Inter3D: A Benchmark and Strong Baseline for Human-Interactive 3D Object Reconstruction

Gan Chen[1] , **Ying He**[1,2] , **Mulin Yu**[3] , **F. Richard Yu**[1,2] , **Gang Xu**[2] , **Fei Ma**[2] , **Ming Li**[2] and **Guang Zhou**[2]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3]Shanghai Artificial Intelligence Laboratory
chengan2022@email.szu.edu.cn, heying@szu.edu.cn, yumulin@pjlab.org.cn, richard.yu@carleton.ca, {xugang, mafei, liming}@gml.ac.cn, maxwell@deeproute.ai

Abstract

Recent advancements in implicit 3D reconstruction methods, *e.g.,* neural rendering fields and Gaussian splatting, have primarily focused on novel view synthesis of static or dynamic objects with *continuous* motion states. However, these approaches struggle to efficiently model a human-interactive object with n movable parts, requiring 2^n separate models to represent all *discrete* states. To overcome this limitation, we propose *Inter3D*, a new benchmark and approach for *novel state synthesis* of human-interactive objects. We introduce a self-collected dataset featuring commonly encountered interactive objects and a new evaluation pipeline, where only individual part states are observed during training, while part combination states remain unseen. We also propose a strong baseline approach that leverages Space Discrepancy Tensors to efficiently modelling all states of an object. To alleviate the impractical constraints on camera trajectories across training states, we propose a Mutual State Regularization mechanism to enhance the spatial density consistency of movable parts. In addition, we explore two occupancy grid sampling strategies to facilitate training efficiency. We conduct extensive experiments on the proposed benchmark, showcasing the challenges of the task and the superiority of our approach. The code and data are publicly available at https://github.com/Inter3D-ui/Inter3D.

1 Introduction

3D object reconstruction has been a long-standing focus of research in computer vision and graphics [Berger *et al.*, 2017]. Recently, neural radiance fields (NeRFs) [Mildenhall *et al.*, 2021] pioneer volumetric rendering to enable high-fidelity modelling of 3D objects with fine-grained geometric structures, facilitating its widespread applications across various domains [Barron *et al.*, 2022; Fridovich-Keil *et al.*, 2023a; Pumarola *et al.*, 2021; Poole *et al.*, 2022; Liu *et al.*, 2024].
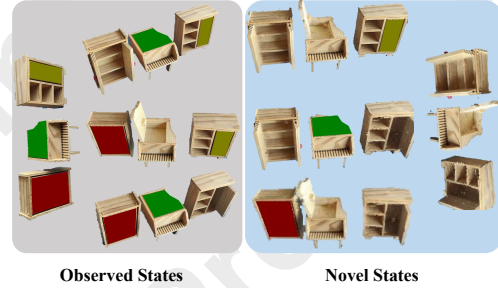

Figure 1: Illustration of reconstructing a human-interactive object, Furniture, in our *Inter3D*. The object consists of $n = 3$ movable parts, highlighted in red, green, and olive-green. With $2^n = 8$ *discrete* states, the task requires significant computational and memory resources, making it impractical for existing methods to train separate models for each state. Furthermore, ensuring consistency in external appearances and internal structures across states poses a significant challenge when states are trained independently. In contrast, our approach efficiently synthesizes *novel combination states* by observing only the canonical and individual part states.

3D Gaussian splatting [Kerbl *et al.*, 2023] extends the principles of NeRFs by representing an object with a set of 3D Gaussian primitives and synthesizes novel views via rasterization, offering a more expressive and rendering-efficient 3D representations [Luiten *et al.*, 2024].

Despite their advancements, existing methods predominantly focus on modeling static objects or dynamic ones with simple *continuous* motions, where the states of objects exhibit no abrupt changes. Also, these methods primarily focus on synthesizing novel views of the observed states during training. However, many objects commonly encountered in daily life are interactive and have numerous *discrete* states. A human-interactive object with n movable parts results in a total of 2^n states. Efficiently modeling such objects holds significant potential for applications in virtual reality [Wohlgenannt *et al.*, 2020], gaming [Cameirão *et al.*, 2009], and embedded AI [Machinery, 1950].

As illustrated in Figure 1, the Furniture object consists of three movable parts, highlighted in red, green, and olive-green. It features one canonical state and three individual states where only one part is open, along with four combi-

nation states where two or more parts are open. Existing methods fail to efficiently model all possible states of the Furniture object. A straightforward solution would be to train a separate 3D model for each object state. However, this solution incurs prohibitive computational and memory overhead. Moreover, different states of the same human-interactive object inherently share consistent external appearances and internal structures, which this solution fails to capture.

To address these limitations, we propose *Inter3D*, a novel benchmark and method for human-interactive 3D object reconstruction. *Inter3D* introduces a self-collected dataset comprising commonly encountered yet challenging interactive objects in daily life, such as Car, Display Case, Drawer, and Furniture. It also introduces a new evaluation scenario, *i.e., novel state synthesis*, in which only the canonical and individual part states of an object are observed during training, while all combination states remain unseen and are reserved for evaluation, as illustrated in Figure 1. This setup is particularly challenging as no views of combination states are used during training, yet the synthesized novel states must remain consistent with the observed states.

