REMOVE360: BENCHMARKING RESIDUALS AFTER OBJECT REMOVAL IN 3D GAUSSIAN SPLATTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding what semantic information persists after object removal is critical for privacy-preserving 3D reconstruction and editable scene representations. In this work, we introduce a novel benchmark and evaluation framework to measure semantic residuals—the unintended semantic traces left behind—after object removal in 3D Gaussian Splatting. We conduct experiments across a diverse set of indoor and outdoor scenes, showing that current methods can preserve semantic information despite the absence of visual geometry. We also release Remove360, a dataset of pre/post-removal RGB images and object-level masks captured in real-world environments. While prior datasets have focused on isolated object instances, Remove360 covers a broader and more complex range of indoor and outdoor scenes, enabling evaluation of object removal in the context of full-scene representations. Given ground truth images of a scene before and after object removal, we assess whether we can truly eliminate semantic presence, and if downstream models can still infer what was removed. Our findings reveal critical limitations in current 3D object removal techniques and underscore the need for more robust solutions capable of handling real-world complexity. Dataset is available at https://huggingface.co/datasets/simkoc/Remove360.

1 INTRODUCTION

Trainable scene representations, such as neural radiance fields (NeRFs) (Mildenhall et al., 2020; Barron et al., 2022; Reiser et al., 2021; Müller et al., 2022; Chen et al., 2024a; Kulhanek & Sattler, 2023; Martin-Brualla et al., 2021) or 3D Gaussians (Kerbl et al., 2023; Lin et al., 2024; Yu et al., 2024; Zhang et al., 2024; Kulhanek et al., 2024; Chen et al., 2024c; Wang et al., 2024) enable photorealistic 3D reconstructions from images, and can be easily enriched with semantic features (Kerr et al., 2023; Shi et al., 2024; Ye et al., 2024a; Wu et al., 2024a; Zhou et al., 2024; Hu et al., 2024; Jain et al., 2024). This allows intuitive search via natural language prompts (Peng et al., 2023; Huang et al., 2024; Takmaz et al., 2025; Liang et al., 2024; Koch et al., 2024), e.g., asking 'find the remote control'. Similarly, it enables intuitive editing (Ye et al., 2024a; Zhou et al., 2024; Chen et al., 2024b; Gu et al., 2024; Choi et al., 2024a), e.g., by asking to 'remove the red armchair in the living room'.

The growing availability of learning-based 3D reconstruction software accessible to non-expert users (sca; Tancik et al., 2023; Yu et al., 2022; RealityCapture2023, 2023; pol; Ye et al., 2024b; pos; lum), coupled with intuitive edit operations based on natural language (Radford et al., 2021; Achiam et al., 2023; Schuhmann et al., 2022), opens up exciting possibilities: Using data casually captured by a smart phone (sca; pol; rea; lum), a user can easily create and edit photorealistic 3D models. At the same time, this raises privacy concerns: users may wish to remove private objects, such as photos, documents, or decorations, before sharing reconstructions online[1]. A key question is whether current editing methods truly remove objects, or whether they leave semantic residuals from which one can still infer what was removed.

This paper is dedicated to this privacy aspect of (mask and language-based) editing of trainable scene representations. We focus on removing objects from scenes and investigating whether state-of-the-art

[1]As an example, IKEA provides an app that allows users to scan rooms. The captured data is uploaded to IKEA's servers. Hence, IKEA recommends to physically remove private parts before scanning. In contrast, the systems envisioned in this paper allow to perform the removal virtually after the scan, which is more practical.

Figure 1: **Detecting (semantic) traces left behind after removing an object from a 3DGS reconstruction.** When there remain residuals of the object, and they can be reasoned over, the object removal is imperfect. We measure the presence of residuals with off-the-shelf semantic models and with depth. Top-Bottom: 3DGS Kerbl et al. (2023) scene before and after table removal. Left-Right: RGB renderings, SAM Kirillov et al. (2023) masks overlay with pseudo-ground-truth object outline, GroundedSAM Kirillov et al. (2023); Liu et al. (2023); Ren et al. (2024) overlay, depth renderings.

removal methods leave residuals that enable us to reason about what content was removed. Contrary to previous works (Cen et al., 2023a; Choi et al., 2024b; Mirzaei et al., 2023) that focus on foreground / background segmentation, we are interested in whether the residuals of the objects remain in the scene after removal and whether the residuals can be reasoned over (see Fig. 1). To the best of our knowledge, we are the first to consider this aspect of trainable scene representations. We introduce an evaluation framework to quantify how well current state-of-the-art methods remove objects from scenes. Our evaluation combines four complementary metrics based on semantics (equation 1), segmentation (equation 2, equation 3) and depth (equation 4), capturing whether removed objects remain detectable at different levels of granularity. Experiments on indoor and outdoor scenes show that the proposed metrics are consistent in their ranking of the evaluated methods.

To complement our evaluation, we introduce Remove360, a new dataset, a dataset of diverse indoor and outdoor scenes with real pre-/post-removal captures, and corresponding object masks serving as ground truth. Unlike existing datasets such as 360-USID Wu et al. (2025), centered on staged single-object removals, Remove360 features multi-object, naturalistic scenes that better reflect real-world complexity and expose challenging residual artifacts. Initial experiments show that current state-of-the-art 3D removal methods often fail to generalize to Remove360, underscoring the open challenges and the relevance of this benchmark for future research.

In summary, our contributions are as follows: i) We propose an evaluation that measures how well scene removal operations remove objects in the context of privacy. To the best of our knowledge, this is the first work to explore this aspect; ii) We define quantitative metrics that support this evaluation and demonstrate their consistency and reliability across state-of-the-art methods for trainable 3D scenes; iii) We release a new dataset of real-world indoor and outdoor scenes with pre-/post-removal images, and masks of the removed objects. The dataset reveals failure cases in state-of-the-art methods—such as residual artifacts, incomplete removal, and over-smoothing—that are not exposed in existing benchmarks. These challenges make it a valuable resource for advancing research on robust, privacy-aware scene editing.

