3D Scene Change Modeling With Consistent Multi-View Aggregation

Zirui Zhou[1] Junfeng Ni[1,2] Shujie Zhang[1] Yixin Chen[2,†,✉] Siyuan Huang[2,✉]
[†] Project lead [✉] Corresponding author
[1] Tsinghua University [2] State Key Laboratory of General Artificial Intelligence, BIGAI
https://zr-zhou0o0.github.io/SCaR3D/

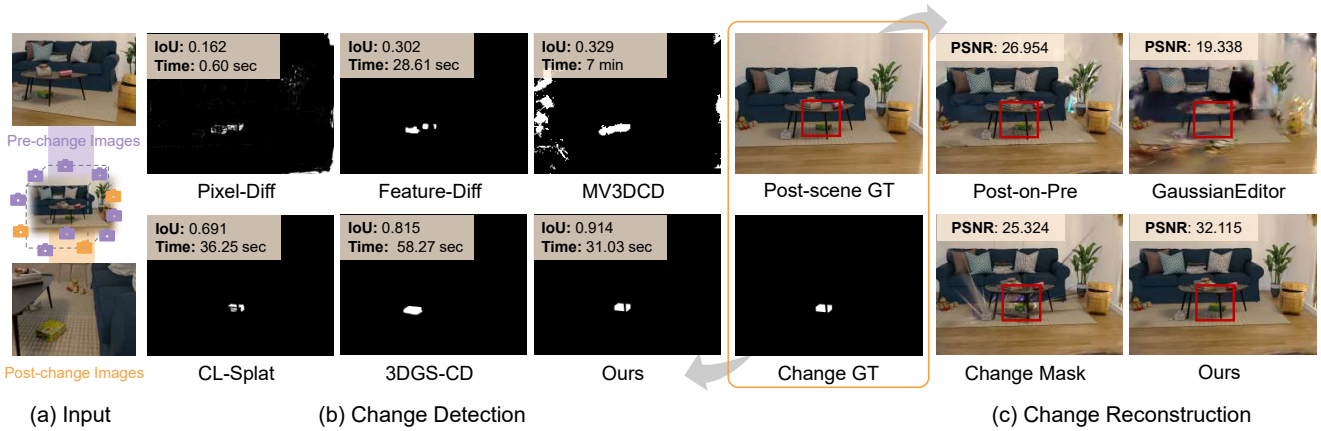

Figure 1. We propose **SCaR-3D**, a 3D scene change modeling framework that detects changes from dense-view pre-change images and sparse-view post-change images, while seamlessly reconstructing the post-change scene. **SCaR-3D** significantly outperforms existing 3D change detection methods in change mask accuracy and computational efficiency, and delivers high-quality continual reconstructions.

Abstract

Change detection plays a vital role in scene monitoring, exploration, and continual reconstruction. Existing 3D change detection methods often exhibit spatial inconsistency in the detected changes and fail to explicitly separate pre- and post-change states. To address these limitations, we propose **SCaR-3D**, *a novel 3D scene change detection framework that identifies object-level changes from a dense-view pre-change image sequence and sparse-view post-change images. Our approach consists of a signed-distance–based 2D differencing module followed by multi-view aggregation with voting and pruning, leveraging the consistent nature of 3DGS to robustly separate pre- and post-change states. We further develop a continual scene reconstruction strategy that selectively updates dynamic regions while preserving the unchanged areas. We also contribute* **CCS3D**, *a challenging synthetic dataset that allows flexible combinations of 3D change types to support controlled evaluations. Extensive experiments demonstrate that our method achieves both high accuracy and efficiency, outperforming existing methods.*

1. Introduction

3D reconstruction [27, 39, 47, 55, 59] is a fundamental task in computer vision, playing a crucial role in visual perception, embodied artificial intelligence (EAI), environment monitoring, and AR/VR [22, 32, 33, 40, 61]. Real-world environments are inherently dynamic, where objects may appear, disappear, or translate and rotate over time. Much like a Sherlockian observer piecing together a scene from the smallest clues, a robust 3D reconstruction system must detect and interpret subtle environmental changes from the sparse, new observations through the lens of its 3D representation. Thus, reliable 3D change detection is essential to maintain an up-to-date and accurate representation of the evolving scenes.

Change detection aims to identify objects in a scene that have changed between two time points, given multi-view images captured before and after the change. Previously, 2D change detection has been extensively studied, particularly in remote sensing applications such as monitoring land use changes, including the construction of buildings or roads [31]. However, these methods face significant lim-

itations when applied directly to 3D scenes. First, most 2D approaches rely on supervised learning with annotated datasets, which are costly to create and often lack generalizability across different environments. Besides, these methods often struggle to maintain consistency across multiple views due to random noise and visual ambiguities, limiting their effectiveness in identifying coherent 3D changes.

Recent studies introduce 3D representations into change detection, with 3D Gaussian Splatting (3DGS) [27] emerging as a particularly prominent approach. 3DGS enables efficient rendering of pre-change scenes from novel viewpoints in the post-change, and its explicit and editable structure facilitates the seamless identification and modification of the changed regions. Building on this, methods [1, 18] encode change indicators directly into Gaussian primitives, yielding a unified representation of altered regions. However, such Gaussian-level representations lack holistic object-awareness, often producing fragmented change masks and view-dependent inconsistencies when representing the same object. To mitigate these issues, 3DGS-CD [34] proposes identifying pre-change object masks using segmentation confidence, followed by learning pose transformations between pre- and post-change states. However, accurately matching masks of the same object consistently across multi-views remains challenging, leading to notable performance degradation under diverse change types such as translation, insertion, and removal.

