopal[1] Rishabh Dabral[2] Vladislav Golyanik[2] Christian Theobalt[2]
[1]Saarland University, SIC [2]Max Planck Institute for Informatics, SIC

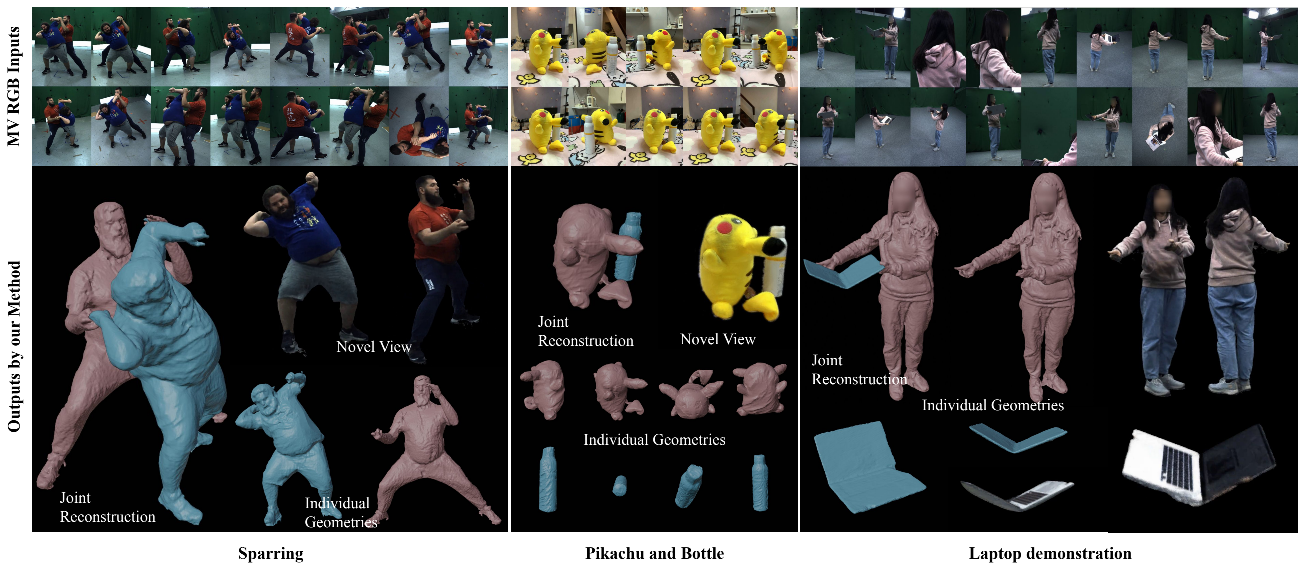

Figure 1. **Our method reconstructs humans and objects in 3D from segmented multi-view (MV) RGB images (top) in a separable way**, i.e. with clean boundaries and no inter-penetration. (Bottom:) For each of the three scenes (*Sparring*, *Pikachu* and *Laptop Demonstration*), we show the two join recovered geometries (left), individual novel view renderings (top right) and individual geometries.

Abstract

Separable 3D reconstruction of multiple objects from multi-view RGB images—resulting in two different 3D shapes for the two objects with a clear separation between them—remains a sparsely researched problem. It is challenging due to severe mutual occlusions and ambiguities along the objects' interaction boundaries. This paper investigates the setting and introduces a new neuro-implicit method that can reconstruct the geometry and appearance of two objects undergoing close interactions while disjoining both in 3D, avoiding surface inter-penetrations and enabling novel-view synthesis of the observed scene. The framework is end-to-end trainable and supervised using a novel alpha-blending regularisation that ensures that the two geometries are well separated even under extreme occlusions. Our reconstruction method is markerless and can be applied to rigid as well as articulated objects. We introduce a new dataset consisting of close interactions between a human and an object and also evaluate on two scenes of humans performing martial arts. The experiments confirm the effectiveness of our framework and substantial improvements using 3D and novel view synthesis metrics compared to several existing approaches applicable in our setting[1].

1. Introduction

The world we live in is compositional. A typical office desk, for example, would consist of a monitor, a keyboard, a few cups of coffee, mobile phones, and so on. Needless to say, we rarely encounter scenes comprising of one and only one object. Yet, most 3D reconstruction research [11, 20, 25, 33, 35, 40] has focused on scenes with only one object (e.g. the famous caterpillar scene). When more than one object is present in the scene (e.g. the GTA Truck scene), the compositionality of the scene is ignored and the entire scene is reconstructed jointly.

Recent works that addressed the challenge of compositional scene reconstruction have either used object tem-

[1]Project page: https://vcai.mpi-inf.mpg.de/projects/separable-recon/

plates [2, 5, 30, 43], or parametric models of humans or hands [9, 41, 42]. A few works that propose a *generalised* solution [36, 37] suffer from inter-penetration of the two or more interacting geometries. In this work, we focus on the generalised compositional reconstruction setting (as shown in Fig. 1) while mitigating the penetration artefacts of the existing literature. This mandates addressing the challenges posed by severe occlusion of the objects during interaction (e.g. a person holding a cup), as well as accounting for the difference in object scales while sampling.

With these considerations in perspective, we propose a new markerless, template-free approach for compositional 3D reconstruction of arbitrary objects undergoing interactions in a scene observed from multiple views. We represent the object geometries as separate Signed Distance Fields (SDFs) and the appearance with the corresponding Neural Radiance Fields. The 3D scene is encoded jointly for the objects by using a *shared* multi-resolution hashgrid [21] which can be decoded into separate SDFs of the target objects (e.g. a hand and a book). Crucially, we propose a novel alpha-blending loss which enforces that a point lying inside one object has a high opacity only for the corresponding object's SDF while suppressing the opacity for the other. This incentivises clean separation boundaries and reduces the penetration volume between the two SDFs, even if the queried point is poorly observed (e.g. due to occlusion).