Our method effectively tackles this challenging scenario by leveraging Space Discrepancy Tensors in combination with multi-resolution hash encoding (MHE) [Müller *et al.*, 2022] to efficiently model all states of an interactive object. Specifically, MHE is initially employed to represent the canonical state of the object, where all movable parts remain untouched. The Space Discrepancy Tensors capture the inherent variations between each manipulated part state and the canonical state, providing a compact and expressive representation for the decomposition of individual movable parts. To mitigate rendering quality degradation caused by trajectory perturbations across training states, we introduce a Mutual State Regularization mechanism to ensure consistency of movable parts across states. Additionally, we explore two new occupancy grid sampling schemes to optimize training efficiency. After training on the canonical state and individual part states, our method can synthesize arbitrary combination states by selecting the maximum density difference.

Our key contributions in this work are summarized as follows:

- We introduce a novel benchmark for human-interactive 3D object reconstruction, supported by a self-collected dataset of commonly encountered interactive objects and a novel state synthesis testing scenario. This benchmark provides a foundation for advancing research in this field.

- We propose a novel approach that integrates Space Discrepancy Tensors with multi-resolution hash encoding, enabling efficient modeling of all possible states of interactive objects. This approach provides a strong baseline for future studies.

- We develop a Mutual State Regularization mechanism to mitigate rendering artifacts caused by trajectory inconsistencies and design two occupancy grid sampling strategies to balance memory usage and training speed.

2 Related Works

Static 3D Reconstruction. Recent advances in implicit 3D reconstruction, particularly with neural radiance fields (NeRFs) [Mildenhall *et al.*, 2021], have enabled high-quality novel view generation by predicting volume density and radiance from 3D coordinates and viewpoint directions. Variants like mip-NeRF [Barron *et al.*, 2022] optimize performance in more complex scenarios. 3D Gaussian splatting [Kerbl *et al.*, 2023] uses a set of Gaussian primitives with learnable parameters, *i.e.,* positions, covariance matrices, opacity, and spherical harmonics coefficients, and differentiable tile rasterizer to improve synthesis quality and training speed. However, these methods primarily focus on modeling static objects and can only be applied to a single state of a human-interactive object. To model all possible states, a series of independent 3D models would be required, significantly increasing computational costs and failing to ensure consistency across states.

Dynamic 4D Reconstruction. Dynamic 4D reconstruction methods, such as K-Plane [Fridovich-Keil *et al.*, 2023a], D-NeRF [Pumarola *et al.*, 2021], and 4D Gaussians [Wu *et al.*, 2024], extend vanilla NeRFs or Gaussian splatting by incorporating motion factors to model geometric and texture changes of time-varying objects. These methods excel at modeling dynamic objects with continuous states and synthesizing novel views of the training states. However, our approach focuses on modeling human-interactive objects with numerous discrete states and emphasizes novel state synthesis, a task that is infeasible for these methods.

3 Human-Interactive Benchmark

3.1 Data Collection

To build the benchmark, we collect a diverse dataset comprising four objects with multiple movable parts commonly encountered in daily life: Car, Display Case, Drawer, and Furniture. These objects are arranged in increasing order of complexity. The **Car** represents smooth-surfaced objects with two movable parts, *i.e.,* front and rear doors, providing a straightforward yet realistic interaction scenario. The **Display Case**, with its two movable *glass* doors and intricate internal contents, introduces additional complexity, especially for assessing the effectiveness of our approach under low-density conditions. The **Drawer**, featuring a typical three-drawer design commonly seen in daily environments, serves as a representative example of structured objects with discrete states. Finally, the **Furniture**, consisting of three types of cabinets, each with one movable part, features a challenging sandy background that introduces additional noise and interference, making it the most complex object in our dataset.

