Semantics-Controlled Gaussian Splatting for Outdoor Scene Reconstruction and Rendering in Virtual Reality

Hannah Schieber[1,4*] Jacob Young[2†] Tobias Langlotz[2,3‡] Stefanie Zollmann[2,3§] Daniel Roth[1¶]

Technical University of Munich
Human-Centered Computing and Extended Reality Lab
TUM University Hospital
Clinic for Orthopedics and Sports Orthopedics
Munich, Germany[1]

Department of Computer Science,
University of Otago,
Dunedin, New Zealand[2]

Department of Computer Science,
Aarhus University
Aarhus, Denmark[3]

Friedrich-Alexander-Universität
Erlangen-Nürnberg (FAU)
Erlangen, Germany[4]

Figure 1: **SCGS generates large-scale 3D assets for a variety of virtual reality applications.** Our approach enables experiencing a virtual environment captured by a continuous camera stream using a Gaussian Splatting representation. Using semantic segmentation we can replace Gaussian Splats of different classes, e.g., "water" for more interactive experiences.

Abstract

Advancements in 3D rendering like Gaussian Splatting (GS) allow novel view synthesis and real-time rendering in virtual reality (VR). However, GS-created 3D environments are often difficult to edit. For scene enhancement or to incorporate 3D assets, segmenting Gaussians by class is essential. Existing segmentation approaches are typically limited to certain types of scenes, e.g., "circular" scenes, to determine clear object boundaries. However, this method is ineffective when removing large objects in non-"circling" scenes such as large outdoor scenes.

We propose Semantics-Controlled GS (SCGS), a segmentation-driven GS approach, enabling the separation of large scene parts in uncontrolled, natural environments. SCGS allows scene editing and the extraction of scene parts for VR. Additionally, we introduce a challenging outdoor dataset, overcoming the "circling" setup. We outperform the state-of-the-art in visual quality on our dataset and in segmentation quality on the 3D-OVS dataset. We conducted an exploratory user study, comparing a 360-video, plain GS, and SCGS in VR with a fixed viewpoint. In our subsequent main study, users were allowed to move freely, evaluating plain GS and SCGS. Our main study results show that participants clearly prefer SCGS over plain GS. We overall present an innovative approach that surpasses the state-of-the-art both technically and in user experience.

Index Terms: Gaussian Splatting, Semantic Gaussian Splatting, Novel View Synthesis, Virtual Reality.

*e-mail: hannah.schieber@tum.de
†e-mail: jacob.young@otago.ac.nz
‡e-mail: tobias.langlotz@cs.au.dk
§e-mail: stefanie.zollmann@cs.au.dk
¶e-mail: daniel.roth@tum.de

1 Introduction

Allowing people to explore virtual replicas of physical environments has captivated interest for years. There are countless interesting places in the world worth capturing and exploring. Either to experience them from afar, to archive and document them, or to use them in applications for education or even games. However, high-quality experiences usually require talented 3D artists or expensive equipment such as laser scanners. Recent advances in neural rendering and radiance fields enable the creation of 3D worlds from photos alone e.g., neural radiance fields (NeRF) [33], Neural Graphics Primitives (NGP) [35], or Gaussian Splatting (GS) [20]. These approaches can create high-quality representations of objects or entire 3D scenes. GS especially reduces rendering time [20], making it particularly suitable for virtual reality (VR).

By integrating GS in VR, users can experience nearly photo-realistic environments. Novel view synthesis (NVS) enables the generation of renderings from novel viewpoints without the need to directly capture that specific part of the scene. This is particularly of interest for VR, where users want to move freely. As Gaussian Splatting (GS) builds on primitives (splats), also used in traditional rendering, they can be seamlessly integrated with 3D modeled objects in Game Engines. This integration combines the strengths of GS and Game Engines, allowing parts of the scene to be enhanced or made more interactive by replacing them with the content of the Game Engine.

For almost any editing of Gaussians and integration of 3D assets, Gaussians must be separated into different classes so that they can be individually edited, removed, or replaced. Current approaches separating Gaussians primarily focus on "circling" or forward-facing scenes [22, 60, 47, 32]. Moreover, existing datasets concentrate on NVS evaluation rather than offering pleasing VR experiences. Applying NVS to non-"circling" scenes introduces unique challenges for GS segmentation. Non-"circling" scenes enable users to be surrounded by the 3D environment instead of viewing an isolated reconstruction. However, the splats are seen by less

camera views and the scene composition results in a different challenge when removing Gaussians. GS on scenes with individual objects captured in a circular motion of the camera can use conventional classifiers and a convex hull to extract objects [60]. These extracted scene parts can be used in VR [57, 60]. When applying GS to scenes captured in a forward motion, a simple convex hull or similar envelope-based segmentation is not applicable. Using classic removal approaches an object may contain neighbour splats of other classes due to the different scene composition. Scenes recorded in forward motion only provide selected views (e.g. the front view) of parts of the scene. If object boundaries are not clear in this view, a removal over classifiers and hulls tends to contain neighboring objects. For outdoor scenes, the similarity of features in the outdoor environment (e.g., reflective water) makes segmentation increasingly more difficult than segmentation of human-created scenes.