To demonstrate the effectiveness of our method, we capture a new real-world dataset consisting of several scenes of human-object interactions. For this, we ask the subjects to naturally interact with various small and mid-sized objects in a large capture dome. In summary, the technical contributions of this paper are as follows:

- A novel markerless and category-agnostic approach for high-quality 3D reconstruction of two interacting objects from multi-view RGB inputs;
- A shared neuro-implicit representation that can be jointly optimised for the geometries of interacting objects while also supporting separable free-viewpoint rendering;
- An interaction-aware alpha-compositing of opacity values for each SDF enforcing clean separation boundaries and mitigating inter-object penetration in 3D;
- A new multi-view dataset for human-object interactions.

In addition to the captured dataset, we also evaluate our method on the publicly available WildRGB-D [38] (object-object) and AffordPose [12] (hand-object) datasets, and demonstrate its effectiveness on marker-based 3D reconstruction datasets like NeuralDome [43]. We also evaluate human-human interactions on scenes of the ReMoCap [7] dataset. These scenes involve practitioners performing martial arts poses, thereby leading to challenging interactions. The proposed approach performs better than previous state-of-the-art methods such as ObjectSDF++ [37] and NeuS2 [34]. Although not the main objective of this work, we also observe better performance on the related task of segmented novel-view synthesis [20, 21].

2. Related Works

We discuss the related works from three perspectives: (1) neural scene representations, (3) implicit models for multi-object segmentation and reconstruction and, finally, (3) generic human-object, hand-object and human-human interaction works.

2.1. Neural Scene Representations

Recent advances in neural implicit representations and NeRF-based techniques [20, 31] have enabled high-quality novel view synthesis and reconstruction of complex scenes from multi-view images. Extensions of those for surface reconstruction by [22, 33, 40] and enhancements in speed by works like [21, 25, 34] have shown that it is possible to reconstruct high-quality geometry, in a reasonable amount of time, such that it can be applied to even short videos on per-frame basis. They have also been applied for human rendering [17, 24, 29, 35, 45], that extend to dynamic scenes as well as provide pose-conditioned animation capabilities. Some works also extend it to multi-person scenarios [11, 19, 27]. The most relevant work to ours [28, 43] also uses implicit representation for the reconstruction of human-object interaction. They both use an SMPL prior [18, 23] for the human body and an object template for a layer-wise representation. HOI-FVV [28] uses sparse-view RGB input to predict occupancy values of the human undergoing interaction with objects. However, the sparse inputs limit the reconstruction quality, and the object geometries have to be tracked assuming a template is available. On the other hand, Zhang et al. [43] use dense RGB inputs, and obtain separate NeRFs for both human and object, using SMPL and an object template as priors, which are then blended to obtain the final reconstruction. In contrast, we implicitly learn to separate the two objects by using only the image segmentation masks as additional supervision.

While similar to Neus2 [34] in employing hashgrid encoding with NeuS, we do not adopt their approximate second-derivative formulation. Instead of the coarse-to-fine training strategy, which caused missing reconstructions for small objects, we optimize all hashgrid levels from the start. Additionally, we condition the separate SDF MLP heads on hashgrid feature vectors derived via another MLP.

2.2. Multi-Object Segmentation/Reconstruction

While most methods focus on entire scene reconstruction, some methods [36, 37, 46] focus on individual objects in the scene. Semantic NeRF [46] uses a semantic head to predict labels for each position, supervised by the segmentation mask. A similar idea is also used in [9], where the method predicts a label for each position and then uses it to separate the SDF of the human and the object. However, this approach—while recovering correct labels at the surface—produces incorrect labels *within* the surface, making it feasible only for minimal occlusion and contact. The approach closest to ours is ObjectSDF++ [37], which predicts different SDFs for each object in the scene. Whereas they supervise the SDF separation using only opacity, we use the separately rendered colour of the two objects for supervision. Importantly, even though ObjectSDF++ proposes an extra SDF distinction regularisation term, it does not guarantee non-penetrating geometries. In contrast, we use an opacity regularisation term that incentivises the two SDFs to have disjoint opacities, thereby resulting in no (in most cases) or minimal interpenetration. Please refer to Sec. 4.1 for more details.

Human-Object Interaction A widely arising scenario is human-object interaction, which our method can also handle since it is applicable to arbitrary objects. Human-object interaction has been extensively studied in the literature. While some previous works [3, 4, 6, 8, 15, 26, 47] focus only on hands interacting with objects, many of the recent works [2, 5, 10, 13, 14, 16, 30, 32, 39, 43, 44] consider whole body interacting with the object. In many scenarios of humans represented by entire bodies interacting with objects, the latter are often substantially smaller than humans. Methods falling into this category can be broadly divided into two groups: Methods based on template fitting or multi-view reconstruction methods. For template fitting, most methods, leverage SMPL or SMPL-X [18, 23], along with a pre-acquired template of the object to fit marker-based motion capture data [5, 30] or (sparse) multi-view RGB-D data [2, 10]. Some methods even attempt to fit templates to a single RGB image [39, 44]. On the other hand, multi-view reconstruction methods, most relevant to our work, can recover accurate geometry and high-quality textures using multi-view RGB(D) images. Similarly to our method, NeuralDome [43] and Neural-HOFusion [14] use segmentation masks to separate humans and objects from multi-view images. However, in contrast to our work, both use a layer-wise NeRF representation that reconstructs humans and objects in isolation and then fuses them to recover final outputs. In contrast, our formulation uses a unified density and color field for the whole scene, but decodes separate geometries (as Signed Distance Fields) for the human and object. NeuralDome [43] uses SMPL-X and object template as prior for tracking. NeuralHOFusion [14] also uses object templates for better reconstruction. Note that our method does not require any 3D object templates free per default (the supplement discusses the case when one is available).

3. Background

Our method is based on an implicit surface representation and uses volumetric rendering for image supervision. It, therefore, builds on top of existing surface representation and volume rendering methods like NeuS [33] and Instant-NGP [21]. We briefly discuss them below.