2 Related Work

3D reconstruction from images builds 3D models that capture the scene's geometry and appearance. Popular scene representations are point clouds (Schonberger & Frahm, 2016; Yunhan Yang & Liu, 2023; Huang et al., 2024; Peng et al., 2023; Liu et al., 2024a; Yin et al., 2024), meshes (Schönberger et al., 2016; Furukawa & Ponce, 2009; Lazebnik et al., 2001; Kundu et al., 2020), and, more recently, Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020; Barron et al., 2022; Reiser et al., 2021; Müller et al., 2022; Chen et al., 2024a; Kulhanek & Sattler, 2023; Martin-Brualla et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023; Lin et al., 2024; Yu et al., 2024; Zhang et al., 2024; Kulhanek et al., 2024; Chen et al., 2024c; Wang et al., 2024).

NeRFs represent the scene implicitly with a colored volumetric field: for each 3D point in the space, an MLP outputs a volumetric density and a view-dependent color value. In 3DGS (Kerbl et al., 2023),

the scene is represented explicitly with a set of 3D Gaussians with learnable parameters (positions, orientation, scale, opacity, view-dependent color), generating a rendering similar to the original view.

Linking 3D reconstructions and semantics. NeRFs and 3DGS can easily be extended to embed semantic features that can be prompted via natural language Kerr et al. (2023); Tschernezki et al. (2022), pixel locations Kerr et al. (2023); Tschernezki et al. (2022) or semantic labels Ye et al. (2024a). The prompt allows humans to search for elements in the 3D reconstruction and then edit that location.

In NeRFs, the MLP is extended to output semantic features that are rendered in a differential manner and supervised with off-the-shelf 2D semantic features. Examples are NeRFs extended with text features Wang et al. (2022); Mirzaei et al. (2022) like CLIP (Radford et al., 2021), text and semantic features Kerr et al. (2023); Kobayashi et al. (2022), and unsupervised features Tschernezki et al. (2022) like DINO (Caron et al., 2021; Oquab et al., 2024). Prompting in reconstruction means locating the NeRF features most similar to the prompt. The feature field can also be trained to render segmentation masks consistent with 2D masks Zhi et al. (2021); Vora* et al. (2022); Cen et al. (2023b); Liu et al. (2024b) derived with semantic models Chen et al. (2017); Graham & Van der Maaten (2017); Li et al. (2022), panoptic models Siddiqui et al. (2023); Bhalgat et al. (2024), foundation models Ravi et al. (2024); Kirillov et al. (2023); Zou et al. (2023); Li et al. (2024); Liu et al. (2022), or user annotations Mirzaei et al. (2023). The prompt search then looks for the features associated with a given semantic label.

Although these methods perform well in locating prompted elements, the implicit nature of NeRFs makes it complex to associate their location with a set of abstract MLP parameters to edit. Hence, editing the reconstruction is not as straightforward as in 3DGS (Kerbl et al., 2023) representations that are explicit: removing a prompted element amounts to deleting the 3D Gaussians at its location. 3DGS can also be embedded with semantic features Cen et al. (2023a); Choi et al. (2024b); Jain et al. (2024); Wu et al. (2024b); Gu et al. (2024); Zhou et al. (2024); Wu et al. (2025), text features Shi et al. (2024); Liao et al. (2024); Qin et al. (2024), and unsupervised features Zuo et al. (2024). This motivates this paper to focus on 3D reconstructions represented with 3DGS.

Evaluating 3D reconstruction approaches. 3D reconstructions are evaluated based on the accuracy of the 3D geometry and whether the color renderings and the semantic (features) renderings are similar to those in the training views. Often, scene operations, such as editing and removal, are reported only as illustrative examples with qualitative results Ye et al. (2024a); Zhou et al. (2024); Gu et al. (2024); Wu et al. (2024b); Jain et al. (2024); Choi et al. (2024b). Some works Cen et al. (2023a); Choi et al. (2024b); Mirzaei et al. (2023) report a quantitative evaluation of 'background' removal by comparing the renderings of the searched element against its appearance in the original view. This paper addresses a different problem - are there residuals of the removed object present in the scene, if so, can they still be recognized or reasoned over? We propose an evaluation framework to answer this question in a quantitative and automatic manner. To the best of our knowledge, there is no previous work that addresses such a question.

Privacy challenges. The use of extensive public data in the recent scientific breakthroughs Rombach et al. (2022); Saharia et al. (2022); Brooks et al. (2024); Achiam et al. (2023) has drawn attention to the protection of user data in the research community Raina et al. (2023); Speciale et al. (2019); Pittaluga et al. (2019); Chelani et al. (2023); Moon et al. (2024); Nasr et al. (2023), in companies Rubinstein & Good (2013); Grynbaum & Mac (2023), and governments Illman & Temple (2019); Voigt & Von dem Bussche (2017).

This challenge will grow as the deployment in households of new types of sensors, eg. Augmented Reality / Virtual Reality (AR / VR) glasses (spe; Engel et al., 2023; xre), and autonomous systems will become the norm. An efficient way to make systems privacy-preserving is to consume data that has already been made privacy-preserving by the user. One relevant line of work proposes anonymizing the images Liu et al. (2024c); Weder et al. (2023) with inpainting before the scene reconstruction rather than editing the reconstruction later.

However, this method is more computationally complex: it involves editing many images as opposed to a single reconstruction and can introduce artifacts in the image that reduce the quality of the reconstruction. Hence, operating on the 3D model offers privacy at a reasonable computational cost and better reconstruction quality.