Controlling the Gaussians, via segmentation, therefore remains challenging, as the individual class of the Gaussians is difficult to assign. In this paper, we propose a novel Semantics-Controlled GS approach (SCGS) that enables the precise segmentation of the elements of the scene. We leverage semantics-controlled Gaussians by assigning a learnable ID. This semantic-controlled IDs allow editing of the scene by removing or replacing objects with other 3D assets. Examples include replacing large scene parts, like static reconstructed water or sky, with matching (dynamic) 3D assets, facilitating a more personalized experience, see Figure 1. Allowing for the replacement of the sky, our approach can target inconsistent or unwanted weather conditions that may occur using precaptured images. Furthermore, replacing a cloudy sky with a clear blue sky allows a more appealing VR experience.

We demonstrate and evaluate our novel approach using our challenging non-"circling" outdoor dataset. Examples for the various challenges posed to NVS are small leaves or reflections in the water. Specifically, we provide a technical evaluation, showing that our approach outperforms the state-of-the-art in 3D separable GS. Our segmentation performance is on par with other semantic NVS approaches on the established 3D-OVS dataset. We also explore the advantages and disadvantages of combining our large-scale 3D asset generation technique with 3D assets from a Game Engine, where a significant and consistently dynamic element is replaced by an asset from the Game Engine. With respect to user experience, we compared video-based scene experience, plain GS, and SCGS bound to the camera capture path in an exploratory study. In our main study, we then compared plain GS and SCGS. Therein, the user was allowed to move freely and was thus able to take a closer look at the environment.

Our work makes the following contributions:

- A state-of-the-art approach for Semantic-Controlled Gaussian Splatting, namely (SCGS), surpassing existing work.
- A novel publicly available and challenging outdoor NVS dataset with semantic labels[1].
- A comprehensive technical evaluation of our approach.
- A user study evaluating the user experience and user perception of SCGS in two phases.

Overall, our work is particularly relevant when using GS to generate large-scale virtual environments beyond individual objects, such as 3D reconstructions, which can be used for cultural and environmental purposes. The scope of application ranges from historical sites or regions threatened by climate change to exploring VR as a sustainable alternative to physical tourism, enabling users to explore destinations from the comfort of their own space. Moreover, our work generally contributes to the rapidly progressing improvements of GS with potential applications extending beyond content generation for VR such as films, or games.

[1]Project Page: https://hannahhaensen.github.io/SCGS/

2 Related Work

Experiencing, creating, and exploring a virtual space can either be done classically, using video replay [4], panorama images [54, 52], single-image-based depth enhancement [3, 36], or in 3D using classic 3D reconstruction [13, 43] or radiance fields [20, 33].

2.1 Virtual Reality Scene Content

Panoramas are a widely used approach for exploring static VR content [52, 4, 64, 54]. Plain panoramas lack immersion as they miss depth information [6, 5]. Bertel et al. [6] optimize this using 3D proxy fitting. Ajisa et al. [3] propose inpainting to view an indoor or outdoor scene based on a single panorama image, thus limiting the area of movement in the scene.

Other approaches for enriching outdoor photographs [49, 31, 7, 14] do not directly address VR. Freer et al. [14] separate people in front of sightseeing attractions and utilize neural rendering to inpaint the area. Zhao et al. [7] integrate the capture of one person in sight and extrapolate it using online data.

Apart from simply replaying a scene, advances in deep learning allow the generation of neural content for VR. Campos et al. [8] utilize procedural content generation based on agents and decision trees to enable a unique user experience. Large language models (LLMs) [1, 59, 61] and other foundation models [9] have advanced content generation, enabling the creation of assets from simple text to 3D [59, 61]. While standard LLMs are challenged to create VR scenes, Yin et al. [61] propose Text2VRScene, generating synthetic non-photorealistic, but content aware VR scenes.

2.2 Novel View Synthesis

Classically, radiance field-based approaches do not directly target large-scale scene extraction for VR. Scenes created with radiance fields can be used for virtual content [12, 26, 17, 40] and multiple mixed reality (MR) devices allow to generate such virtual content [42]. However, using radiance field-based rendering in VR/extended reality (XR) challenges include scene representation and the underlying data structure. For example, changing 3D scene content may prove difficult as editing a NeRF is not trivial [10, 55, 58]. ClipNeRF [55] addresses this by adapting an existing NeRF in a separate training. Although NeRF works well for NVS, its real-time rendering capacity for VR has been outperformed by GS [20].