3.1. Neural Implicit Surfaces

NeuS [33] is an implicit multi-view reconstruction method that extends NeRF [20] by representing the surface and appearance of a scene as a Signed Distance Function (SDF), $\Phi(\mathbf{x}) : \mathbb{R}^3 \to \mathbb{R}$ and a radiance field $c(\mathbf{x}, \mathbf{v}) : \mathbb{R}^5 \to \mathbb{R}^3$, respectively. The surface $\mathcal{S}$ is defined by the zero level set of the SDF, $\mathcal{S} = \{\mathbf{x} \in \mathcal{R}^3 | \Phi(\mathbf{x}) = 0\}$ and the pixel colour is obtained by volumetrically rendering the colours $\hat{C}$ along the ray, $\mathbf{p} = \mathbf{o} + i \cdot \mathbf{v}$, shot from the camera's origin $\mathbf{o}$ in the direction $\mathbf{v}$, through pixel p using the rendering equation:

$$\hat{\mathbf{C}}(p) = \sum_{i}^{N} T_i \alpha_i c_i, \tag{1}$$

where T_i is the accumulated transmittance, α_i is the opacity and c_i is the colour of i^{th} sample along the ray. α_i and T_i are obtained directly from Φ using

$$\alpha_i = \max\left(\frac{\sigma(\Phi_i) - \sigma(\Phi_{i+1})}{\sigma(\Phi_i))}, 0\right), \tag{2}$$

where $\sigma(\Phi_i) = (1 + e^{-\beta \Phi_i})^{-1}$ is the sigmoid function and β is a learnable parameter.

We extend this formulation for two SDFs and, thereby, opacities of two interacting objects (see Sec. 4).

3.2. Hashgrid Encoding

Training NeuS in the originally proposed fashion is slow and requires hours. Recent works [21, 25, 34, 45] accelerate it using multi-resolution hash grid encoding of 3D points. Müller *et al.* [21] as first proposed a multi-resolution voxel grid such that the grids of different resolutions are represented by a hash table that maps a 3D point $\mathbf{x}$ to a learnable feature vector $h_l(\mathbf{x})$, with l being the resolution level. All the feature vectors are concatenated to obtain the hash-encoded feature as $h(\mathbf{x}) = \{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_L(\mathbf{x})\}$, where L is the number of resolution levels. This representation (along with the CUDA implementation), speeds up the training substantially by three orders of magnitude and has been used to accelerate several surface reconstruction methods [34, 45]. We also adopt it in our method.

4. Method

Our goal is to separately recover the 3D geometry and the appearance of each object in the scene from multiple RGB

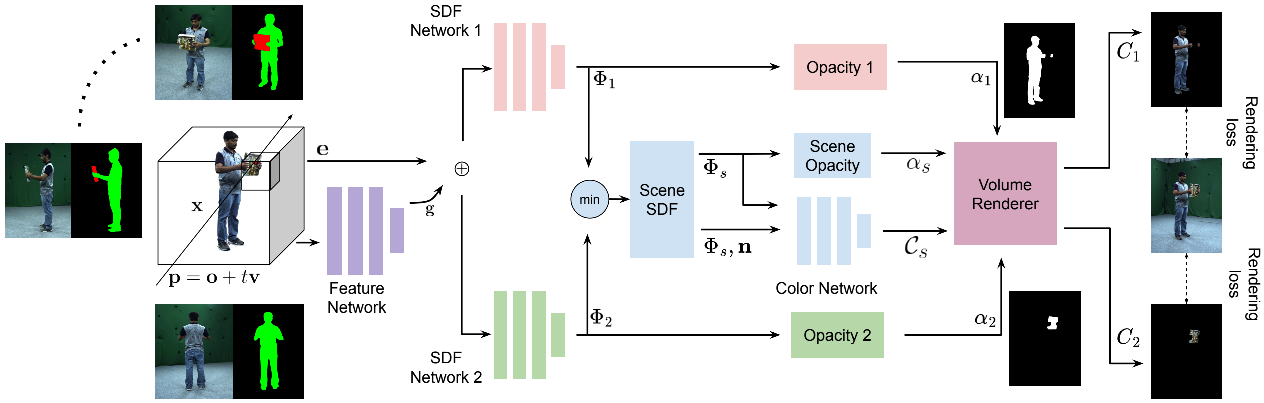

Figure 2. **Schematic overview of our framework.** We semantically segment the input multi-view images into the background and the areas corresponding to two interacting objects. The scene is encoded using a shared, multi-resolution hash grid encoding $\mathbf{e}$ and the shared features are decoded using two separate SDF MLPs to produce corresponding SDFs Φ_1 and Φ_2. The per-point colour $\mathcal{C}_s$ is estimated from the joint scene SDF composed using $\Phi_s = \Phi_1 \cup \Phi_2$. Finally, we integrate the colours of the sampled points in the ray by α-blending the individual opacities, α_1 and α_2, ensuring clean separation boundaries between the two (see Eq. (7)). The entire framework is supervised using the rendering loss and additional regularisers (see Eq. (10)).

views. With close interactions (leading to severe mutual occlusions), the key challenge is to recover a clean surface boundary while preventing inter-penetrations. We address the problem with a new neural approach illustrated in Fig. 2. The scene observed from multiple views is first encoded using a shared multi-resolution hash grid. In the second step, the scene features are decoded as two SDFs using two MLP heads, one for each interacting object (Sec. 4.1). Next, the individual opacity of each object, the overall scene opacity and the scene colour for each point in the ray are forwarded to the volume renderer that renders the images. The separation boundaries between different objects are obtained using ray colour integration with α-blending (Sec. 4.2). We next discuss each step in detail.