Our method, Spotting Changes and Reconstruction in 3D Scenes (**SCaR-3D**), is a multi-view voting-and-validation-based framework for efficient and consistent change detection in complicated and large-scale 3D scenes. Given two image sets captured from arbitrary viewpoints before and after scene changes, we first register their camera poses within a unified coordinate system. We then identify the feature-level difference between the pre-change and post-change observations by computing signed distance metrics. Utilizing a voting-based approach, we aggregate 2D differences from multiple perspectives, and suppress noise and ensure geometric coherence via multi-view voting and pruning operations. The pruning strategy also robustly separates the pre-change and post-change difference. Finally, we leverage EfficientSAM's segmentation capability to validate the 3D differences and extract high-confidence change masks. By integrating these masks into the 3D reconstruction pipeline, our method enables continual reconstruction in regions where changes have occurred, while preserving the integrity of unchanged areas.

We introduce a new synthetic dataset tailored for 3D scene change detection, featuring complex and diverse indoor environments beyond simple tabletop settings. The dataset is fully editable, allowing flexible combinations of change types to support controlled evaluations. To assess the effectiveness of our method, we conduct experiments on both real-world datasets and synthetic datasets. Compared to existing methods, our approach produces more accurate and view-consistent change masks with higher efficiency.

In summary, our key contributions are as follows:

- We propose a novel 3D scene change detection framework that leverages a 3D difference map and a multi-view consistency validation mechanism to accurately and efficiently identify object-level changes from two sequences captured under arbitrary viewpoints.
- We construct a high-quality synthetic dataset, **CCS3D**, comprising editable indoor scenes for controlled evaluation of various 3D change types in complex environments.
- Extensive experiments demonstrate that our method outperforms previous approaches in terms of detection accuracy, change mask quality, and computational efficiency.

2. Related Work

2.1. Change Detection

Change detection involves identifying regions or objects that exhibit differences by comparing images taken before and after the changes occur. 2D change detection from paired images has been a long-studied problem, with traditional methods such as [5, 8, 35], and deep learning approaches [3, 13, 15, 21].

Early 3D change detection methods, including image-based geometric approaches [43, 44, 46, 50, 51] and TSDF-based detection [16], provide important foundations but rely heavily on cadastral models or satellite-derived imagery, where pose and alignment errors are common. More recent geometric-consistency approaches also exhibit limitations in the types of changes they can handle: [2, 34] assume that all changed objects appear in both pre- and post-change views to estimate rigid transformations, while [29] does not account for object removals.

With the emergence of NeRF [39] and 3DGS [27], change detection can now operate on well-reconstructed scenes. For instance, [23, 38] train separate NeRFs on pre- and post-change images to detect changes from aligned views.[34] aggregates 2D change masks into a 3D point cloud to learn pose changes, while [18] embeds change channels in 3DGS. Our method introduces an effective voting strategy to initialize a 3D difference map on 3DGS, validated by multi-view checks and segmentation confidence, enabling fast and accurate change detection.

2.2. Continual Scene Reconstruction

Continual 3D reconstruction aims to model a continuously updated 3D scene or its static background from an image sequence taken in dynamic environments [12]. However, directly training a scene representation over the sequence causes catastrophic forgetting [12, 30] and degradation [54]. To mitigate this, Li et al. [30] and Cai et al. [6] introduce a

keyframe database for historical image replaying. Another key topic for continual 3D reconstruction is to identify the transient regions to be excluded during model update. Traditional methods rely on depth residuals [45] and pixel difference [17]. Learning based methods include [6, 28, 30] masking the transient objects with a learned classifier to maintain reconstruction consistency. Others [1, 28] exploits off-the-shelf vision model [7, 57] to identify the dynamic region. These methods highlight that effective change detection is essential for maintaining accurate and up-to-date 3D scene reconstructions over time, motivating our approach to integrate change detection with continual reconstruction.

2.3. 3D Editing

3D editing refers to modifying specific parts of a reconstructed scene. Traditional 3D editing relies on human-operated tools such as Maya and Blender. For neural implicit representations such as NeRF and 3DGS, existing approaches primarily focus on text-driven [14, 20, 41, 42, 52, 56, 58, 62] and image-based [4, 24, 52, 62] 3D editing. While existing methods provide stable edits and user-friendly interaction, they lack precise object insertion capabilities and depend on manual initiation. An alternative paradigm for 3D editing involves segmenting all objects in the scene, followed by selective editing of the targeted objects [9, 19, 20, 26, 49, 60]. However, when only a few objects in a cluttered scene require editing, this approach leads to significant computational overhead. Our method leverages scene change detection to automatically trigger precise 3D edits, enabling efficient and targeted modifications by localizing updates to the detected change regions.

3. Method

An overview of our method is shown in Fig. 2. Given pre-change and post-change image sets of a scene, we aim to detect object-level 3D changes. We first estimate camera poses and render the paired pre-post images (Sec. 3.2). We then compute signed-distance maps for coarse 2D differences (Sec. 3.3), which are aggregated into 3D differences through multi-view voting and validation (Sec. 3.4). The resulting change masks then guide 3D updates, enabling stable and accurate continual reconstruction (Sec. 3.5).

3.1. Problem Setup

The input consists of two image sets: $\mathcal{I}_{pre} = \{I_i \mid i = 1, \dots, n_{pre}\}$ captured from the pre-change scene under n_{pre} arbitrary viewpoints, and $\mathcal{I}_{post} = \{I'_i \mid i = 1, \dots, n_{post}\}$ from the post-change scene under n_{post} viewpoints. We emphasize that $\mathcal{I}_{pre}$ represents a densely sampled set of views, whereas $\mathcal{I}_{post}$ corresponds to a sparsely sampled one. Our goal is to generate a set of change masks $\mathcal{C} = \{C_i \mid i = 1, \dots, n_{test}\}$ under specified target viewpoints, and reconstruct the 3D scene $\mathcal{G}_{post}$ in 3DGS.

3.2. Image Registration

For $\mathcal{I}_{pre}$ and $\mathcal{I}_{post}$, we first leverage the structure-from-motion (SFM) algorithm [47], *e.g.*, COLMAP [48], to simultaneously estimate their camera poses $\mathcal{P}_{pre}$ and $\mathcal{P}_{post}$. Assuming that the majority of scene features remain unchanged, we jointly register both image sets in a single SfM process to ensure that all estimated poses lie within a unified coordinate system. Additional implementation details are provided in *supplementary*.