We capture object videos using handheld, forward-facing mobile shots and utilize SAM2 [Ravi *et al.*, 2024b] to segment movable parts that remain in place. For each object, 60 images are extracted through uniform sampling at regular intervals to represent the canonical state and each individual state. As illustrated in Figure 3, we collect forward-facing data for a toy car with two movable parts, *i.e.,* its front and rear doors. Images are captured for the canonical state S_0 and for individual part states S_1 and S_2, where only one door is opened at a time. Camera poses are extracted using
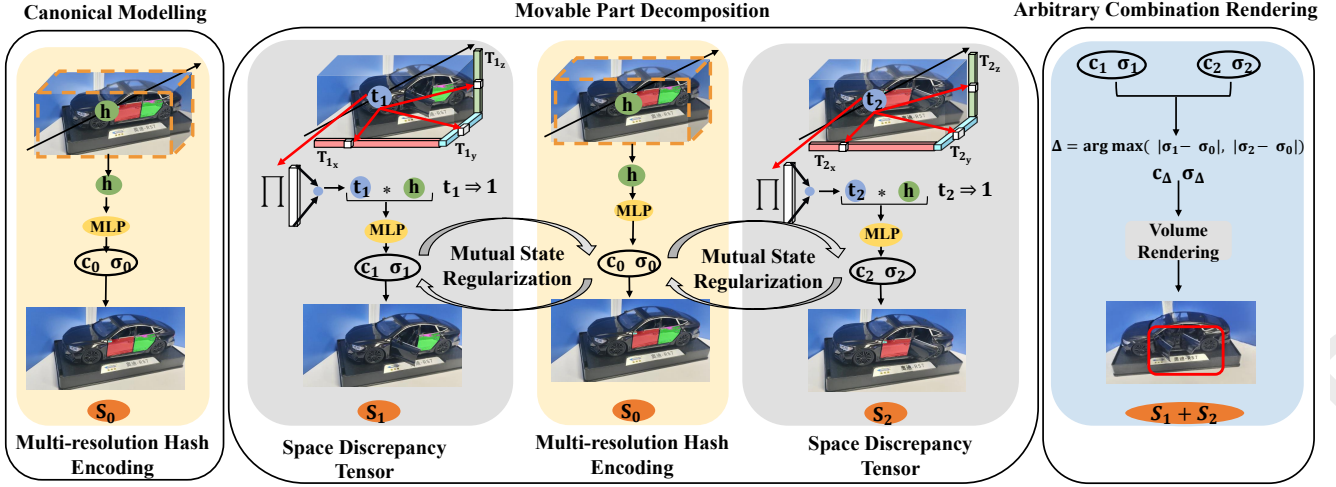
Figure 2: Overview of our approach, which comprises three key stages, *i.e.,* Canonical Modelling, Movable Part Decomposition, and Arbitrary Combination Synthesis. The inactive movable parts in each state are covered by colored masks. In **Canonical Modelling**, the canonical state S_0 of the interactive object, with all n movable parts closed, is reconstructed using InstantNGP. The spatial sample point features are retrieved via multi-resolution hash encoding, represented as $\mathbf{h}$, with attributes color c_0 and density σ_0 projected through a multi-layer perceptron (MLP). In **Movable Part Decomposition**, each movable part is manipulated sequentially, resulting in states where only one part is open, denoted as $S_i, i \in \{1, ..., n\}$. For each S_i, the sample point features are represented as $\mathbf{h} * \mathbf{t}_i$, where $\mathbf{t}_i$ is derived from the proposed Space Discrepancy Tensors, encoding the differences between S_i and S_0. To address camera trajectory perturbations across states, we introduce a Mutual State Regularization mechanism, which mitigates rendering artifacts caused by movable part misalignment. In **Arbitrary Combination Synthesis**, to render the combined novel state $S_i + S_j$, the sample point features with the maximum density difference, selected from $\{|\sigma_i - \sigma_0|, |\sigma_j - \sigma_0|\}$, are used for volumetric rendering, which enables the efficient modelling of arbitrary combinations of movable parts.

Figure 3: Data collection example of the object **Car** in our human-interactive benchmark *Inter3D*. A sequence of forward-facing images are captured for the canonical state S_0 with doors closed, the state S_1 with the front door open, and the state S_2 with the rear door open.

COLMAP [Fisher *et al.*, 2021], ensuring precise alignment for subsequent learning.

3.2 Novel State Evaluation

In traditional 3D or 4D reconstruction scenarios, objects are either static or exhibit *continuous* states. Furthermore, models typically observe all possible states during training, with only novel views used for evaluation. In contrast, our benchmark introduces interactive objects with numerous *discrete* states, where the evaluating states are novel and remain en-

tirely unseen during training.

For an object with n movable parts, there are 2^n possible states. Each part is manipulated individually, resulting in an individual state $S_i, i \in \{1, \ldots, n\}$, which is observed during training. When multiple parts are manipulated simultaneously, the resulting state becomes a *spatial* combination of individual states. For instance, $S_i + S_j, i, j \in \{1, \ldots, n\}$, represents the simultaneous manipulation of the ith and jth parts. These combinatorial states are entirely unseen during training and reserved for evaluation rendering, significantly increasing the complexity of the task. This challenging setup requires models to achieve a higher level of generalization and adaptability, thereby pushing the boundaries of existing 3D reconstruction techniques.

4 Human-Interactive 3D Reconstruction

Inter3D establishes a new benchmark for synthesizing diverse states of human-interactive objects, offering a powerful baseline method. Unlike classical approaches, which require 2^n separate models to represent an object with n movable parts, our method efficiently represents the object using a unified model. Figure 2 illustrates the framework of our approach. In this section, we first describe the details of Instant-NGP [Müller *et al.*, 2022] in Sec. 4.1. Then, we detail the initialization Canonical Modelling in Sec. 4.2, which defines the canonical state of a human-interactive object and provides a solid foundation for the subsequent steps. Sec. 4.3 introduces the Space Discrepancy Tensors, which efficiently models discrepancies between individual part states and canonical state.