4.1. Scene Representation

We are given a set of K calibrated multi-view images $\mathcal{I} = \{I_i\}_{i=1}^{K}$ capturing two interacting objects (subscripted with 1 and 2) along with their corresponding segmentation masks $\mathcal{M}_1 = \{M_1^1, M_1^2, \ldots, M_1^K\}$ and $\mathcal{M}_2 = \{M_2^1, M_2^2 \ldots, M_2^K\}$. One can recover the foreground mask as their union: $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$. A naïve way to perform 3D reconstruction would be to use the set of masks corresponding to each object in isolation in an attempt to recover the corresponding surfaces using a multi-view reconstruction method such as NeuS [33] or VolSDF [40]. Unfortunately, such a solution is sub-optimal as it does not jointly account for all the densities, resulting in large gaps due to occlusion and poor separation boundary as shown in **??** and Tab. 1.

Hence, our approach uses a *shared* hash-encoding for both objects in the scene. For any 3D point $\mathbf{x}$—encoded using a *shared* multi-resolution hashgrid as $\mathbf{e} = (\mathbf{x}, h(\mathbf{x}))$—we estimate the signed distance to the two objects in the scene using two separate SDFs, Φ_1 and Φ_2, respectively, parameterised using two separate MLP heads. As shown in Fig. 2, the feature extraction network provides hashgrid features $\mathbf{g} \in \mathbb{R}^{d_g}$ for each encoded position $\mathbf{e}$. These features, along with the hashgrid encodings are the input to the SDF MLPs that produce $\Phi_1(\mathbf{e}, \mathbf{g})$ and $\Phi_2(\mathbf{e}, \mathbf{g})$. The SDF of the joint scene, $\Phi_s(\mathbf{e}, \mathbf{g})$, can now be composed using:

$$\Phi_s(\mathbf{e}, \mathbf{g}) = \min(\Phi_1(\mathbf{e}, \mathbf{g}), \Phi_2(\mathbf{e}, \mathbf{g})). \tag{3}$$

We also recover the colour of the scene, $\mathcal{C}_s(\mathbf{x}, \mathbf{v}, \mathbf{n}, \Phi_s, \mathbf{g}) : \mathbb{R}^{38} \rightarrow \mathbb{R}^3$, which can be encoded as an MLP and conditioned on the position $\mathbf{x} \in \mathbb{R}^3$, spherical-harmonic encoded view direction $\mathbf{v} \in \mathbb{R}^{16}$, the surface normal $\mathbf{n} \in \mathbb{R}^3$, SDF value $\Phi_s \in \mathbb{R}$ and the hashgrid features $\mathbf{g} \in \mathbb{R}^{15}$. For brevity, we denote this colour MLP as $\mathcal{C}_s(\mathbf{x}, \mathbf{v})$ in future sections. Using the scene SDF, Φ_s, and the colour values, $\mathcal{C}_s$, one can apply the volume rendering proposed in [33] to render the scene for each camera pose and each object separately, which is visualised in our supplementary video.

4.2. Interaction-aware Training

Given the per-object segmentation masks, we optimise the scene parameters defined above using the following loss formulation. Let $\mathbf{C}$ represent the segmented foreground ground-truth colour, while $\mathbf{C}_1 = \mathbf{C} \circ \mathcal{M}_1$ and $\mathbf{C}_2 = \mathbf{C} \circ \mathcal{M}_2$ be the segmented objects' ground-truth colours, where "$\circ$" indicates the Hadamard product. The rendering loss function $\mathcal{L}_{\text{color}}$ is then defined as:

$$\hat{\mathcal{L}}_{\text{color}} = \sum_p |\hat{\mathbf{C}}_1(\mathbf{p}) - \mathbf{C}_1(\mathbf{p})|_s + \sum_p |\hat{\mathbf{C}}_2(\mathbf{p}) - \mathbf{C}_2(\mathbf{p})|_s, \tag{4}$$

where $|\cdot|_s$ denotes the Smooth-L1 loss.

Training Stabilization: In practice, since the individual objects can be relatively smaller than the overall scene scale, using only these segmented colours for SDF supervision makes the training unstable, especially in the beginning. To stabilise the training, especially in the earlier stages, we use the entire scene colour for supervision as well, with predicted scene colour calculated using Eq. (1). The modified final loss now reads as:

$$\mathcal{L}_{\text{color}} = \hat{\mathcal{L}}_{\text{color}} + \sum_p |\hat{\mathbf{C}}_s(\mathbf{p}) - \mathbf{C}(\mathbf{p})|_s. \tag{5}$$

While the estimated scene colour $\mathcal{C}_s(\mathbf{x}, \mathbf{v})$ can be directly supervised with the RGB colour loss, it is insufficient to enforce that the learned object and the human SDFs are *separate*. Hence, the next question is how to ensure that the colour loss leads to separation between the two SDFs without arbitrarily entangling the two geometries. Towards this goal, we introduce an α-blending colour loss and a regularisation term that constrains the opacities of the two fields. Recall that we construct the scene SDF $\Phi_s(\mathbf{e})$ as a union of the individual object SDFs, $\Phi_1(\mathbf{e})$ and $\Phi_2(\mathbf{e})$, as in Eq. (3). We can, therefore, recover the opacity of the individual objects, α_1^i and α_2^i, at position i using the respective SDFs (as in Eq. (2)). Now, to recover the joint scene opacity α_s^i, we α-composite the opacity contributions from both α_1^i and α_2^i:

$$\alpha_s^i = \alpha_1^i + \alpha_2^i - \alpha_1^i \alpha_2^i. \tag{6}$$

After substituting α_s^i from Eq. (6) in the rendering equation (1), we obtain:

$$\begin{aligned}\hat{\mathbf{C}}_{\mathbf{s}}(\mathbf{p}) &= \sum_{i=1}^{N} T_s^i \left(\alpha_1^i + \alpha_2^i - \alpha_1^i \alpha_2^i\right) c_s^i \\ &= \sum_{i=1}^{N} T_s^i \alpha_1^i c_s^i + T_s^i \alpha_2^i c_s^i - T_s^i \alpha_1^i \alpha_2^i c_s^i. \end{aligned} \tag{7}$$

Here, the first two terms, $\hat{\mathbf{C}}_1(\mathbf{p}) = \sum_{i=1}^{N} T_s^i \alpha_1^i c_s^i$ and $\hat{\mathbf{C}}_2(\mathbf{p}) = \sum_{i=1}^{N} T_s^i \alpha_2^i c_s^i$, represent the visible part of the two objects. Note, however, that the transmittance T_s^i and colour c_s^i terms correspond to the entire scene, and T_s^i reaches close to 0, when obstructed by either of the objects, thus ensuring that the final colour output is occlusion-aware.