We then train a 3DGS model using the pre-change image set $\mathcal{I}_{pre}$ to obtain a pre-change 3DGS $\mathcal{G}_{pre}$. We render $\mathcal{G}_{pre}$ from the post-change camera poses $\mathcal{P}_{post}$, producing $\mathcal{I}_{ren}$, where each rendered image in $\mathcal{I}_{ren}$ is paired with its corresponding real image in $\mathcal{I}_{post}$.

3.3. 2D Difference Generation

Feature Extraction We utilize EfficientSAM [57] to extract image features f from given image I:

$$f = \mathcal{F}(I), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the image encoder of EfficientSAM and $f \in R^{h \times w \times d}$. We bilinearly upsample EfficientSAM's raw feature maps to the image resolution while keeping the embedding dimension d.

Signed Distance-Based Change Localization To capture the directionality of changes, we adopt a signed distance formulation in the feature space. Specifically, for each feature map pair (f_i, f'_i), obtained from pairs of rendered pre-change image and post-change image via Eq. (1), we apply Principal Component Analysis (PCA) [37] to determine the dominant direction of variation. All pixel-wise feature vectors from f_i and f'_i are collected, and the first principal component vector v is extracted as the direction of maximum variance. For each pixel p, the signed distance between pre- and post-change features is computed by projecting the feature difference onto v:

$$D_i^p = \frac{(f_i^p - f_i'^p) \cdot v}{\|v\|}. \tag{2}$$

Since the signed distance D_i^p separates foreground and background [11, 36], we threshold it to obtain two directional binary change masks:

$$\mathcal{M}_{i,1} := \mathbf{1}\{\, D_i^p > \epsilon_1 \,\}, \tag{3}$$

$$\mathcal{M}_{i,2} := \mathbf{1}\{\, D_i^p < \epsilon_2 \,\}. \tag{4}$$

$\mathbf{1}\{\cdot\}$ is indicator function, and $\epsilon_1 \geq 0 \geq \epsilon_2$ are thresholds.

3.4. 3D Difference Aggregation

Although the 2D difference masks $\mathcal{M}_{i,1}$ and $\mathcal{M}_{i,2}$ incorporate image-level semantic features and are more robust than raw pixel-level comparisons, they still suffer from noise and

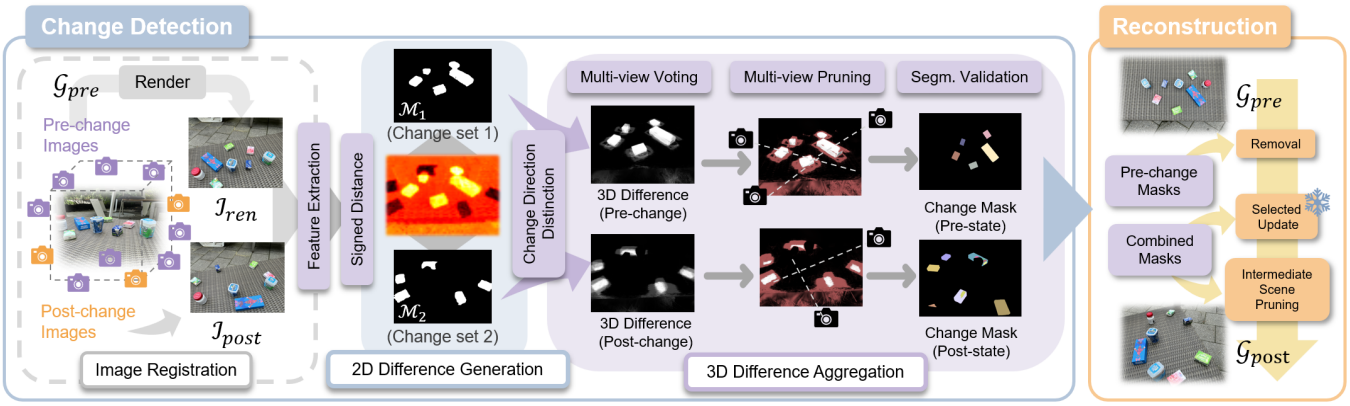

Figure 2. **Overview of SCAR-3D.** We first employ COLMAP for image registration, producing paired pre-change renders and post-change captures. In the *2D Difference Generation* stage, features are extracted and a signed distance metric is applied to separate the change regions into two sets. After that, the *3D Difference Aggregation* stage integrates multi-view differences through voting, pruning, and segmentation validation. Finally, the change masks are applied to the reconstruction process to update the 3D scene selectively.

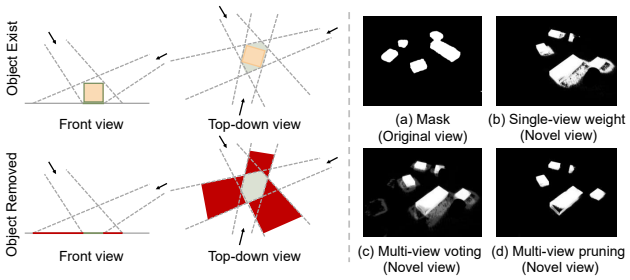

Figure 3. **Multi-view pruning.** Left: When objects exist (top), masks are scattered and few Gaussians are removed; when objects are removed (bottom), background tracing leads to extensive pruning. Right: Novel-view 3D difference visualization for (b) single-view weighting in Eq. (5), (c) multi-view voting in Eq. (7), and (d) multi-view pruning.

viewpoint-induced ambiguities. To mitigate these issues, we aggregate the 2D differences into a unified 3D representation, leveraging spatial consistency across multiple views.

Multi-view Voting To aggregate the 2D differences into 3D, we initialize a 3D difference representation based on the pre-trained pre-change 3DGS model $\mathcal{G}_{pre}$. Specifically, we embed an additional difference channel into each Gaussian to indicate whether it has changed.