GS, is an explicit scene representation, allowing easier adaptation compared to the implicit NeRF representation. GS starts with a sparse point cloud. Using photometric loss as well as densification and pruning steps for refining. The initial GS representation can be challenged by large simultaneous localization and mapping (SLAM) like scenes [20, 21]. To overcome this, Kerbl et al. [21] introduce hierarchical GS, enabling a block-wise optimization depending on the camera location at rendering time.

For VR, Jian et al. propose VR-GS [17] for indoor scenes. VR-GS integrates inpainting in the GS training process, using mesh exports and manual post-processing for each object. Thus, the objects are movable in VR. Chen et al. [10] reconstruct dynamic urban scenes using Gaussian scene graphs. Each graph holds information about individual parts of the scene.

Another approach, besides separating a scene into statics and dynamics, is semantic GS. Semantic GS enables extracting parts of a scene as 3D assets. Feature 3DGS [63] utilizes Segment-Anything (SAM) embeddings to improve NVS quality. Similarly, Gaussian Grouping (GG) [60] introduces semantic features into a GS structure, proposing identity encoding allowing to group 3D Gaussians. This outperforms LangSplat [37] building upon CLIP embeddings.

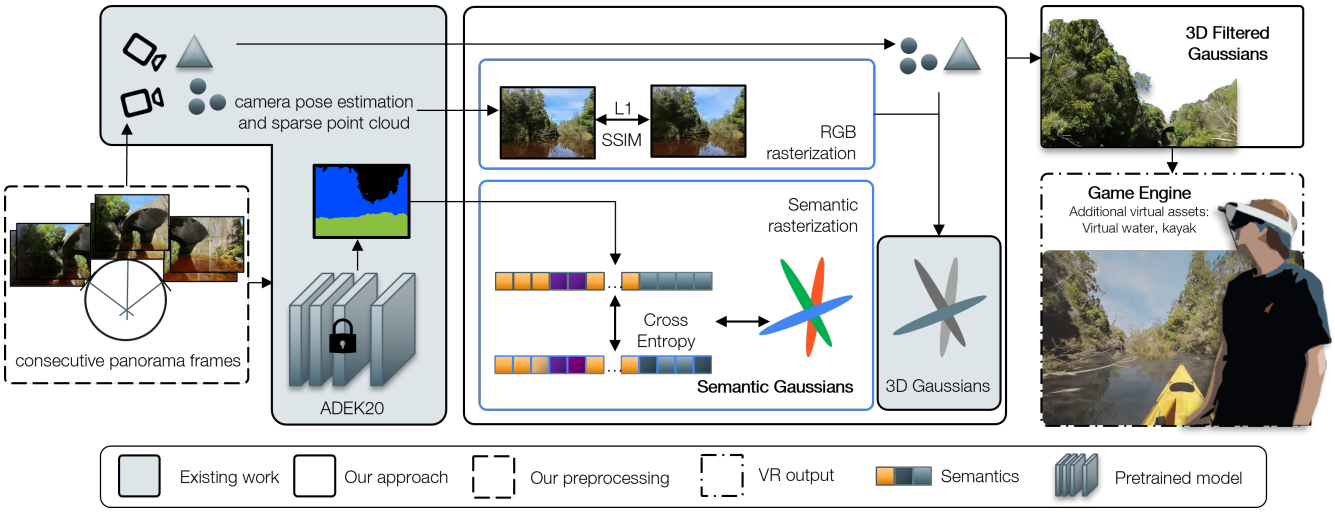

Figure 2: **Architecture of SCGS.** We first extract images from our continuous panoramic stream. Using COLMAP to estimate the camera positions, we obtain the sparse point cloud for the initialization of GS. To enable 3D filtering, the data is preprocessed with a segmentation model. During 3D Gaussian training, we use CE-loss, L1-loss and SSIM-loss to fit our scene into the RGB and segmentation space. The final 3D representation can be viewed in the viewer, or individual parts of the scene be extracted and used in VR.

Building on SAM-DEVA [11], Contrastive GG [47] extends the idea of GG by omitting the tracking step in pre-processing and identifying consistence labels through a contrastive learning step. Disadvantageously, when interacting in 3D and especially when exporting the scene content to VR not all approaches concentrate on integrability with Game Engine plugins.

2.3 User Experience of Reconstructed Environments

NVS has been explored in XR, specifically for MR [34, 41] with screen-based applications, and in VR with individual 3D reconstructed parts of a scene [23].

Sakashita et al. [41] visualize a point cloud and NeRF using a head-mounted camera and a PC for shared interactions. The PC displays the point cloud overlaid on the NeRF. In a preliminary user study, they found a preference for NeRFs combined with point cloud overlays over video or pure point cloud visualization.