3 METRICS DEFINITION

We evaluate whether object removal in 3DGS (Kerbl et al., 2023) leaves identifiable traces of the removed content, with a focus on privacy when sharing scene representations. In this context, an element is private if it cannot be identified (Illman & Temple, 2019; Voigt & Von dem Bussche, 2017; Raina et al., 2023). Our metrics assess whether removed elements remain identifiable, assuming access to the ground-truth mask of the target object and focusing on changes within this region.

3.1 SEMANTIC OBJECT RECOGNITION

Semantic segmentation identifies objects by classifying each pixel into categories (Chen et al., 2017; Graham & Van der Maaten, 2017; Li et al., 2022). Here, it is used to test whether an object can still be recognized after removal by rendering the scene from multiple views and evaluating a segmentation model on those renderings (see Fig. 2). Comparing the segmentation performance before and after removal provides information on the removal quality. A drop in performance indicates that the object is removed. We thus define the semantic recognition metric as the segmentation's performance gap on the renderings before and after removal. The semantic segmentation is evaluated with the standard Intersection over Union (IoU) Chen et al. (2017); Badrinarayanan et al. (2017) that measures how well the estimated semantic mask overlaps with the ground-truth mask.

Specifically, IoU_{pre} and IoU_{post} are computed on the predicted semantic masks from renderings before and after removal. To reduce false positives, we only keep predictions overlapping with the ground-truth mask. The IoU_{drop} is defined as:

$$\text{IoU}_{\text{drop}} = \text{IoU}_{\text{pre}} - \text{IoU}_{\text{post}}, \tag{1}$$

ranging from -1 to 1, with higher values (↑) indicating better removal.

A low absolute value of the IoU_{drop} implies $\text{IoU}_{\text{post}} = \text{IoU}_{\text{pre}}$, which can be interpreted in two ways. (1) Both $\text{IoU}_{\text{post,pre}}$ are high, so the object is recognized even after removal (failure). (2) Both $\text{IoU}_{\text{post,pre}}$ are low, meaning the model could not segment the object even in the original scene. No conclusion about removal quality can be drawn.

To handle this ambiguity, we add a complementary semantic metric defined in the next section. Still, IoU_{drop} is useful on its own as a warning signal, especially in interactive systems where human oversight is possible, e.g., active labeling. In the experiments, we also report a more intuitive metric, the performance of the segmentation after removal, and analyse its correlation IoU_{drop}. We define the accuracy acc_{seg} as the ratio of images after removal in which the semantic element is not recognized anymore. The element is not recognized if IoU_{post} is smaller than a given threshold ξ_{IoU}, and we report this metric over multiple thresholds (see Supp. Tab. 6b, 9).

acc_{seg} ranges from 0 to 1 and the higher acc_{seg}, the better the object removal, as indicated by the ↑.

$$\text{acc}_{\text{seg},\xi_{\text{IoU}}} = \frac{\|\#\text{ images with } \text{IoU}_{\text{post}} < \xi_{\text{IoU}}\|}{\|\#\text{ images }\|} \tag{2}$$

3.2 ANYTHING RECOGNITION

The previous sections defined the IoU_{drop} that can be interpreted in two ways when it is low. To address such ambiguity, we complement the IoU_{drop} with a second semantic metric based on finer segmentations, i.e., object parts or instances instead of semantic categories. These segmentations are derived with the foundational SegmentAnything (SAM) Kirillov et al. (2023); Ravi et al. (2024), a prompt-based segmentation model that can be prompted with image locations, bounding boxes, or texts. The model can also operate without prompts, which results in a set of masks that cover all the semantic elements in the image (see Fig. 3).

In this evaluation, we assess whether an element has been removed by comparing the SAM Kirillov et al. (2023); Ravi et al. (2024) masks of the scene renderings before and after removal (in case we do not have ground truth after removal), and the scene renderings with ground truth after removal. When the object is removed or partially removed, SAM segments what is behind the object so the masks should change (see rows in Fig. 3).

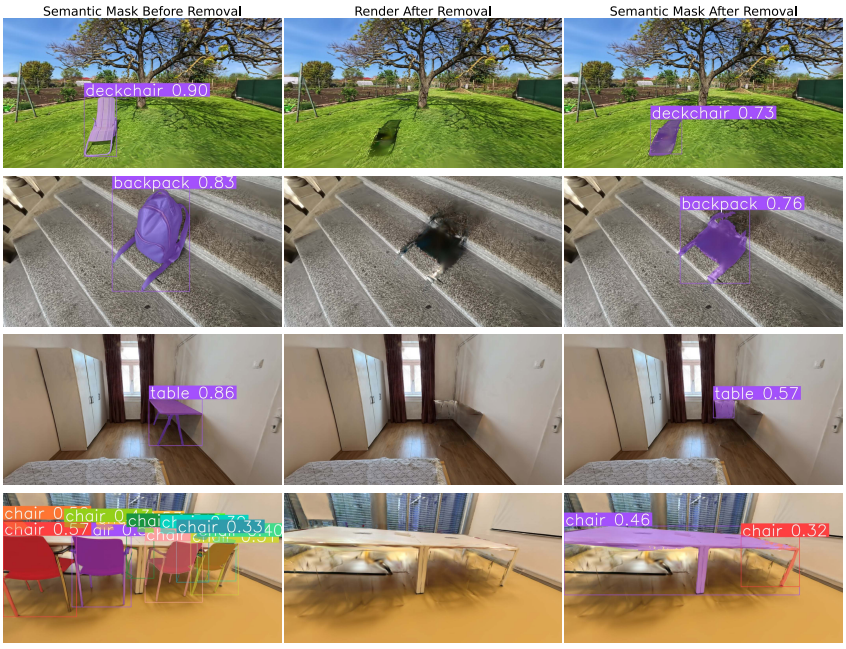

Figure 2: **Semantic segmentation changes before and after removal on Remove360.** Left-right: GroundedSAM2 Kirillov et al. (2023); Liu et al. (2023); Ren et al. (2024) overlay on the rendering before removal, rendering after removal, overlay after removal. These semantic masks are used to calculate change in semantic segmentation in equation 1 and its accuracy equation 2. Rows: Different object removals. Even though the object can not be recognized by a human, the segmentation model still finds it. One explanation can be that the pixel distribution on the edited area still exhibits patterns characteristic of the object, similar to what occurs in adversarial attacks.