For every single view, following the semantic tracing method introduced in GaussianEditor [14], we identify and update the relevant Gaussians for each 2D mask by computing their contribution:

$$w_i = \sum_p o_i(p) \cdot T_i(p) \cdot M(p), \tag{5}$$

where w_i represents the weight of the i-th Gaussian, $o_i(p)$, and $T_i(p)$ denote the Gaussian's opacity, transmittance from pixel p, and $M(p)$ the 2D mask of pixel p. To normalize the weights, we define $\tilde{w}_i = \frac{w_i}{w_{\max}}$, where $w_{\max}$ is the maximum weight across all Gaussians in the current view, ensuring that $\tilde{w}_i \in [0, 1]$.

For the multi-view setting, a straightforward approach is to aggregate weights from all post-change views. Let $S_i^k = \sum_p o_i^k(p)\, T_i^k(p)\, M^k(p)$ denote the aggregated contribution of the i-th Gaussian from the k-th view. A simple normalization by the total number of post-change views is:

$$w_i = \frac{1}{n_{\text{post}}} \sum_{k=1}^{n_{\text{post}}} \frac{S_i^k}{w_{\max}^k} \tag{6}$$

However, this uniform normalization by n_{post} introduces bias against Gaussians visible in fewer views due to occlusions or restricted fields of view, assigning them disproportionately low weights. To address this, we adopt a visibility-aware strategy, normalizing each Gaussian's weight by the actual number of views in which it is observed, n_i^{seen}:

$$w_i = \frac{1}{n_i^{\text{seen}}} \sum_{k=1}^{n_i^{\text{seen}}} \frac{S_i^k}{w_{\max}^k}. \tag{7}$$

Our method leverages the fast GPU Radix sort algorithm already implemented in the 3DGS rendering pipeline, enabling us to process a scene in just a few seconds, whereas methods such as GaussianCut [25] require several minutes per scene and incur additional memory overhead.

Multi-view Pruning We observe from Fig. 3 that the accumulated weights lack object-level awareness. In some cases, the weights erroneously bleed through foreground objects and are projected onto the background, leading to inconsistent aggregation. To mitigate this issue, we perform a multi-view consistent pruning step after the voting

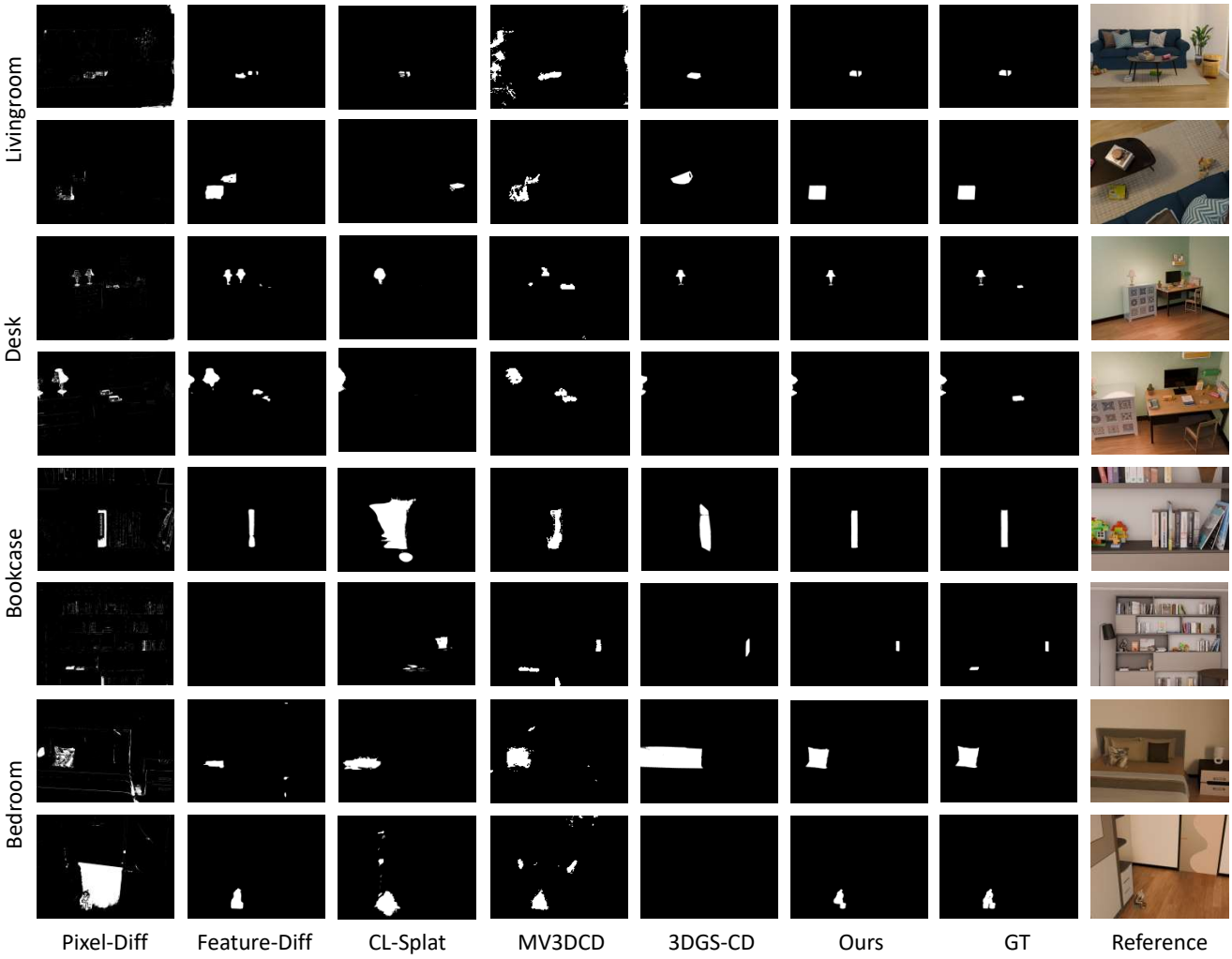

Figure 4. **Qualitative change detection results on the CCS3D dataset.** Each pair of rows corresponds to a single scene captured from different viewpoints. The last column, labeled *Reference*, shows the post-change images from the matched viewpoints.

process. Specifically, we remove Gaussians whose centers are projected outside the 2D masks in more than τ out of the total n_{post} views.