Sec. 4.4 presents the pipeline for representing novel combination states, and elaborates how the Mutual State Regularization alleviates the impractical constraints on camera trajectories across training states. Finally, Sec. 4.5 discusses two occupancy grid sampling strategies designed to enhance training efficiency.

4.1 Preliminaries

InstantNGP [Müller *et al.*, 2022] boosts up the rendering and training speed of NeRFs. It proposed a multi-resolution voxel grid, implicitly encoded using hash functions, enabling efficient optimization and rendering with compact models. A ray passing through three-dimensional space is defined as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ is the ray's origin, $\mathbf{d}$ is its direction, and t represents the distance along the ray. InstantNGP employs a density occupancy grid with step-based indexing to acquire sampling points. Then, based on the 3D coordinates of a sampling point, feature vectors of multiple surrounding voxels are retrieved from the hash table. These voxel feature vectors are then linearly interpolated to obtain the feature vector $\mathbf{h}$ for the sampling point. An MLP is used to compute the color c and density σ of the sampling points. The pixel color C is determined by the following volume rendering equation applied to the N sampling points along the ray:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d}) \, dt \text{ with,}$$
$$T(t) = exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(t')) \, dt'). \tag{1}$$

$\mathcal{L}_{MSE}$ is employed to minimize the discrepancy between the predicted pixel color and the ground truth C^{gt}:

$$\mathcal{L}_{MSE} = \mathbb{E}\left[\, ||C - C^{gt}||_2^2 \,\right]. \tag{2}$$

4.2 Canonical Modelling

Canonical Modelling reconstructs the canonical state of the object, providing a solid foundation for the subsequent movable part decomposition, where a part of the object is manipulated into an individual state. As shown in Figure 2, we leverage InstantNGP to model the canonical state S_0.

To enhance the compactness of spatial points and reduce the presence of fuzzy clouds in space, we employ the distortion loss function from MipNeRF360 [Barron *et al.*, 2022]:

$$\mathcal{L}_{\text{dist}}(s, w) = \sum_{i,j} w_i w_j \left| \frac{s_i + s_{i+1}}{2} - \frac{s_j + s_{j+1}}{2} \right| + \frac{1}{3}\sum_i w_i^2 \delta_i, \tag{3}$$

where $w_i = (1 - e^{-\sigma_i \delta_i})T_i$, and s represents the distance from the sample point to the ray origin. Additionally, we incorporate a loss term aimed at minimizing the entropy of the cumulative weights, which helps to concentrate the distribution of spatial sample points:

$$\mathcal{L}_{\text{opacity}} = \mathbb{E}\left[-O \log O\right], \tag{4}$$

where O is calculated as the sum of the weights w_i of N sample points along the ray, given by: $O = \sum_i^N w_i$.

4.3 Movable Part Decomposition

In this section, we aim to identify the movable components from the interactive object, and learn their individual states. As illustrated in Figure 2, we train the model across all individual states, starting with the initialized canonical state S_0. Instead of explicitly identifying the moving parts, we propose Space Discrepancy Tensors to implicitly encode the individual states, enhancing the scalability and flexibility of our method.

Specifically, the Space Discrepancy Tensors are three spatially correlated feature tensors in the Cartesian coordinate system, denoted as $\mathbf{T}_{i_\mathbf{x}}$, $\mathbf{T}_{i_\mathbf{y}}$, and $\mathbf{T}_{i_\mathbf{z}}$, each with a shape of 512×32, where 512 represents the spatial resolution and 32 denotes the feature dimension. These tensors are leveraged to capture the discrepancies introduced by the individual state S_i. We normalize the 3D coordinates of a random sampling point p and project them onto these three feature tensors. Using nearest-neighbor sampling, we extract the spatially relevant features along each axis, and then compute the final feature vector $\mathbf{t}_i$ of p by element-wise multiplication:

$$\mathbf{t}_i = \prod_{\mathbf{a} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}} \psi(\mathbf{T}_{i_\mathbf{a}}, \pi_\mathbf{a}(p)), \tag{5}$$

where $\pi_\mathbf{a}(p)$ represents projecting p onto the axis $\mathbf{a}$, and $\psi(\mathbf{T}_{i_\mathbf{a}}, \cdot)$ denotes the nearest-neighbor sampling of a point feature from the tensor $\mathbf{T}_{i_\mathbf{a}}$. To ensure the non-interactive part consistency between the canonical state and individual states, similar to K-Planes [Fridovich-Keil *et al.*, 2023b], we employ $\mathcal{L}_{consis}$ to constrain $\mathbf{T}_i$ towards 1:

$$\mathcal{L}_{consis}(\mathbf{T}_i) = \sum_{\mathbf{a} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}} ||1 - \mathbf{T}_{i_\mathbf{a}}||_1. \tag{6}$$

We also employ the Laplacian (second-order derivative) filter $\mathcal{L}_{smooth}$ in K-Planes [Fridovich-Keil *et al.*, 2023b] to encourage the smooth variations in the state S_i.