Alpha-Blending Regularisation. To achieve separable reconstruction, we assume that all the objects in a scene are opaque. Therefore, at any point, at least one of the two opacities, (α_1^i, α_2^i), should be 0, such that. $\alpha_1^i \alpha_2^i = 0$ in Eq. (7). This observation is key to ensuring clean separation boundaries between the different objects. However, we cannot *explicitly* enforce this constraint as there is no way to know which of the two opacities should be 0. Thus, we introduce the following α-regularisation which ensures that each position is opaque due to the influence of only one of the two SDFs, thereby preventing SDF penetration:

$$\mathcal{L}_{\text{alpha}} = \sum_p \left(\exp\left(\frac{\beta}{\lambda_t} \cdot \alpha_1(\mathbf{p}) \cdot \alpha_2(\mathbf{p}) \right) - 1 \right), \tag{8}$$

where β is the learnable parameter from Eq. (2), λ_t is a hyperparameter controlling the temperature of the exponential curve and $\alpha_1, \alpha_2 \geq 0$. Here, β increases as the training converges, thereby regularising more for overlapping opacities at the later stages of training. We empirically find that the above-proposed α-regularisation performs the best and show an ablation in Fig. 8. Finally, we employ the commonly used Eikonal regularisation term to obtain the correct SDFs:

$$\begin{aligned}\mathcal{L}_{\text{eik}} = \sum_{\mathbf{x}} &\|\nabla_{\mathbf{x}} \Phi_1(\mathbf{e}, \mathbf{g}) - 1\|^2 + \|\nabla_{\mathbf{x}} \Phi_2(\mathbf{e}, \mathbf{g}) - 1\|^2 + \\ &+ \|\nabla_{\mathbf{x}} \Phi_s(\mathbf{e}, \mathbf{g}) - 1\|^2 . \end{aligned} \tag{9}$$

The resulting total loss can now be written as:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{color}} + \lambda_\alpha \mathcal{L}_{\text{alpha}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}}. \tag{10}$$

We use $\lambda_\alpha = 0.1$, $\lambda_{\text{eik}} = 0.01$ and $\lambda_t = 100.0$ as hyperparameters in all our experiments.

Opacity *vs.* direct SDF regularisation. While we ensure separability by regularising the per-object opacities, another possible approach would be to regularise at the SDF level: Specifically, one could enforce both SDFs Φ_1 and Φ_2 to be not negative at the same point. We empirically observe that the proposed α-regularisation performs best and show the corresponding ablation in Fig. 8.

5. Experiments

Metrics: We report novel-view synthesis evaluations on commonly used metrics such as peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and learned perceptual image patch similarity (LPIPS). We use bi-directional Chamfer distance to evaluate the 3D reconstruction quality.

As our method is agnostic to the object types, we demonstrate its effectiveness on four kinds of interactions: Human-object, hand-object, and object-object and human-human interaction. For the hand-object and object-object scenarios, we use scenes from the AffordPose [12] and WildRGB-D [38] datasets, respectively. AffordPose is a synthetic dataset with ground-truth geometry. We use these ground-truth meshes to generate a synthetic multi-view dataset of 60 views for six objects. Human-object interaction is especially challenging due to the large differences between the scales of the two entities. To evaluate our method comprehensively, we capture and record a new dataset with various human-object interactions with

Scene	Seq ID	Segmented NeuS2		ObjectSDF++		Ours	
		Overall scene	Object	Overall scene	Object	Overall scene	Object
Box	1	**4.86**	26.06	11.08	10.14	5.65	**7.9**
	2	**11.18**	42.87	18.96	**12.62**	11.41	38.46
Book	1	6.78	-	13.31	9.13	**5.35**	**7.33**
	2	7.30	-	11.96	9.35	**5.95**	**8.28**
Birdhouse	1	4.70	-	9.27	13.2	**4.50**	**10.99**
	2	4.96	-	8.74	14.33	**4.38**	**11.13**
Spray Bottle	1	**3.21**	-	9.23	10.89	4.40	**7.72**
	2	**3.04**	-	9.60	9.67	4.19	**7.48**
Hanoi Tower	1	**4.13**	-	10.06	10.86	4.30	**8.59**
	2	**3.66**	-	9.47	10.53	4.30	**8.64**
Cupid	1	8.11	119.20	36.16	16.76	**5.79**	**11.71**
	2	6.96	51.84	18.81	63.34	**5.92**	**8.66**
Mean		5.74	59.99	13.89	15.90	**5.51**	**11.40**

Table 1. 3D reconstruction accuracy of the *overall scene* and individual *object*, on human-object evaluation dataset, using Chamfer Distance (lower the better).

objects of different scales and complexity (further details in supplementary **??**) and also evaluate our method on a few scenes of human-human interaction from ReMoCap[7] dataset. The evaluation datasets also differ in the relative scales of the objects in the scene. For example, the human-object evaluation dataset consists of small scale objects in a large capture dome, as opposed to the Wild-RGBD dataset.

Comparisons: We compare our method against ObjectSDF++ [37] and "Segmented Neus2". As NeuS2 [34] is not designed to be instance-specific, we train it separately for each object in the scene by providing the corresponding segmentation masks (henceforth referred to as "Segmented NeuS2"). ObjectSDF++, on the other hand, is a state-of-the-art method that reconstructs objects in a scene separately.