Let μ_i denote the center of the i-th Gaussian in 3D space, and p_k^i be its projection, *i.e.*, the pixel location, onto the k-th view. For each view k, we define an indicator $\delta_k^i = 1$ if the 2D mask value $M^k(p_k^i)$ at pixel p_k^i is zero (*i.e.*, outside the mask), and $\delta_k^i = 0$ otherwise. The total number of views in which the Gaussian center lies outside the mask is:

$$n_{\text{out}}^i = \sum_{k=1}^{n_i^{seen}} \delta_k^i. \tag{8}$$

Finally, the i-th Gaussian is pruned if $n_{\text{out}}^i > \tau \cdot n_i^{seen}$.

Change Direction Distinction As introduced in Sec. 3.3, we use signed distances to separate change regions into two directional categories, but we cannot yet determine which category indicates object addition and which indicates object removal. Since added objects can only be found in post-change scenes, and removed objects can only be found in pre-change scenes, we combine the multi-view pruning strategy in Sec. 3.4 with the following procedure to further infer whether each directional change mask corresponds to objects in the pre-change scene or the post-change scene.

Let N denote the total number of Gaussians selected during the voting process, and let N_p denote the number of Gaussians pruned during the consistency check. We define the retention rate as $R = \frac{N-N_p}{N}$. As illustrated in Fig. 3, when an object exists at the hypothesized location, the projected mask aligns well with the actual object surface, resulting in a higher weight concentration and a higher retention rate R. Conversely, when the object is absent, weights

are more likely to scatter and get pruned, leading to a lower R. Applying 3D difference tracing to the pre-change Gaussians ($\mathcal{G}_{pre}$) associates a higher retention rate (R) with the pre-change state and a lower rate with the post-change state.

Segmentation Validation To estimate the reliability of the 3D difference map and refine it into a high-quality 2D mask, we validate the difference against predictions from EfficientSAM. For each view, we simply select the EfficientSAM mask with the highest IoU against the projected 3D difference.

3.5. Continual Scene Reconstruction

Updating $\mathcal{G}_{pre}$ to $\mathcal{G}_{post}$ presents several challenges. First, given the sparsity of post-change views, it is crucial to prevent degradation of Gaussians that are unseen in the post-change images. A straightforward approach of mixing post-change images with pre-change images mitigates this issue, but it significantly increases computational cost and introduces view-dependent ambiguities within the changed regions. Alternatively, freezing Gaussians outside the change masks is a common strategy; however, insufficient supervision near the boundaries of the 2D change masks often leads to drifting Gaussians along the edges of changed areas.

To mitigate these problems, we adopt a 2D change mask–based strategy that replaces pre-change objects with their post-change counterparts during reconstruction. Specifically, we first apply the pre-change masks to $\mathcal{G}_{pre}$ to remove the corresponding objects. Then, we use the masked post-change images $\mathcal{I}_{post}$ to locally update $\mathcal{G}_{pre}$, producing an intermediate scene. In this process, Gaussians outside the 2D change masks are frozen, while loss computation and gradient descent are restricted to the masked regions in screen space. Finally, leveraging the voting and pruning method introduced in Sec. 3.4, we replace the change region in $\mathcal{G}_{pre}$ with the corresponding regions from the intermediate scene. This results in the updated post-change scene $\mathcal{G}_{post}$ with fewer drifting artifacts from view-dependent degradation while accurately relocating the changed objects to their new position.

4. Experiments

4.1. Experiment Setup

Datasets We first introduce a new dataset, Controllable Change in 3D Scenes (**CCS3D**), which comprises four diverse and comprehensive synthetic scenes: Desk, Bookcase, Livingroom, and Bedroom, constructed with Blender [10]. Compared with existing 3D change detection datasets, **CCS3D** is not restricted to tabletop scenarios with simple camera trajectories (*e.g.*, face-forward or fixed 360° rotation). The Bookcase scene features a multi-floor bookshelf, where the camera navigates from a distant view to a close-up inspection, sequentially exploring each shelf. The Livingroom and Bedroom scenes offer complete 360° environments containing both large-scale furniture (*e.g.*, chairs, tables) and small tabletop items (*e.g.*, books, pencil cases), designed to support fine-grained change detection. They incorporate complex, human-like camera navigation patterns, simulating natural walking and exploration. Furthermore, our dataset enables controlled experiments on change detection with varying numbers of objects and change types, where such control is difficult to achieve in existing 3D change detection datasets.

We also evaluate our model on the real-world dataset, 3DGS-CD [34] dataset, which focuses on tabletop scenes with complex object changes, including object removal, insertion, and movement in cluttered environments.

Baselines and Metrics We compare our method with 2D-based approaches, *i.e.*, MTP [53], and 3D-based methods 3DGS-CD [34], MV3DCD [18], and the change detection module of CLSplat [1]. In addition, we evaluate two baselines based on pixel difference (Pixel-Diff) and feature difference (Feature-Diff). Following the evaluation protocols in C-NeRF [23] and 3DGS-CD [34], we report Precision, Recall, F1-score, and IoU as our quantitative metrics.

4.2. Evaluation Results

CCS3D Results in Tab. 1 show that our method consistently achieves the highest Precision and IoU across all scenes. 2D pixel and feature differences yield stable but worse results due to noise from Gaussian splatting artifacts. MV3DCD's Gaussian-wise change representation often leads to fragmented detections lacking object integrity. Furthermore, 3DGS-CD's performance is highly sensitive to its object matching and pose estimation steps, especially in challenging scenarios. For instance, it fails in the Bedroom scene, where limited viewpoints lead to critical matching errors, and struggles to distinguish visually similar books in the Bookcase scene, causing a significant drop in IoU. In contrast, our approach uses 3D difference aggregation and segmentation validation to effectively suppress such noise and recover complete, accurate change masks.