4.4 Arbitrary Combination Synthesis

We propose a simple yet effective strategy of synthesizing a novel combination state on top of the training states. For example, the state $S_1 + S_2$ in Figure 4 can be generated by spatially merging S_1 and S_2. Specifically, our goal is to combine the varying components of S_1 and S_2 relative to S_0, along with the non-interactive parts of S_0, to form a new state $S_1 + S_2$. The sample point density σ_i associated with the individual state S_i are employed to capture the differences *w.r.t.* the canonical state S_0 and the state with the maximum differences dominates the volume rendering:

$$\Delta = \arg \max_i |\sigma_i - \sigma_0| \quad i \in \{1, \ldots, n\}. \tag{7}$$

The corresponding density and color c_Δ and σ_Δ are then queried to render the new state $S_1 + S_2$ via Equation 1. It is straightforward to adapt this synthesis strategy to an interactive object with more movable parts.

Mutual State Regularization. In practice, it is infeasible to maintain absolutely identical camera trajectories across different training states, which degrades the synthesis quality of
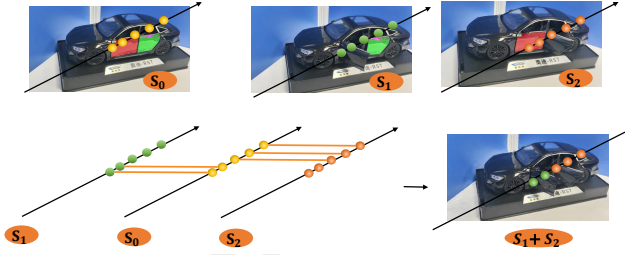
Figure 4: Illustration of synthesizing the novel combination state on the **Car** object. Given the features of the same sample point across the training states, its optimal feature on the novel state is determined by comparing the individual state S_i with the canonical state S_0. The inactive movable parts in each state are covered by colored masks.

movable parts. To address this issue, we propose the Mutual State Regularization (MSR) to enhance the spatial consistency of movable parts between training states.

As shown in Figure 5.a, we match the inactive movable part, *i.e.,* the back door, of S_1 to that of S_0. Specifically, we randomly sample a set of points along a ray passing through the back door of S_0 (segmented by SAM2 [Ravi *et al.*, 2024a]) and obtain the pixel value $C_0[\text{Mask}]$ via integration. We then project the points onto the S_1 space and query the pixel value $C_1[\text{Mask}]$. The $\mathcal{L}_{MSR}$ loss is employed to align them, ensuring the back door consistency:

$$\mathcal{L}_{MSR} = ||C_1[Mask] - C_0[(Mask)]||_1. \quad (8)$$

Reversely, in Figure 5.b, we align the pixel values of the back door of S_0 with those of S_1. The alignment process can be easily generalized to other individual state S_i.
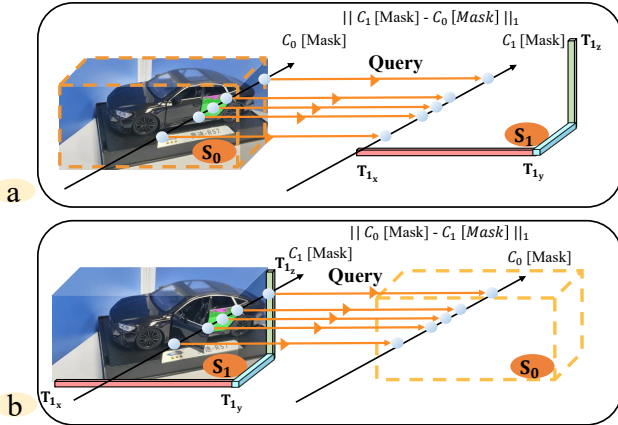

Figure 5: Illustration of our Mutual State Regularization mechanism on the **Car** object. MSR regularizes the sampling point features of the inactive movable part in the individual state S_i with the canonical state S_0 through pixel alignment, and vice versa. $\mathcal{L}_1$ loss is applied to ensure consistency.

4.5 Training Efficiency

As aforementioned, in Canonical Modelling, we employ InstantNGP to reconstruct the canonical state of an interactive

object. In the Movable Part Decomposition stage, we model various individual states, each sharing non-interactive parts but featuring different movable parts. To achieve this, we propose two occupancy grid sampling strategies that balance training memory and speed.

Shared Occupancy Grid. We reuse the occupancy grid from Canonical Modelling by only increasing its learned density values, which represent the manipulated movable parts in each individual state. This strategy saves the memory usage, but experimentally increases the number of sampling points in each ray, which slows down the training process.

Independent Occupancy Grid. In this strategy, we set independent occupancy grids for individual states, initialized with the canonical state, and update them separately. While this approach significantly increases memory usage, it greatly enhances training speed by reducing the number of sampling points.

With these strategies, we can choose the better option depending on computational resources and preferences.