For human-object, human-human and object-object datasets, we evaluate 3D reconstruction quality for the overall scene. As we do not have a ground truth here, we consider a NeuS2 [34] model trained on the entire scene (agnostic of the objects) as the pseudo ground truth. This allows us to compare the compositionally reconstructed scene with the non-compositional reconstruction of the scene, which can be treated as an upper bound on the reconstruction quality. In the case of human-object interaction, we also evaluate the object reconstruction separately with respect to 3D scanned templates–we first align the pre-scanned 3D object shape with the reconstructed mesh (extracted using marching cubes) using the rigid ICP [1] algorithm and then compute the Chamfer distance. We use the ground-truth meshes provided in AffordPose to compute the Chamfer distance metric for both the hand and the object separately.

5.1. Geometric Evaluation

Human-Object Interaction: Table 1 tabulates the quantitative comparisons of our method with ObjectSDF++ and Segmented Neus2 for the whole scenes and also individual objects separately. While we outperform ObjectSDF++ on most scenes, an interesting pattern emerges: The *Overall scene* in Tab. 1 shows Segmented NeuS2 achieves consistently lower Chamfer distance for the scenes involving

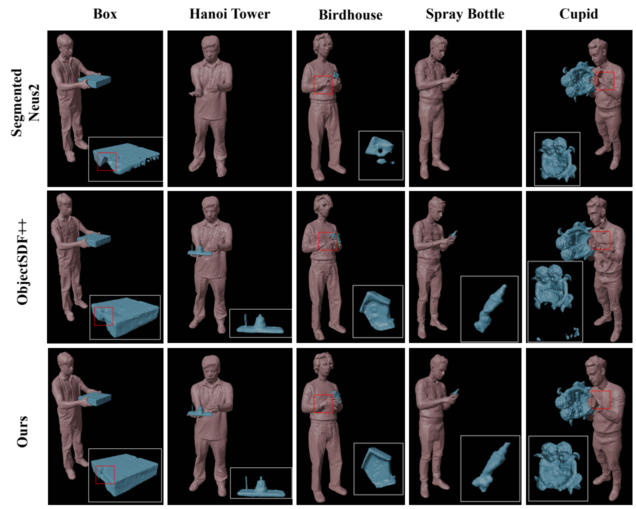

Figure 3. Qualitative comparison of the reconstructed geometry. In most scenes, we obtain better geometry, with fewer deformations near the contact regions. Best viewed when zoomed.

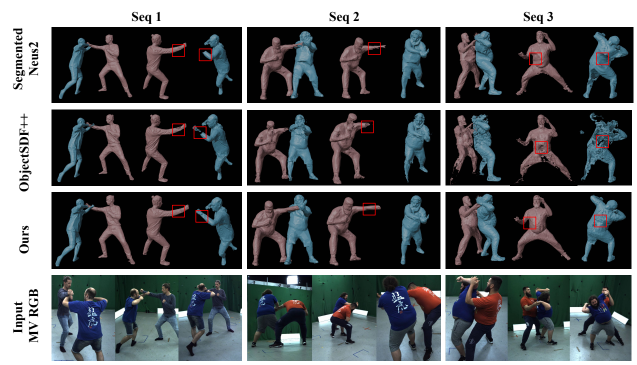

Figure 4. Qualitative comparison of 3D scene reconstructions with human-human interaction along with selected multi-view (MV) input images. Digital zoom recommended.

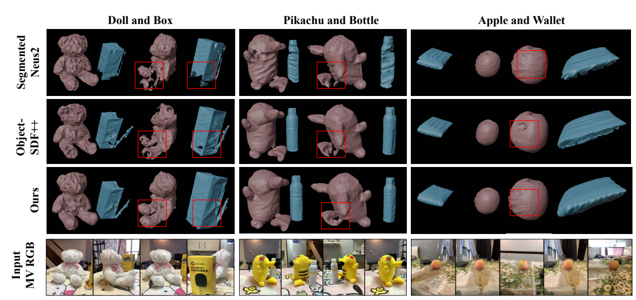

Figure 5. Qualitative comparison of reconstruction of scenes involving two objects in proximity, along with samples from the multi-view (MV) input images. Digital zoom recommended.

small objects like *Spray Bottle* and *Hanoi Tower*. The full-scene results are dominated by the reconstruction of the human. However, these results deteriorate with a relatively larger object like *Cupid*. As the occlusions on the human body grow (due to the larger object size), the Segmented

Scene	Segmented Neus2		ObjectSDF++		Ours	
	Hand	Object	Hand	Object	Hand	Object
Bag	7.51	**4.16**	5.86	11.30	**5.83**	4.62
Bottle	8.28	8.28	6.28	2.76	**5.40**	**1.77**
Earphone	7.18	7.15	5.85	6.18	**5.44**	**6.17**
Knife	6.06	4.10	**5.64**	4.32	5.68	**3.24**
Pot	-	3.13	**6.94**	11.97	7.49	**2.79**
Scissors	6.00	28.75	6.52	4.24	**5.42**	**4.20**
Mean	7.00	9.26	6.18	6.79	**5.87**	**3.80**

Table 2. 3D reconstruction accuracy for different object sequences on the AffordPose dataset. Chamfer distance (lower the better) is calculated against the ground-truth meshes provided. Segmented Neus2 reconstruction fails for the hand in the *Pot* scene, which is represented with "-".

Seq ID	Segmented Neus2	ObjectSDF++	Ours
1	**4.36**	25.05	4.73
2	9.8	21.14	**6.19**
3	13.71	43.26	**6.67**

Table 3. Comparison of 3D reconstruction quality for human-human interaction. Chamfer distance metric (lower the better) is calculated against the overall scene reconstructed using Neus2.