Controlled Experiments We evaluate our method under varying change types and object counts. Tab. 2 shows that performance remains strong for simple scenarios (*e.g.*, single-object cases) across all change types. However, as the number of changed objects increases, performance degrades, especially for Rotation and Translation. Mixed changes are the most challenging overall, yielding lower scores and compounded difficulty under heterogeneous transformations. This also validates the utility of our dataset in examining algorithm robustness in more complex cases.

3DGS-CD Dataset Results on the 3DGS-CD dataset (Tab. 3) show our method achieves significantly higher F1 and IoU scores than prior approaches. While MTP attains

Table 1. **Quantitative change detection results on the CCS3D dataset.** The best, second-best, and third-best scores are highlighted in red, orange, and yellow, respectively. Our method demonstrates the best overall performance.

	Livingroom		Desk		Bookcase		Bedroom		Average	
Method	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
Pixel-Diff	0.273	0.162	0.398	0.254	0.315	0.201	0.286	0.176	0.318	0.198
Feature-Diff	0.420	0.302	0.480	0.323	0.320	0.256	0.705	0.584	0.450	0.343
CL-Splat	0.789	0.657	0.567	0.399	0.294	0.199	0.501	0.341	0.538	0.399
MV3DCD	0.478	0.329	0.291	0.178	0.449	0.295	0.547	0.413	0.441	0.304
3DGS-CD	0.897	0.815	0.525	0.408	0.477	0.353	0.148	0.089	0.512	0.416
Ours	0.955	0.914	0.610	0.477	0.423	0.377	0.909	0.834	0.724	0.650

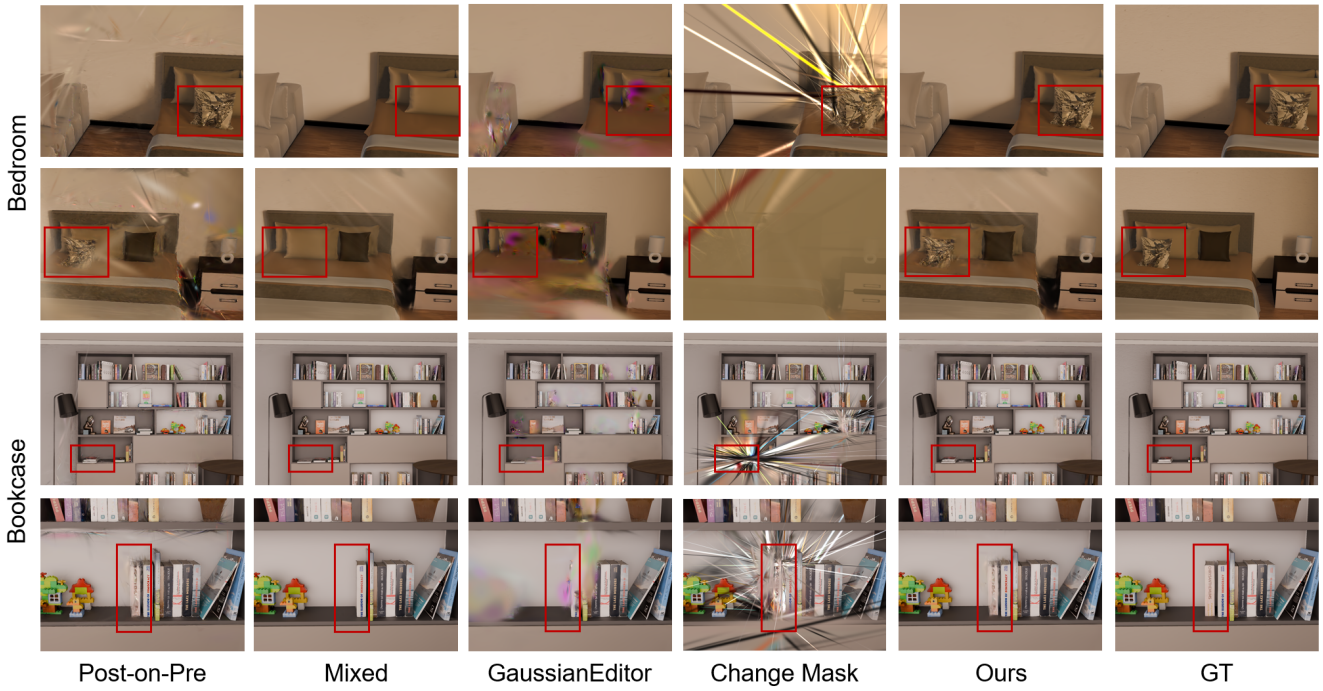

Figure 5. **Qualitative reconstruction results on novel views.** Each pair of rows corresponds to a single scene captured from different novel viewpoints. The changed regions are highlighted with red boxes.

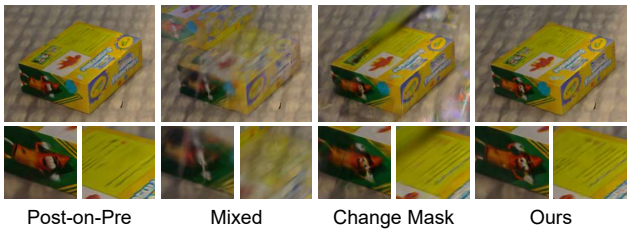

Figure 6. **Reconstruction results on post-change views.**

Table 2. **Controlled Evaluation on the Livingroom scene.** We report F1/IoU across change types and number of changed objects.

	# Obj = 1		# Obj = 2		# Obj = 4	
Change Type	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**
In/Out	0.953	0.911	0.770	0.742	0.764	0.732
Translation	0.985	0.970	0.613	0.498	0.585	0.472
Rotation	0.982	0.964	0.690	0.613	0.383	0.253
Mixed	–	–	0.604	0.582	0.408	0.260

high precision, its lower recall limits performance. The 3DGS-CD method localizes changes accurately but suffers from instability due to reliance on 2D detection and object association steps, especially in cluttered scenes like Mustard and Bench. In contrast, our approach delivers more consistent results by leveraging 3D difference voting and validation to reduce errors from 2D difference detection.