4.6 Overall Objectives

The first two stages of our approach involves training. The learning objective in Canonical Modelling $\mathcal{L}_{CM}$ is as follows:

$$\mathcal{L}_{CM} = \lambda_1 \mathcal{L}_{\text{dist}} + \lambda_2 \mathcal{L}_{\text{opacity}} + \mathcal{L}_{MSE}, \quad (9)$$

where λ_1 and λ_2 are the weighting coefficients, both set to 1e-3 in our experiments. In Movable Part Decomposition, an individual state S_i is randomly selected for training. We employ both $\mathcal{L}_{CM}$ and the learning objective $\mathcal{L}_{MPD}$ to optimize the model:

$$\mathcal{L}_{MPD} = \lambda_3 \mathcal{L}_{consis} + \lambda_4 \mathcal{L}_{smooth} + \lambda_5 \mathcal{L}_{MSR}, \quad (10)$$

where λ_3, λ_4 and λ_5 are set as 1e-4, 1e-3, and 1e-2, respectively.

5 Experiments

5.1 Implementation Details

We employ the Adam optimizer [Kingma and Ba, 2014] with hyperparameters $lr = $ 1e-2, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon =$1e-15 to train our model. The image resolution in our dataset is 1280×680. Each epoch consists of 1,000 iterations and 8,192 pixels are randomly selected for each iteration. The Canonical Modelling stage requires two epochs to reconstruct the canonical shape of an interactive object. During the Movable Part Decomposition stage, a random individual state is selected for each iteration to learn the corresponding movable part. This phase takes 13 epochs. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

5.2 Case Study

Car. As shown in Figure 6, the training states of the Car object consist of the canonical state S_0, the front door open state S_1, and the rear door open state S_2. The combination state $S_1 + S_2$, where both doors are open simultaneously, is

entirely synthesized. The view visualizations showcase our efficacy.

Display Case. Reconstructing a transparent object is inherently challenging, even without interactive components. The Display Case, illustrated in Figure 7, features two glass doors as movable components and contains numerous small gadgets, further complicating the scenario. Synthesized views from various poses demonstrate the effectiveness of our method in handling this complex setting.

Drawer. The Drawer object, depicted in Figure 8, consists of three vertically arranged drawers, resulting in four novel combination states. The synthesized views across different states demonstrate consistent internal structures and external appearances.

Furniture. Figure 9 illustrates the Furniture scene, which includes three types of cabinets. Each cabinet features a movable part located on the front or the top and is surrounded by a pile of sand. The synthesized combination states demonstrate realistic cabinet appearances and consistent background rendering. Additionally, Figure 10 presents rendered images from varying camera poses of multiple synthesized combination states, demonstrating that our method effectively generates geometrically consistent arbitrary novel states.
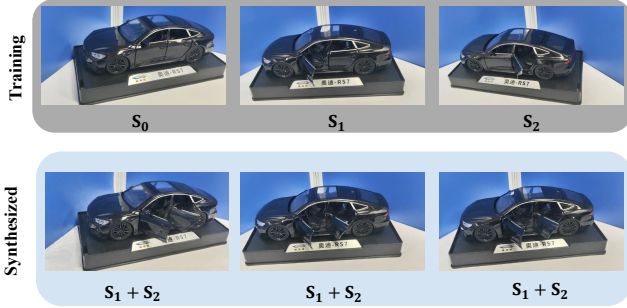

Figure 6: Training and synthesized states for the Car object in *Inter3D*. The canonical state and the individual states of two movable parts are observed during training. Our method successfully synthesizes various views of the novel combination state.
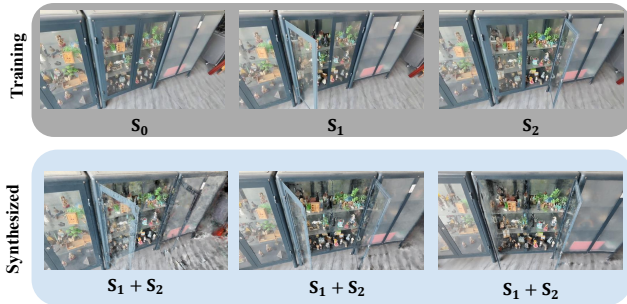

Figure 7: Training and synthesized states for the Display Case in *Inter3D*. Clearly, our method effectively handle the transparent object containing intricate gadgets.
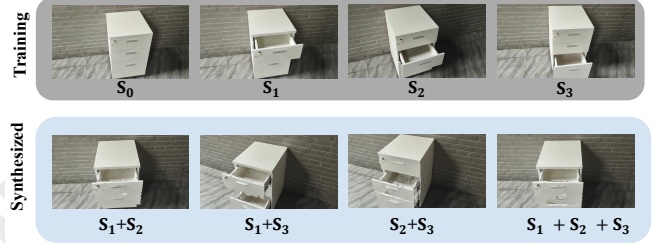

Figure 8: Training and synthesized states for the Drawer in *Inter3D*. Clearly, our method effectively model the complex object with more stacked movable parts.
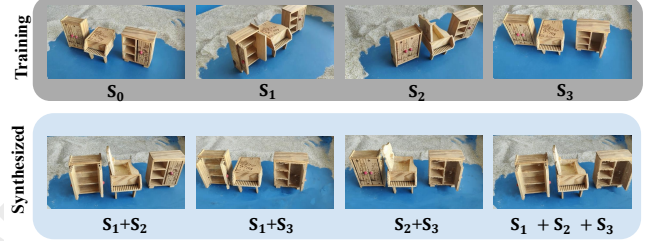

Figure 9: Training and synthesized states for the Furniture in *Inter3D*. The results demonstrate that our method can effectively model a highly challenging scene featuring three types of cabinets and a cluttered background.