NeuS2 struggles to maintain artefact-free reconstruction. Further results in the *Object* in Tab. 1 indicate that indeed, the Segmented NeuS2 does not recover the object geometries in most cases from the supervision using segmented objects alone. We believe that this is because Neus2 tries to *carve away* regions that are segmented out because of occlusion. This causes conflicting optimisation goals for different camera views depending on whether the object is visible. Since this can be a significant volume relative to the total volume for smaller objects, Neus2 fails to converge. We show a qualitative comparison of our 3D reconstructions in Fig. 3.

Hand-Object Interaction: We show quantitative and qualitative comparisons in Tab. 2 and **??** in the supplement, respectively. ObjectSDF++ suffers from denting artefacts in the occluded areas, whereas our method generates a smoother surface. Both methods are at par when the object is only mildly in contact (as in the case of comparably thin scissors).

Human-Human Interaction: We show qualitative and quantitative comparisons in Fig. 4 and Tab. 3, respectively. Similarly to the previous section, the Segmented Neus2 reconstructs humans with missing sections, while reconstruction near contact areas in ObjectSDF++ shows severe artefacts. Note that for the case of Seq 1—even though numerically Segmented Neus2 appears to be slightly better—we can see in the qualitative results that our method reconstructs the hand near occlusions significantly better.

Object-Object Reconstruction: The qualitative and quantitative results for two interesting objects are shown in Fig. 5 and Tab. 4, respectively. Our method excels in two cases out of three.

	Segmented Neus2	ObjectSDF++	Ours
Doll and Box	7.44	11.88	**4.44**
Pikachu and Bottle	4.11	3.96	**2.99**
Apple and Wallet	**2.72**	6.28	3.83

Table 4. Comparison of the 3D reconstruction quality for overall scene, on the WilDRGBD dataset using Chamfer Distance metric (lower the better).

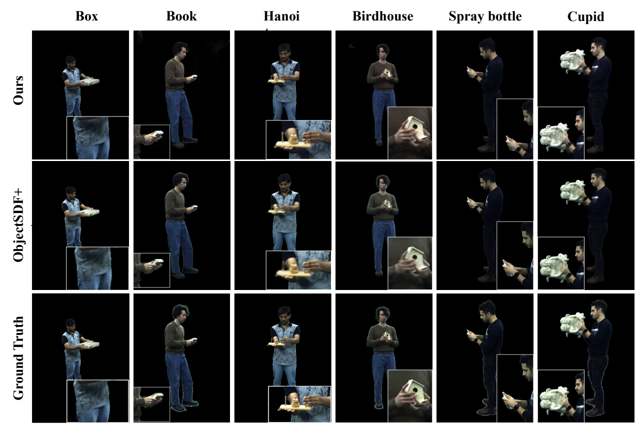

Figure 6. Qualitative comparison of novel view synthesis on human-object scenes. The results of ObjectSDF++ are blurrier than our rendered views, especially around the object. Digital zoom recommended.

Scene	Seq ID	PSNR↑		SSIM ↑		LPIPS ↓	
		Object-SDF++	Ours	Object-SDF++	Ours	Object-SDF++	Ours
Box	1	27.28	**30.39**	0.95	**0.96**	0.08	**0.07**
	2	28.26	**29.64**	0.95	**0.96**	0.08	**0.07**
Book	1	25.61	**33.06**	0.95	**0.96**	0.08	**0.06**
	2	28.70	**32.12**	0.95	**0.96**	0.08	**0.07**
Birdhouse	1	29.82	**33.71**	0.96	**0.97**	0.08	**0.05**
	2	26.37	**32.69**	0.93	**0.96**	0.10	**0.06**
Spray Bottle	1	29.90	**32.73**	0.97	**0.98**	0.07	**0.04**
	2	26.32	**36.38**	0.94	**0.98**	0.10	**0.04**
Hanoi Tower	1	27.25	**30.36**	0.95	**0.96**	0.09	**0.07**
	2	26.30	**29.63**	**0.96**	**0.96**	**0.07**	**0.07**
Cupid	1	29.92	**34.10**	0.96	**0.98**	0.10	**0.05**
	2	30.50	**34.20**	**0.97**	**0.97**	**0.08**	**0.08**
Mean		28.02	**32.42**	0.95	**0.97**	0.08	**0.06**

Table 5. Quantitative comparison of view synthesis on held-out views for the human-object dataset. We consistently outperform ObjectSDF++.

5.2. Appearance Evaluation

To evaluate the quality of novel view synthesis, we render the entire scenes into a held-out set of views. We report the human-object novel-view results in Tab. 5. Again, we achieve consistently better performance than ObjectSDF++; see Fig. 6 for the visualisations. One can observe blurring artefacts in ObjectSDF++ renderings, which are especially pronounced around the object. As it supervises only the individual opacities, we hypothesize that the colour network of Wu *et al.* [37] assigns colours to any residual opacity, which is more likely to exist at the transition boundaries. We also provide appearance evaluation for human-human

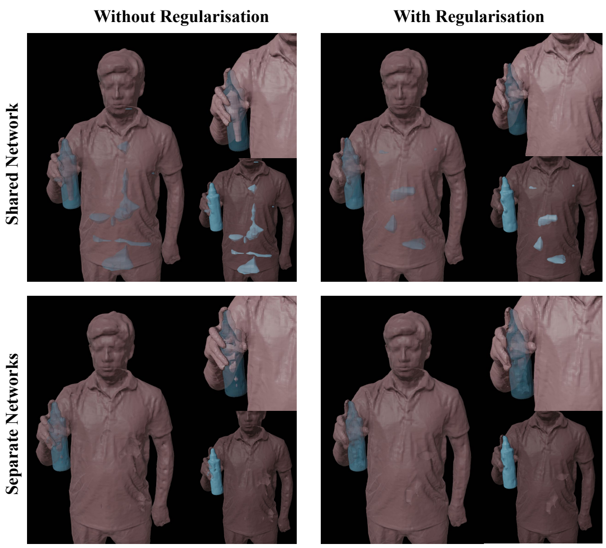

Figure 7. Qualitative comparison (ablations). We observe that the overall scene reconstruction largely remains the same, though the individual object and human reconstruction quality deteriorates because of phantom blobs formed underneath the surface (as highlighted inside the transparent surface) when we have a shared MLP or we do not use the alpha-regularisation.