Table 3. **Quantitative change detection results on the 3DGS-CD dataset.** Our method consistently achieves the best performance in terms of F1 score and IoU.

Scene	Method	Precision	Recall	F1	IoU
Mustard	MTP	0.949	0.231	0.371	0.228
	3DGS-CD	0.315	0.104	0.155	0.085
	Ours	0.794	0.573	0.583	0.507
Desk	MTP	0.957	0.344	0.506	0.339
	3DGS-CD	0.967	0.961	0.964	0.930
	Ours	0.995	0.968	0.981	0.964
Swap	MTP	0.942	0.246	0.390	0.243
	3DGS-CD	0.983	0.989	0.986	0.973
	Ours	0.998	0.992	0.995	0.990
Bench	MTP	0.902	0.887	0.895	0.809
	3DGS-CD	0.851	0.796	0.817	0.691
	Ours	0.995	0.867	0.915	0.863
Sill	MTP	0.483	0.308	0.376	0.232
	3DGS-CD	0.981	0.974	0.977	0.956
	Ours	0.998	0.972	0.982	0.970
Average	MTP	0.846	0.403	0.508	0.370
	3DGS-CD	0.819	0.765	0.780	0.727
	Ours	0.956	0.874	0.891	0.859

Table 4. **Quantitative reconstruction results on novel views.** Change-centric crops refer to test-view images cropped using bounding boxes that tightly enclose the ground-truth change regions, while Full scenes correspond to the entire test-view images.

Method	PSNR↑	SSIM↑	LPIPS↓
Change-centric crops			
Post-on-Pre	21.996	0.606	0.333
Mixed	18.251	0.576	0.550
GaussianEditor	17.356	0.554	0.585
Change Mask	13.318	0.445	0.628
Ours	21.837	0.608	0.356
Full scenes			
Post-on-Pre	27.680	0.900	0.173
Mixed	27.982	0.944	0.117
GaussianEditor	22.348	0.867	0.252
Change Mask	16.114	0.730	0.479
Ours	30.304	0.939	0.116

Continual Scene Reconstruction We further evaluate the image quality of our continual reconstruction results. As reported in Tab. 4, our method achieves the highest rendering quality on full-scene novel views and delivers the second-best performance on change-centric crops, demonstrating strong overall robustness across both global and change-focused regions. Directly fine-tuning the pre-change Gaussians $\mathcal{G}pre$ with post-change images $\mathcal{I}post$ suffers from severe view-dependent artifacts in full scenes. Training with mixed pre- and post-change images tends to retain pre-change object states because post-change views are far fewer, leading to lower performance on change-centric crops. Qualitative comparisons are provided in Fig. 5 and Fig. 6. As illustrated in Fig. 5, GaussianEditor [14] often fails to precisely identify the correct new location or state of the changed objects.

4.3. Ablation Study

Tab. 5 presents the results of our ablation study, averaged over the **CCS3D** dataset. The first five rows correspond to vanilla 2D change detection methods commonly adopted in prior works [1, 17, 18, 34], indicating that feature difference is more robust than pixel- and SSIM-based difference, while combining them via multiplication does not yield significant improvement. Compared to 2D difference methods, the 3D difference approaches consistently achieve better performance, demonstrating the effectiveness of multi-view voting in suppressing noise. Our full method, which integrates 2D difference, 3D difference, and segmentation validation, attains the highest accuracy overall.

Table 5. **Ablation study.** We evaluate pixel-, feature-, and SSIM-based differences, as well as their combinations. They are compared against 3D difference and our full method.

Method	Precision	Recall	F1	IoU
2D difference				
Pixel-Diff	0.303	0.458	0.318	0.198
Feature-Diff	0.436	0.538	0.450	0.343
SSIM-Diff	0.316	0.637	0.394	0.256
Pixel+Feature	0.519	0.241	0.308	0.199
SSIM+Feature	0.481	0.355	0.388	0.271
3D difference				
3D-Diff	0.522	0.617	0.487	0.366
Ours (Full)	0.836	0.680	0.724	0.650

5. Conclusion

We propose **SCAR-3D**, a multi-view voting and validation framework for 3D change detection and reconstruction in complex and large-scale 3D scenes. Our method generates accurate and consistent change masks, enabling localized updates to dynamic 3D scenes. Extensive experiments, together with the proposed dataset **CCS3D**, demonstrate that **SCAR-3D** outperforms state-of-the-art methods in both accuracy and efficiency. Future work may address current limitations, including handling a greater number of changed objects within a scene and mitigating the effects of varying lighting conditions and shadows. Furthermore, exploring how to effectively model non-rigid deformations or significant topological changes within the 3D Gaussian framework remains a key challenge.

References

[1] Jan Ackermann, Jonas Kulhanek, Shengqu Cai, Xu Haofei, Marc Pollefeys, Gordon Wetzstein, Leonidas Guibas, and Songyou Peng. Cl-splats: Continual learning gaussian splatting with local optimization. *ICCV*, 2025. 2, 3, 6, 8

[2] Aikaterini Adam, Torsten Sattler, Konstantinos Karantzalos, and Tomas Pajdla. Objects can move: 3d change detection by geometric transformation consistency. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 2

[3] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 2

[4] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023. 3

[5] Francesca Bovolo and Lorenzo Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1):218–236, 2006. 2

[6] Zhipeng Cai and Matthias Müller. Clnerf: Continual learning meets nerf. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23185–23194, 2023. 2, 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[8] Turgay Celik. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE geoscience and remote sensing letters*, 6(4): 772–776, 2009. 2

[9] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1971–1979, 2025. 3

[10] Blender Online Community. Blender - a 3d modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[11] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 2024. 3

[12] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021. 2

[13] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. 2

[14] Jiemin Fang, Junjie Wang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *CVPR*, 2024. 3, 4, 8

[15] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *arXiv preprint arXiv:2209.08290*, 2022. 2