Figure 3: **SAM Kirillov et al. (2023); Ravi et al. (2024) mask comparison on Remove360.** Object removal alters SAM masks, and smaller changes relative to ground-truth masks indicate better removal. These differences are used to compute the similarity score equation 3. Left to right: SAM overlay before removal, after removal, and ground-truth with the object mask (green outline).

We next define sim_{SAM} that measures the similarity between two sets of SAM Kirillov et al. (2023); Ravi et al. (2024) masks based on how well they spatially overlap. We first match the masks that overlap the most between the two sets. Then sim_{SAM} is the average overlap between the mask matches. We enforce a 1-to-1 matching, i.e., a mask in one set is matched to at most one mask in the other. We do so by defining that two masks match if they overlay, and if one mask gets matched to more than one, we keep the match that leads to the maximum overlay over all matches. This is derived by solving an assignment problem that maximizes the overlay over all matches with the Hungarian algorithm Munkres (1957).

More formally, let $A = (a_i)_{i\in[1,N]}$ and $B = (b_j)_{j\in[1,M]}$ be the sets of comparing SAM masks,and $(a_k, b_k)_{k=1,K}$ be the K matching masks. The similarity between these two sets is:

$$\text{sim}_{\text{SAM}} = \frac{\sum_{k=1}^{K} \text{IoU}(a_k, b_k)}{\max(N, M)} \tag{3}$$

sim_{SAM} lies in $[0, 1]$. Based on set of comparing mask, we aim for higher sim_{SAM} score, expecting masks to be more similar, or lower score, expecting masks to be less similar. Having ground truth after removal, comparison between after removal mask with the ground truth mask should yield high

Figure 4: **Depth changes before and after removal.** Left to right: rendered depth before removal, rendered depth after removal, thresholded depth difference and ground-truth outline of the object to be removed in green. This depth difference is used for evaluation in equation 4. Results on Remove 360 showing localized changes in the depth maps, indicating not fully inaccurate object removal, somewhere over removed other under removed.

score, indicating no visual difference in the rendering, as indicated by the ↑. When comparing masks of renderings before and after removal, the score should be be lower, the less similar the masks are, hence the better the removal, as indicated by the ↓.

Note that we normalize the score with the highest number of masks $\max(N, M)$ instead of the number of mask matches K. We do so to account not only for the difference in overlay (in the numerator) but also for the difference in the number of masks. To reflect the changes related to the removed object, the metric is derived only over the masks that overlay with the object with an IoU of at least 0.1.

3.3 Spatial Recognition

We complete the previous metrics with one that depends only on the 3D scene before and after removal, hence increasing the robustness of the evaluation against possible errors in the segmentations.

Inspired by recent works in scene change detection (Adam et al., 2022), we measure how well an object is removed based on the changes in the rendered depths before and after removal: a strong change in the depth maps indicates a change in the scene.

Hence, if the depth of the object changes enough, then the object is well removed.

More formally, we report the ratio of the object's pixels which depth changes by more than a threshold ξ_{depth}. The threshold ξ_{depth} is derived automatically with Generalized Histogram Thresholding (Barron, 2020) on the histogram of depth differences over the whole image (see Fig. 4).

The depth maps are derived from the scene's rendering before and after removal so they have consistent scales. The defined ratio can be interpreted as the accuracy in depth change and is noted $\text{acc}_{\Delta\text{depth}}$:

$$\text{acc}_{\Delta\text{depth}} = \frac{\#\text{object pixel with depth change} > \xi_{\text{depth}}}{\#\text{object pixels}} \tag{4}$$

The object's pixel locations are specified by the object's mask in the image that we assume is available.

4 Object Removal Dataset Remove360

To facilitate the evaluation of object removal methods, we introduce Remove360, a new dataset featuring RGB images of scenes both before and after object removal, along with accurate object masks. Unlike 360-USID Wu et al. (2025), which focuses on single-object, carefully aligned captures with one reference view, Remove360 targets complex, real-world scenarios with multiple interacting objects and rich scene context. This makes it a more challenging and realistic benchmark for object removal.

Dataset Composition. Remove360 contains 11 scenes: 5 indoor and 6 outdoor (see Fig. 5). Each scene includes: (1.) Training views: RGB images and object masks before removal. (2.) Testing views: RGB images of the same scene after removal, providing ground truth for novel view synthesis and residual detection. Each scene contains between 150 and 300 training views, with a comparable number of testing views. The number of removed objects per scene ranges a single item (e.g., backpack, bicycle), to pairs (e.g., white plastic chairs), and up to several objects (e.g., pillows

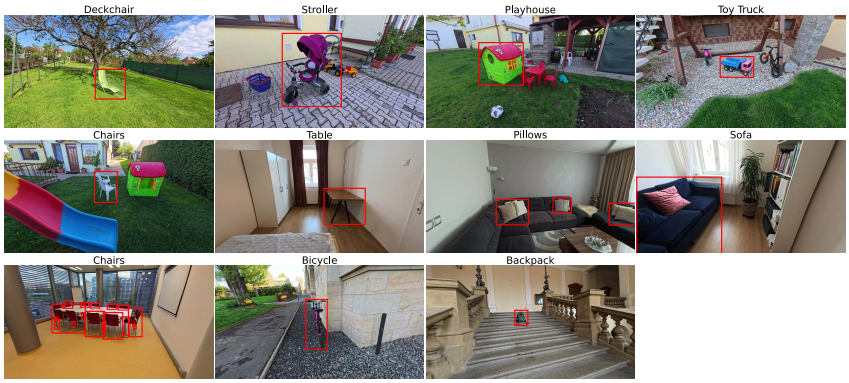

Figure 5: **Overview of the Remove360 dataset.** Samples from 11 scenes (5 indoor, 6 outdoor) with varied object counts, layouts, and interactions. Removed objects are shown with bounding boxes.