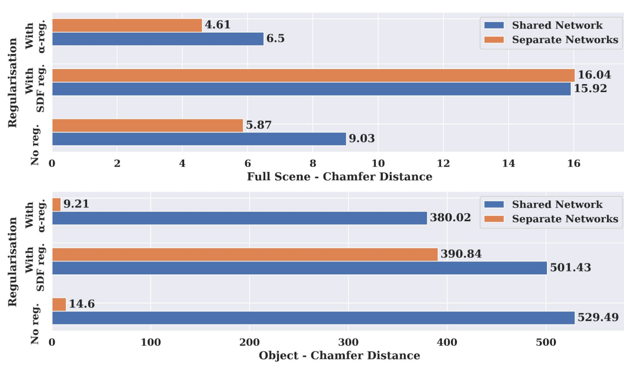

Figure 8. Quantitative evaluation with ablated components shows that having separate MLPs for human and object, and the proposed alpha-regularisation are important for high-quality reconstruction.

interaction scenes in the supplementary **??**.

5.3. Ablations

We also perform an ablation study to evaluate the design choices. In particular, we evaluate the importance of having separate MLPs for the human and the object, instead of a single, shared MLP predicting both the SDFs as done in ObjectSDF++. We also compare the proposed alpha-regularisation against SDF level regularisation $\sum_p \left(\exp(\frac{\beta}{\lambda_t} \cdot \max(-\Phi_1, 0) \cdot \max(-\Phi_2, 0) \right)$, such that both SDFs are not negative at the same position as mentioned in Sec. 4.2. The differences in the results are shown

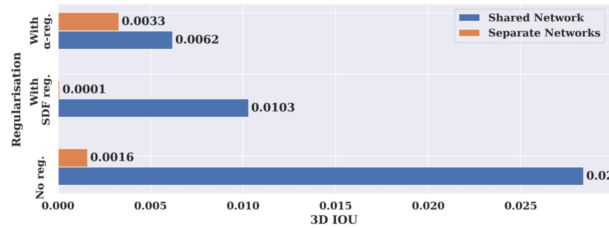

Figure 9. 3D IOU to assess the intersection of segmented objects for the ablated components.

in Fig. 7, and the quantitative results are shown in Fig. 8. The overall scene reconstruction largely remains the same, but the individual object and human reconstruction qualities deteriorate because of phantom blobs that get formed underneath the surface of the complimentary SDF when we have shared MLP or we do not use the alpha-regularisation. Since these blobs are underneath the surface, they are invisible in renderings, thereby satisfying rendering losses. To, demonstrate that the high Chamfer distance is due to the intersecting phantom blobs, we also show results by calculating the 3D IOU between the human and the object. Ideally, we do not want any penetrations, hence the IOU should be as low as possible. Thus, we can see that having separate MLPs for humans and objects and the proposed alpha-regularisation are important for high-quality reconstruction.

6. Conclusion

We introduced a novel method for separable 3D reconstruction of two-object interaction in a multi-view setting considering the challenges of occlusion and difference in object scales. Following the insight that the opaque objects in the scene must have non-overlapping opacities in the implicit network, we showed that the proposed α-blending regulariser can indeed incentivise the network to learn disjoint opacities ensuring that the object boundaries remain separate. Through comprehensive experiments, our approach demonstrated the suitability and high accuracy, both on 3D and novel view synthesis metrics and across several datasets. Our simple yet effective regularization strategy demonstrated in a *generalized* setting, can potentially be applied to specific use cases such as template-based human performance capture, or compositional scene generation. We hope that the newly recorded datasets will allow researchers to make further progress in studying the challenging problem of multi-view compositional 3D scene reconstruction. In the future, we intend to refine our method for larger-scale and multi-object scenes and use it for markerless dataset collection.

References

[1] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, 1992. 6

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 2, 3

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[5] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[6] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[7] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 6

[8] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[9] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2

[10] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299. Springer, 2022. 3

[11] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *ACM SIGGRAPH*, 2021. 1, 2

[12] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *International Conference on Computer Vision (ICCV)*, pages 14713–14724, 2023. 2, 5

[13] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction, 2022. 3

[14] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. Neuralhofusion: Neural volumetric rendering under human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2022. 3

[15] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, pages 10138–10148, 2021. 3

[16] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 3

[17] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3

[19] Willi Menapace, Aliaksandr Siarohin, Stéphane Lathuilière, Panos Achlioptas, Vladislav Golyanik, Sergey Tulyakov, and Elisa Ricci. Promptable game models: Text-guided game simulation via masked diffusion models. *ACM Trans. Graph. (ToG)*, 43(2), 2024. 2

[20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2, 3

[22] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[24] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[25] Radu Alexandru Rosu and Sven Behnke. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3

[26] Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. 3

[27] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH Conference Proceedings*, 2022. 2

[28] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2

[29] Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. Metacap: Meta-learning priors from multi-view imagery for sparse-view human performance capture and rendering. In *ECCV*, 2024. 2

[30] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[31] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. Advances in Neural Rendering. *Computer Graphics Forum (EG STAR 2022)*, 2022. 2

[32] Edith Tretschk, Vladislav Golyanik, Michael Zollhöfer, Aljaz Bozic, Christoph Lassner, and Christian Theobalt. Scenerflow: Time-consistent reconstruction of general dynamic scenes. In *International Conference on 3D Vision (3DV)*, 2024. 3

[33] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 1, 2, 3, 4

[34] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 6

[35] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 1, 2

[36] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, 2022. 2

[37] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2, 3, 6, 7

[38] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024. 2, 5

[39] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[40] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 4

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

[42] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[43] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[44] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[45] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph.*, 41(6), 2022. 2, 3

[46] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[47] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *International Conference on Computer Vision (ICCV)*, 2019. 3