5.3 Comparison with State-of-the-Arts

As the first work focusing on the novel state synthesis of a human-interactive object, we compare our method with three typical counterparts for 3D or 4D reconstruction, *i.e.,* Instant-NGP [Müller *et al.*, 2022], 3D Gaussian [Kerbl *et al.*, 2023], and 4D Gaussian [Wu *et al.*, 2024], as shown in Figure 11. All methods are trained on the same states as ours, *i.e.,* the canonical state and individual states. For our method, we present the synthesized combination state where all movable parts are manipulated. For existing methods, we select the closest view to the combination state for comparison, as they
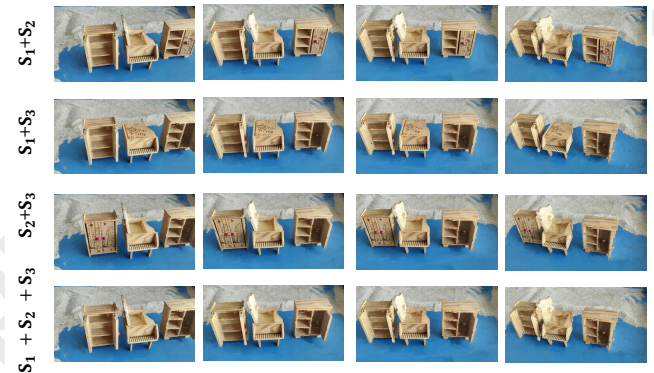

Figure 10: Synthesized combination states of the Furniture object in *Inter3D* under varying viewpoints. The visualizations demonstrate that our method produces realistic and geometrically consistent novel states across different views.

Scene	Car		Displayer Case		Drawer		Furniture	
Strategies | Metric	MA/MB	TT/s	MA/MB	TT/s	MA/MB	TT/s	MA/MB	TT/s
Independent	318.28	**969.44**	318.28	**1 032.37**	335.03	**939.98**	335.03	**1 120.48**
Share	**284.78**	1 264.99	**284.78**	1 717.74	**285.53**	1 386.01	**285.53**	1 627.35

Table 1: Ablation study on the proposed occupancy grid sampling strategies. Across all scenes, the Independent strategy significantly reduces training time, while the Shared strategy requires considerably less memory. The best results are highlighted in bold.

lack the inherent designs of synthesizing novel states.

The highlighted visualizations in Figure 11 reveal that existing methods fail to generate unseen states during training. For instance, in the Car object, InstantNGP simply overlays the movable parts, *i.e.,* the front and rear doors, while 3D Gaussian blends the doors together, failing to "open the doors" and resulting in visible quality degradation and prominent artifacts. Similarly, 4D Gaussian associates each individual state with time but still suffers from state overlap. In contrast, our method successfully renders the combination state where both car doors are open, and produce much higher-quality images. In the Drawer object, InstantNGP reconstructs the background wall but fails to model the open drawers. 3D Gaussian is unable to reconstruct the scene entirely, missing both the environment and the open drawers. 4D Gaussian, on the other hand, produces significant ghosting artifacts in the drawers. In comparison, our method successfully synthesizes the novel state with all drawers open. Furthermore, with the effective Movable Part Decomposition and Mutual State Regularization, our method achieves superior rendering quality on novel combination states.

Figure 12: Ablation study on the proposed Mutual State Regularization and occupancy grid sampling strategies. The first column illustrates the camera trajectory perturbations across training states, which significantly degrade modeling efficacy, particularly for interactive objects with more movable parts. In contrast, the MSR mechanism effectively mitigates these negative effects. Additionally, the last two columns demonstrate that both the Shared and Independent Occupancy Grid sampling strategies perform very well.
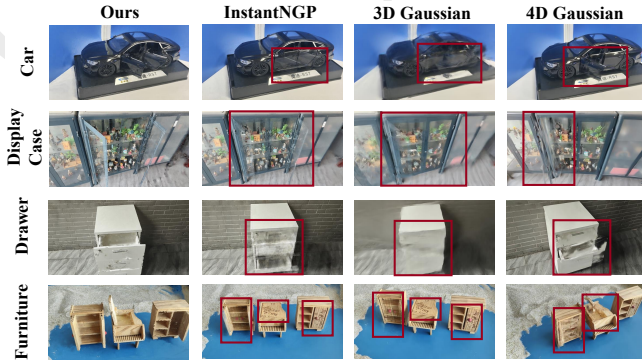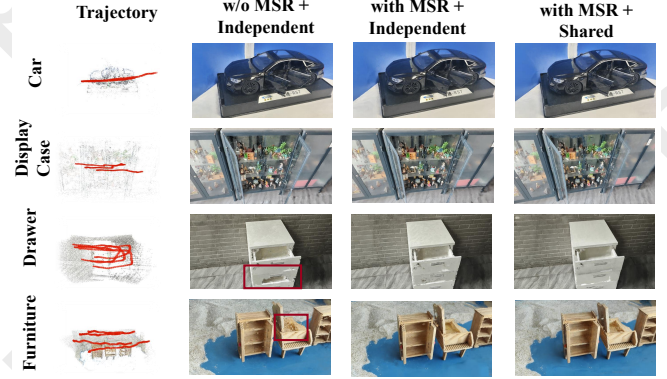