or multiple chairs in a conference room). The removed objects cover a wide range of physical characteristics, varying in size—from small (e.g., toy truck, backpack) to large (e.g., sofa, table)—and shape, from compact (e.g., backpack, playhouse) to complex shapes (e.g., deckchair, bicycle). Object masks were initially generated using SAM Kirillov et al. (2023) and subsequently refined by manual annotation. To measure the impact of inaccuracies in the ground truth masks on the results of the evaluation process, we performed a mask erosion and dilation analysis. The analysis validated that the evaluation truly focuses on removal fidelity and is robust to small boundary inaccuracies (see Supp. A.3).

Dataset Collection Protocol. We recorded 4K 60fps videos using the Insta360 AcePro camera. Each scene was captured over 4–6 minutes. We selected the sharpest frame per second using the variance of the Laplacian method. Camera poses were recovered using the hloc structure-from-motion pipeline, with SuperPoint and LightGlue for feature matching (see Supp. A.2).

5 Experiments

Methods. We evaluate five publicly available methods for object removal. To ensure fair comparison, no additional inpainting or refinement is applied after removal. We focus on Gaussian Splatting due to its explicit and interpretable 3D representation, which allows direct manipulation and evaluation of individual scene components. However, our evaluation framework is not limited to Gaussian Splatting—any 3D representation can be evaluated as long as renderings and depth maps before and after removal are available.

Feature3DGS (FGS) Zhou et al. (2024) distills the LSEG (Li et al., 2022) semantic features aligned with CLIP's text features (Radford et al., 2021). FGS is prompted with a tuple of text entries: one positive query is associated with the object of interest and the others are negative queries. The search compares the Gaussians' feature with the features of each text entry, and their similarity is normalized.

GaussianGrouping (GG) Ye et al. (2024a) distills SAM (Kirillov et al., 2023) features that operate at a finer granularity than LSEG (Li et al., 2024). Also, GG enforces spatial consistency between semantically similar Gaussians so that close-by Gaussians have similar features. A Gaussian is removed if its feature is associated with the prompted instance label. Post-processing then removes all Gaussians within the convex hull of the removed Gaussians.

SAGS Hu et al. (2024) is a training- and feature-free method that removes Gaussians based on their projection overlap with 2D object masks across views. It estimates a removal probability for each Gaussian. The 3D center of the Gaussian is projected on the images and the removal probability is the ratio of images in which the projections land on the object's location. When assigning Gaussians to the object does not account for the Gaussian's opacity, which may lead to over-removal.

GaussianCut (GC) Jain et al. (2024) leverages the spatial and color correlations between Gaussians. It models a trained 3DGS Kerbl et al. (2023) scene as a graph of Gaussians and removes them through graph-cut optimization using 2D object mask prompts, without features or training. The

Scene-	IoU_{drop} ↑					$\text{acc}_{\text{seg, IoU}_{post} < 0.5}$ ↑					$\text{acc}_{\Delta\text{depth}}$ ↑					sim_{SAM} ↑				
Object	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF
Backyard- Deckchair	*	*	*	**0.85**	0.84	*	*	*	**0.99**	**0.99**	*	*	*	**0.67**	0.65	*	*	*	**0.56**	0.54
Backyard- Chairs	*	*	*	0.85	**0.87**	*	*	*	**1.00**	**1.00**	*	*	*	**0.76**	0.67	*	*	*	**0.83**	0.62
Backyard- Stroller	*	*	*	**0.92**	0.91	*	*	*	**1.00**	**1.00**	*	*	*	**0.89**	0.73	*	*	*	**0.85**	0.72
Backyard- Playhouse	*	*	*	0.95	**0.97**	*	*	*	**1.00**	**1.00**	*	*	*	**0.92**	0.87	*	*	*	**0.50**	0.49
Backyard- Toy Truck	*	*	*	**0.95**	0.93	*	*	*	**0.99**	0.98	*	*	*	**0.73**	0.64	*	*	*	**0.22**	0.20
Bedroom- Table	*	*	*	**0.91**	**0.91**	*	*	*	0.98	**1.00**	*	*	*	0.57	**0.58**	*	*	*	**0.48**	0.44
Living Room- Pillows	*	*	*	0.62	**0.76**	*	*	*	0.77	**0.88**	*	*	*	**0.53**	0.51	*	*	*	**0.19**	0.18
Living Room- Sofa	*	*	*	0.57	**0.62**	*	*	*	0.50	**0.64**	*	*	*	**0.62**	**0.62**	*	*	*	**0.17**	0.13
Office- Chairs	*	*	*	**0.69**	0.64	*	*	*	**0.85**	0.76	*	*	*	**0.91**	0.82	*	*	*	**0.34**	0.33
Park- Bicycle	*	*	*	**0.95**	**0.95**	*	*	*	0.99	**1.00**	*	*	*	**0.91**	0.80	*	*	*	**0.68**	0.48
Stairwell- Backpack	*	*	*	**0.89**	0.82	*	*	*	**0.93**	0.85	*	*	*	**0.73**	0.65	*	*	*	**0.37**	**0.37**

(a) Remove360 dataset evaluation results.