[16] Marius Fehr, Fadri Furrer, Ivan Dryanovski, Jürgen Sturm, Igor Gilitschenski, Roland Siegwart, and Cesar Cadena. Tsdf-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 5237–5244. IEEE, 2017. 2

[17] Bin Fu, Jialin Li, Bin Zhang, Ruiping Wang, and Xilin Chen. Gs-lts: 3d gaussian splatting-based adaptive modeling for long-term service robots, 2025. 3, 8

[18] Chamuditha Jayanga Galappaththige, Jason Lai, Lloyd Windrim, Donald Dansereau, Niko Sunderhauf, and Dimity Miller. Multi-view pose-agnostic change localization with zero labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 11600–11610, 2025. 2, 6, 8

[19] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118*, 2024. 3

[20] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting, 2024. 3

[21] Zipeng Qi Hao Chen and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–14, 2021. 2

[22] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1

[23] Rui Huang, Binbin Jiang, Qingyi Zhao, William Wang, Yuxiang Zhang, and Qing Guo. C-nerf: Representing scene changes as directional consistency difference-based nerf. *arXiv preprint arXiv:2312.02751*, 2023. 2, 6

[24] Vishnu Jaganathan, Hannah Hanyun Huang, Muhammad Zubair Irshad, Varun Jampani, Amit Raj, and Zsolt Kira. Ice-g: Image conditional editing of 3d gaussian splats. *arXiv preprint arXiv:2406.08488*, 2024. 3

[25] Umangi Jain, Ashkan Mirzaei, and Igor Gilitschenski. Gaussiancut: Interactive segmentation via graph cut for 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:89184–89212, 2024. 4

[26] 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. 3

[27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 2

[28] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *NeurIPS*, 2024. 3

[29] Edith Langer, Timothy Patten, and Markus Vincze. Robust and efficient object change detection by combining global semantic information and local geometric verification. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8453–8460. IEEE, 2020. 2

[30] Baicheng Li, Zike Yan, Dong Wu, Hanqing Jiang, and Hongbin Zha. Learn to memorize and to forget: A continual learning perspective of dynamic slam. In *European Conference on Computer Vision*, pages 41–57. Springer, 2024. 2, 3

[31] Kaiyu Li, Xiangyong Cao, Yupeng Deng, Chao Pang, Zepeng Xin, Deyu Meng, and Zhi Wang. Dynamicearth: How far are we from open-vocabulary change detection? *arXiv preprint arXiv:2501.12931*, 2025. 1

[32] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. In *Advances in Neural Information Processing Systems*, pages 140903–140936. Curran Associates, Inc., 2024. 1

[33] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[34] Ziqi Lu, Jianbo Ye, and John Leonard. 3dgs-cd: 3d gaussian splatting-based change detection for physical object rearrangement. *IEEE Robotics and Automation Letters*, 2025. 2, 6, 8

[35] Luigi T Luppino, Filippo M Bianchi, Gabriele Moser, and Stian N Anfinsen. Unsupervised image regression for heterogeneous change detection. *arXiv preprint arXiv:1909.05948*, 2019. 2

[36] Yunqiu Lv, Jing Zhang, Nick Barnes, and Yuchao Dai. Weakly-supervised contrastive learning for unsupervised object discovery. *IEEE Transactions on Image Processing*, 33: 2689–2702, 2024. 3

[37] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19 (3):303–342, 1993. 3

[38] Eric Martinson and Paula Lauren. Meaningful change detection in indoor environments using clip models and nerf-based image synthesis. In *2024 21st International Conference on Ubiquitous Robots (UR)*, pages 603–610. IEEE, 2024. 2

[39] 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 *ECCV*, 2020. 1, 2

[40] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. 2024. 1

[41] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *CVPR*, 2025. 3

[42] Francesco Palandra, Andrea Sanchietti, Daniele Baieri, and Emanuele Rodola. Gsedit: Efficient text-guided editing of 3d objects via gaussian splatting. *arXiv preprint arXiv:2403.05154*, 2024. 3

[43] Emanuele Palazzolo and Cyrill Stachniss. Change detection in 3d models based on camera images. In *9th Workshop on Planning, Perception and Navigation for Intelligent Vehicles at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, page 14, 2017. 2

[44] Emanuele Palazzolo and Cyrill Stachniss. Fast image-based geometric change detection given a 3d model. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6308–6315. IEEE, 2018. 2

[45] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 3

[46] Ken Sakurada, Mikiya Shibuya, and Wang Weimin. Weakly supervised silhouette-based semantic scene change detection. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 2

[47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[48] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[49] Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3d: Consistent segmentation lifting via gaussian instance tracing. In *ICCV*, 2025. 3

[50] Aparna Taneja, Luca Ballan, and Marc Pollefeys. Image based detection of geometric changes in urban environments. In *2011 international conference on computer vision*, pages 2336–2343. IEEE, 2011. 2

[51] Aparna Taneja, Luca Ballan, and Marc Pollefeys. City-scale change detection in cadastral 3d models using images. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 113–120, 2013. 2

[52] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *arXiv preprint arXiv:2112.05139*, 2021. 3

[53] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 6

[54] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023. 2

[55] 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. *arXiv preprint arXiv:2106.10689*, 2021. 1

[56] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *European Conference on Computer Vision*, pages 55–71. Springer, 2024. 3

[57] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv:2312.00863*, 2023. 3

[58] Ziyang Yan, Lei Li, Yihua Shao, Siyu Chen, Zongkai Wu, Jenq-Neng Hwang, Hao Zhao, and Fabio Remondino. 3dsceneeditor: Controllable 3d scene editing with gaussian splatting. *arXiv preprint arXiv:2412.01583*, 2024. 3

[59] 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

[60] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *ECCV*, 2024. 3

[61] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. *International Conference on Computer Vision (ICCV)*, 2025. 1

[62] Jingyu Zhuang, Di Kang, Yan-Pei Cao, Guanbin Li, Liang Lin, and Ying Shan. Tip-editor: An accurate 3d editor following both text-prompts and image-prompts. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 3