Figure 11: Comparison between our method and typical 3D and dynamic 4D object reconstruction methods. The highlighted regions demonstrate that existing methods fail to generalize to novel combination states, whereas our method successfully generates high-fidelity images of unseen states.

5.4 Ablation Study

As shown in Figure 12, we conduct ablation experiments on the proposed Mutual State Regularization (MSR) and occupancy grid sampling strategies.
Effect of Mutual State Regularization. The first column shows the camera trajectories used to capture training state images. The next two columns compare the synthesized novel states under the Independent Occupancy Grid. Evidently, tra-

jectory inconsistencies lead to severe modeling failures, especially for objects with multiple movable parts, *i.e.,* Furniture and Drawer. Without MSR, noticeable artifacts appear in the highlighted regions of the Furniture, while missing regions are observed in the Drawer object.
Effect of Occupancy Grid Sampling Strategies. The last two columns of Figure 12 present the ablation results for the Independent and Shared Occupancy Grid sampling strategies. No significant differences in rendering quality are observed, demonstrating that both strategies are well-suited for our task. A detailed comparison of these strategies in terms of Memory Allocation (MA) and Training Time (TT) is provided in Table 1. Across all objects, the Independent strategy significantly reduces training time but requires more memory, while the Shared strategy exhibits the opposite behavior.

6 Conclusion

In this work, we present *Inter3D*, a comprehensive benchmark and a strong baseline method for *novel state synthesis* of human-interactive objects with multiple movable parts. To facilitate this research, we construct a diverse dataset comprising commonly encountered but challenging interactive objects from everyday life. On this dataset, we introduce a novel and practical evaluation scenario, *i.e.,* novel state synthesis, where only canonical and individual part states are observed during training, while combination states are withheld for evaluation. To efficiently model the full range of states of

an interactive object, we leverage Space Discrepancy Tensors with multi-resolution hash encoding, enabling precise learning of individual states. To ensure movable part consistency across training states, even under trajectory perturbations, we propose a Mutual State Regularization mechanism. Furthermore, we explore two occupancy grid sampling schemes to balance memory efficiency and training speed. Extensive experiments demonstrate that our approach achieves high-fidelity novel state synthesis, significantly outperforming existing methods. These results establish a strong foundation for advancing research in interactive 3D object reconstruction and open new avenues for exploring 3D object synthesis.

Acknowledgements

Contribution Statement

This work was a collaborative effort by all contributing authors. Gan Chen and Ying He made equal contributions to this study and are designated as co-first authors. Gang Xu and Ming Li, serving as the corresponding authors, are responsible for all communications related to this manuscript.

References

[Barron *et al.*, 2022] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.

[Berger *et al.*, 2017] Matthew Berger, Andrea Tagliasacchi, Lee M Seversky, Pierre Alliez, Gael Guennebaud, Joshua A Levine, Andrei Sharf, and Claudio T Silva. A survey of surface reconstruction from point clouds. In *Computer graphics forum*, volume 36, pages 301–329. Wiley Online Library, 2017.

[Cameirão *et al.*, 2009] Mónica S Cameirão, Sergi Bermúdez i Badia, Esther Duarte Oller, and Paul FMJ Verschure. The rehabilitation gaming system: a review. *Advanced Technologies in Rehabilitation*, pages 65–83, 2009.

[Fisher *et al.*, 2021] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L Palmer. Colmap: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021.

[Fridovich-Keil *et al.*, 2023a] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[Fridovich-Keil *et al.*, 2023b] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Liu *et al.*, 2024] Xiangyue Liu, Han Xue, Kunming Luo, Ping Tan, and Li Yi. Genn2n: Generative nerf2nerf translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5105–5114, 2024.

[Luiten *et al.*, 2024] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024.

[Machinery, 1950] Computing Machinery. Computing machinery and intelligence-am turing. *Mind*, 59(236):433, 1950.

[Mildenhall *et al.*, 2021] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[Müller *et al.*, 2022] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[Poole *et al.*, 2022] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[Pumarola *et al.*, 2021] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[Ravi *et al.*, 2024a] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Ravi *et al.*, 2024b] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Wohlgenannt *et al.*, 2020] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. Virtual reality. *Business & Information Systems Engineering*, 62:455–461, 2020.

[Wu *et al.*, 2024] 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.