Scene-	IoU_{drop} ↑					$\text{acc}_{\text{seg, IoU}_{post} < 0.5}$ ↑					$\text{acc}_{\Delta\text{depth}}$ ↑					sim_{SAM} ↓				
Object	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF	FGS	GG	SAGS	GC	AF
Counter-																				
Baking Tray	0.34	0.53	0.10	**0.62**	0.60	0.78	0.91	0.48	**0.99**	0.96	**0.99**	0.96	0.21	0.98	0.76	**0.21**	0.35	0.71	0.35	0.37
Plant	0.75	0.84	0.03	0.86	**0.87**	**1.00**	**1.00**	0.17	**1.00**	**1.00**	**1.00**	**1.00**	0.01	0.99	0.74	0.13	**0.12**	0.85	0.13	0.13
Gloves	0.01	0.60	0.10	0.60	**0.65**	0.28	0.84	0.34	0.83	**0.89**	0.01	**1.00**	0.55	**1.00**	0.74	0.99	**0.12**	0.56	0.16	0.17
Egg Box	0.08	**0.63**	0.56	0.62	**0.63**	0.20	**1.00**	0.96	0.99	0.99	0.06	**1.00**	0.86	**1.00**	0.39	0.84	**0.15**	0.47	0.19	0.79
Room-																				
Plant	**0.53**	0.26	0.17	**0.53**	0.23	**1.00**	0.80	0.72	**1.00**	0.96	0.97	0.70	0.33	**0.99**	0.43	0.22	0.33	0.57	0.14	**0.07**
Slippers	0.00	**0.82**	0.25	0.48	0.06	0.02	**0.83**	0.28	0.44	0.67	0.00	**1.00**	0.91	0.98	0.38	1.00	**0.05**	0.45	0.35	0.15
Coffee table	0.57	**0.86**	0.00	**0.86**	0.55	0.62	**0.99**	0.09	**0.99**	0.98	0.67	0.89	0.06	**0.99**	0.53	0.26	0.08	0.86	0.07	**0.05**
Kitchen-																				
Truck	0.62	0.61	0.67	0.66	**0.95**	0.95	0.92	**1.00**	0.99	**1.00**	0.96	**1.00**	**1.00**	0.92	0.86	0.35	0.17	0.22	**0.08**	0.19
Garden-																				
Table	0.67	0.48	0.81	0.86	**0.90**	0.70	0.54	0.88	0.95	**1.00**	0.99	**1.00**	0.98	**1.00**	0.57	0.11	0.14	**0.04**	0.06	0.10
Ball	0.00	0.16	0.41	**0.42**	**0.42**	0.94	**1.00**	**1.00**	**1.00**	**1.00**	0.00	**0.60**	**0.60**	0.53	0.47	0.59	**0.01**	0.21	0.37	0.13
Vase	0.79	0.64	0.96	**0.97**	**0.97**	0.89	0.79	**1.00**	**1.00**	**1.00**	0.99	**1.00**	0.96	**1.00**	0.92	0.12	**0.10**	0.11	0.11	0.11

(b) Mip-NERF360 Barron et al. (2022) dataset evaluation results.

Table 1: **Object removal evaluation with the proposed metrics on two datasets.** The four metrics measure changes in semantics and depth before and after removal: IoU_{drop} measures the drop in semantic segmentation after removal, $\text{acc}_{\text{seg},\xi_{\text{IoU}}}$ measures the ratio of images after removal in which the semantic element is not recognized anymore, having $\text{IoU}_{\text{post}} < 0.5$, $\text{acc}_{\Delta\text{depth}}$ captures changes in the depth maps, and sim_{SAM} quantifies difference in the SAM Kirillov et al. (2023) masks. The **best** and second-best are highlighted each metrics. (a) On the Remove360, GaussianCut (GC) Jain et al. (2024) outperforms AuraFusion (AF) Wu et al. (2025), especially in the instance segmentation similarity sim_{SAM}. (b) On the Mip-NERF360 dataset Barron et al. (2022) GC and Gaussian Groupping (GG) Ye et al. (2024a) mostly outperform Feature3DGS (FGS) Zhou et al. (2024) and SAGS Hu et al. (2024), while AF achieves the best results in semantic segmentation drop (IoU_{drop}). In the Mip-NERF360 we do not have ground truth after removal, therefore we compare instance segmentation of the renders before and after removal, expecting to see lower similarity sim_{SAM} score.

removal probability value is initialized by lifting the 2D prompt mask to 3D and refined via graph-cut optimization where the unary term represents the likelihood of the Gaussian to be removed and the binary term measures the color similarity and spatial distance between two Gaussians.

AuraFusion360 (AF) Wu et al. (2025) is a training-based method using multi-view RGB images and object masks. It removes objects using depth-aware mask generation to handle occlusions.

Datasets. Methods are evaluated on our Remove360 dataset and the Mip-NeRF360 dataset Barron et al. (2022). Since Mip-NeRF360 Barron et al. (2022) lacks semantic masks, we generate pseudo-ground-truth masks using SAM Kirillov et al. (2023) and human annotations (see Supp. A.3). We use the Mip-NeRF360 dataset as it has been commonly used by the approaches we consider in this work.

Implementation details. Due to memory constraints, we use 100–150 images per used scene from Mip-NeRF360 (kitchen, counter, room, garden), which is sufficient per prior studies (Jain et al., 2024; Zhou et al., 2024). All methods are trained on the same image subsets and 3D point clouds reconstructed using Remove360 scenes are used in full. For evaluation, we apply GroundedSAM2 Kirillov et al. (2023); Liu et al. (2023); Ren et al. (2024) to compute metrics based on semantic segmentation, and SAM Kirillov et al. (2023) for instance segmentation metrics.

5.1 Results

Methods comparison. Tab. 1 reports the considered approaches under our proposed metrics. While methods perform well overall, closer inspection and visualizations show persistent semantic residuals, indicating that current removal methods remain imperfect (see Supp. A.1).

On Remove360, GaussianCut (GC) Jain et al. (2024) outperforms AuraFusion (AF)Wu et al. (2025), particularly in instance segmentation similarity (simSAM), suggesting more accurate and complete object removal. AF benefits from training on multi-view masks, however, this characteristic did not translate to ideal performance, where removing objects using AF resulted in less similar instance segmentations after the removal, lower sim_{SAM}. IoU_{drop}, sim_{SAM}, and acc_{seg} appear to correlate, confirming that greater semantic change and segmentation similarity with ground truth novel view after removal, results in better removal. However, some scenes (e.g., living room, office) still show residual traces, and depth accuracy ($\text{acc}_{\Delta\text{depth}}$) remains low—indicating limited depth modification. Notably, several methods (Ye et al., 2024a; Zhou et al., 2024; Hu et al., 2024) fail on Remove360, unable to detect removed objects, showcasing its difficulty. This suggests that these methods might not be altering the depth information of the removed objects as effectively as they are altering the semantic and instance segmentation. On the Remove360, methods such as (Jain et al., 2024; Wu et al., 2025) are trained and evaluated on the full Remove360 dataset. In contrast, other methods (Ye et al., 2024a; Zhou et al., 2024; Hu et al., 2024) could not be evaluated on Remove360, as they failed to learn meaningful object representations and were thus unable to detect removed objects. As a result, their evaluation metrics remain identical before and after object removal, indicating failure to detect changes in Tab. 1.

For the Mip-Nerf360 Barron et al. (2022), the methods' ranks remain stable across the three metrics: AF Wu et al. (2025), GC Jain et al. (2024), and GG Ye et al. (2024a) lead across metrics. In contrast, FGS Zhou et al. (2024) under performs, likely due to prompt sensitivity, and SAGS Hu et al. (2024) shows high variance, performing better in spatially distinct objects-centric scenes (garden, kitchen). Mip-NeRF360 Barron et al. (2022) lacks post-removal ground truth, so only sim_{SAM} between pre- and post-removal renderings is available, therefore we expect to see lower similarity.

Importantly, Remove360 introduces real-world challenges with paired pre/post-removal images and masks, enabling direct measurement of semantic residuals and post-removal segmentation. Unlike MipNeRF360, it supports ground-truth-based evaluation and reveals generalization gaps, offering a more rigorous benchmark for future methods.

Qualitative Results. Figs. 2, 3, and 4 show renderings of the evaluated methods before and after removal. Fig. 2 illustrates an interesting case where the object is not visible to humans anymore, yet GroundedSAM2 Kirillov et al. (2023); Liu et al. (2023); Ren et al. (2024) finds the object. Our dataset was not available until now, and thus is not part of GroundedSAM2's training data. This suggests that barely visible information about the object can remain in the scene, even when the removal is successful to the human eye, and that the proposed metrics can detect such scenarios. This opens interesting future directions on whether a network could be trained to invert the object removal from invisible pixel information and how to prevent it.

Fig. 3 shows the distribution of SAM Kirillov et al. (2023) masks on the rendering before and after removal, compared to ground truth instance segmentation after removal. It provides a visual intuition on how sim_{SAM} behaves: successful 'sofa' removal should reveal new segments behind it. Fig. 4 provides a visual intuition for $\text{acc}_{\Delta\text{depth}}$. Successful removal causes depth differences localized in the object area (object mask outlined in green). More visualizations and quantitative results are provided in Supplementary (for Remove360 see B, and for Mip-NERF360 see C).

Limitations. The metrics rely on off-the-shelf semantic segmentation models that can introduce errors. Although introducing redundancy between the metrics alleviates this issue, it does not fully address it, which calls for further research on robust metrics for the evaluation of object removal.

6 Conclusion

We introduce a novel evaluation framework for assessing object removal in 3D Gaussian Splatting, targeting privacy-preserving scene representations. Our metrics combine off-the-shelf semantic models and depth reasoning to quantify whether removed objects leave detectable residuals. Experiments on state-of-the-art methods reveal persistent semantic traces, underscoring key limitations in current approaches. To enable rigorous, ground-truth-based evaluation, we release Remove360, a challenging real-world dataset with paired pre- and post-removal images and object masks.

We hope this work lays the foundation for future research in privacy-preserving 3D scene manipulation, where removal operations leave no recoverable trace.

REFERENCES

Luma AI Interactive Scenes. https://lumalabs.ai/interactive-scenes.

Polycam. https://poly.cam/spatial-capture.

Postshot. https://www.jawset.com/.

Realityscan - 3d scanning app. URL https://www.capturingreality.com/introducing-realityscan-2.

Scaniverse. niantic, inc. https://scaniverse.com/.

Spectacles, AR glasses powered by Snap OS. https://www.spectacles.com/.

Xreal One, AR for All. https://www.xreal.com/.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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*, pp. 108–124. Springer, 2022.

R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Jonathan T. Barron. A generalization of otsu's method and minimum error thresholding. In *European conference on computer vision*, 2020.

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*, pp. 5470–5479, 2022.

Yash Bhalgat, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *Advances in Neural Information Processing Systems*, 36, 2024.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

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 IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *arXiv preprint arXiv:2312.00860*, 2023a.

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023b.

Kunal Chelani, Torsten Sattler, Fredrik Kahl, and Zuzana Kukelova. Privacy-preserving representations are not enough: Recovering scene content from camera poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13132–13141, 2023.

Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *European Conference on Computer Vision*, pp. 338–355. Springer, 2024a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Dge: Direct gaussian 3d editing by consistent multi-view editing. In *European Conference on Computer Vision*, pp. 74–92. Springer, 2024b.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024c.

Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024a.

Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024b.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023.

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

Benjamin Graham and Laurens Van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27, 2023.

Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. In *European Conference on Computer Vision*, pp. 382–400. Springer, 2024.

Xu Hu, Yuxi Wang, Lue Fan, Junsong Fan, Junran Peng, Zhen Lei, Qing Li, and Zhaoxiang Zhang. Semantic anything in 3d gaussians. *arXiv e-prints*, pp. arXiv–2401, 2024.

Rui Huang, Songyou Peng, Ayca Takmaz, Federico Tombari, Marc Pollefeys, Shiji Song, Gao Huang, and Francis Engelmann. Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels. In *European Conference on Computer Vision*, pp. 278–295. Springer, 2024.

Erin Illman and Paul Temple. California consumer privacy act. *The Business Lawyer*, 75(1): 1637–1646, 2019.

Umangi Jain, Ashkan Mirzaei, and Igor Gilitschenski. Gaussiancut: Interactive segmentation via graph cut for 3d gaussian splatting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19729–19739, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022.

Sebastian Koch, Johanna Wald, Mirco Colosi, Narunas Vaskevicius, Pedro Hermosilla, Federico Tombari, and Timo Ropinski. Relationfield: Relate anything in radiance fields. *arXiv preprint arXiv:2412.13652*, 2024.

Jonas Kulhanek and Torsten Sattler. Tetra-nerf: Representing neural radiance fields using tetrahedra. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18458–18469, 2023.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 518–535. Springer, 2020.

Loic Landrieu and Guillaume Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4):1724–1766, 2017.

Svetlana Lazebnik, Edmond Boyer, and Jean Ponce. On computing exact visual hulls of solids bounded by smooth surfaces. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pp. I–I. IEEE, 2001.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pp. 467–484. Springer, 2024.

Siyun Liang, Sen Wang, Kunyi Li, Michael Niemeyer, Stefano Gasperini, Nassir Navab, and Federico Tombari. Supergseg: Open-vocabulary 3d segmentation with structured super-gaussians. *arXiv preprint arXiv:2412.10231*, 2024.

Guibiao Liao, Jiankun Li, Zhenyu Bao, Xiaoqing Ye, Jingdong Wang, Qing Li, and Kanglin Liu. Clip-gs: Clip-informed gaussian splatting for real-time and view-consistent 3d semantic understanding. *arXiv preprint arXiv:2404.14249*, 2024.

Jiaqi Lin, Zhihao Li, Xiao Tang, Jianzhuang Liu, Shiyong Liu, Jiayue Liu, Yangdi Lu, Xiaofei Wu, Songcen Xu, Youliang Yan, et al. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5166–5175, 2024.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.

Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models. *arXiv preprint arXiv:2402.03327*, 2024a.

Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, pp. 275–292. Springer, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Yichen Liu, Benran Hu, Chi-Keung Tang, and Yu-Wing Tai. Sanerf-hq: Segment anything for nerf in high quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3216–3226, 2024b.

Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024c.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7210–7219, 2021.

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020.

Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. Laterf: Label and text driven object radiance fields. In *European Conference on Computer Vision*, pp. 20–36. Springer, 2022.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20669–20679, 2023.

Heejoon Moon, Chunghwan Lee, and Je Hyeong Hong. Efficient privacy-preserving visual localization using 3d ray clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9773–9783, June 2024.

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.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively using public data in privacy preserving machine learning. In *International Conference on Machine Learning*, pp. 25718–25732. PMLR, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

Francesco Pittaluga, Sanjeev J Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing scenes by inverting structure from motion reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 145–154, 2019.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20051–20060, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. Egoblur: Responsible innovation in aria, 2023.

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. URL https://arxiv.org/abs/2408.00714.

RealityCapture2023. RealityCapture, 4 2023. URL https://www.capturingreality.com/.

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14335–14345, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ira S Rubinstein and Nathaniel Good. Privacy by design: A counterfactual analysis of google and facebook privacy incidents. *Berkeley Tech. LJ*, 28:1333, 2013.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. URL https://github.com/cvg/Hierarchical-Localization.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. URL https://github.com/cvg/Hierarchical-Localization.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

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

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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5343, 2024.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9043–9052, 2023.

Pablo Speciale, Johannes L Schonberger, Sing Bing Kang, Sudipta N Sinha, and Marc Pollefeys. Privacy preserving image-based localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5493–5503, 2019.

Ayca Takmaz, Alexandros Delitzas, Robert W Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3d: Hierarchical open-vocabulary 3d segmentation. *IEEE Robotics and Automation Letters*, 2025.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pp. 443–453. IEEE, 2022.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Suhani Vora*, Noha Radwan*, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *Transactions on Machine Learning Research*, 2022.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3835–3844, 2022.

Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *Advances in Neural Information Processing Systems*, 37:107326–107349, 2024.

Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16528–16538, 2023.

Chung-Ho Wu, Yang-Jung Chen, Ying-Huan Chen, Jie-Ying Lee, Bo-Hsu Ke, Chun-Wei Tuan Mu, Yi-Chuan Huang, Chin-Yang Lin, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Aurafusion360: Augmented unseen region alignment for reference-based 360deg unbounded scene inpainting. In *CVPR*, 2025.

Yanmin Wu, Jiarui Meng, LI Haijie, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024a.

Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. 2024b.

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

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024b. URL https://arxiv.org/abs/2409.06765.

Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3292–3302, 2024.

Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. Sdfstudio: A unified framework for surface reconstruction, 2022. URL https://github.com/autonomousvision/sdfstudio.

Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 43(6):1–13, 2024.

Tong He Hengshuang Zhao Yunhan Yang, Xiaoyang Wu and Xihui Liu. Sam3d: Segment anything in 3d scenes. In *ICCVW*, 2023.

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In *European Conference on Computer Vision*, pp. 341–359. Springer, 2024.

Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15838–15847, 2021.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023.

Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding. *International Journal of Computer Vision*, pp. 1–17, 2024.