The use of 3D assets for task execution planning benefits from 3D assets generated with signed distance field (SDF) based approaches [23, 27]. Kleinbeck et al. [23] create a digital twin of operating rooms in VR. Using SDF-based mesh reconstruction of the scene, an accurate mesh is created. Manual post-processing of scene parts enables a VR experience for participant exploration.

2.4 Research Gap

Recreating the real world through cameras achieves high realism [3, 52, 51], but 3D reconstruction offers greater freedom in VR. An approach is needed that processes GS for VR while providing a dataset supporting a pleasant VR experience and NVS evaluation.

SCGS addresses limitations in semantic GS for VR and NVS, as existing methods focus on "circling" camera setups [24, 60, 47, 37], relying on feature-based separation [60, 47] or language guidance [37]. These approaches work well in controlled environments with low similarity of features. However, outdoor scenes in a non-"circling" camera setup with similar features pose challenges to feature-based editing for VR.

With SCGS and our NVS dataset, we bridge this gap, enabling accurate large-scale scene editing with semantics-controlled GS in a) a forward-moving camera setup and b) complex environment with homogeneous features. Our dataset captures immersive outdoor scenes in a non-"circling" setup, supporting VR experiences. Moreover, our user experience measurements provide insights in the user perception on GS.

3 Method

Our approach, SCGS, separates Gaussians into segmentation classes, directly assigning the respective segmentation class. This enables the editing of the scenes via semantic-controlled Gaussians. To achieve this, we alter the Gaussian rasterization process. This allows the classification of 3D Gaussians in the 2D image space and 3D Gaussian space at almost equal quality, which is advantageous in non-"circling" setups. The direct class assignment of SCGS enables the removal of complete classes at a large-scale, while omitting feature similarity[2].

3.1 Semantics-Controlled Gaussian Splatting

3.1.1 Preliminary 3D Gaussians

3D GS [20] represents an explicit scene representation initialized from a (sparse) point cloud. For the Gaussian representation Σ' represents the 2D rasterized Gaussians, J is the Jacobian of the affine approximation, W is the world-to-camera transformation matrix and Σ is the 3D representation [20].

$$\Sigma' = JW\Sigma W^T J^T \tag{1}$$

Each Gaussian G is represented by its 3D center position (x) and a 3D covariance matrix (Σ) that can be denoted as a rotation matrix and scaling matrix [20]. To represent colors and scene appearance, each Gaussian holds a density value (σ) and spherical harmonics (SH) coefficient to encode RGB information. To retrieve the color (c) of each pixel, alpha (α) blending is used [20].

$$RGB = \sum_{i \in \mathbf{N}} T_i \alpha_i c_i \; with \; T_i = \prod_{j=1}^{i-1} (1 - \alpha) \tag{2}$$

For RGB, the differentiable rendering pipeline first converts SH into RGB values, which are then splatted onto the 2D image plane.

[2]Project Page: https://hannahhaensen.github.io/SCGS/

Figure 3: **Images of our dataset.** Tree, Open Sea, Picnic, Outback, and Kayak (from left to right).

3.1.2 3D Gaussian Segmentation

We enhance the GS representation by integrating semantic information, extending the conventional RGB rasterization process to support semantic rendering, see Figure 2. This adaptation allows us to reformulate the segmentation problem within the Gaussian parameter space, facilitating the direct assignment of class IDs to each Gaussian in 3D space. We introduce a new learnable parameter similar to SH. This results in a segmentation ID per Gaussian which can be expressed as class IDs using *argmax* activation. i denotes the number of Gaussians and c the number classes:

$$ID = [s_{c_1}, ...s_{c_i}]^T \tag{3}$$

For the semantic map (s), we adapt the parallel rasterization process [60], where the SH components are set to zero, instead of rasterizing feature maps we effectively isolating the semantic attributes.

Our approach diverges from traditional classification [60, 63], which typically requires a separate classifier. Instead, we assign class IDs consistently during training, using cross entropy (CE). Our IDs derive from the image space and get assigned via differentiable rendering. This addresses feature space similarities, especially in outdoor scenes with significant reflections. The 3D segmentation is projected on a 2D map using alpha blending (α):

$$\text{SEGMENTATION} = \sum_{c \in C} T_i \alpha_i s_{c_i} \tag{4}$$

i denotes the number of Gaussians and c the number classes in the segmentation map s.

Our approach modifies the rasterization process to facilitate backpropagation of the segmentation map, similarly to how RGB values are handled. After the rasterization step, each Gaussian is associated with a class segmentation ID that is splatted onto the 2D image plane. This enables the application of CE loss to supervise the 3D GS segmentation through a 2D loss function.

The loss function for semantic segmentation is defined as

$$L_{CE} = -\sum i \sum_c s_{c_i} \log \hat{s}_{c_i} \tag{5}$$

where s_{c_i} is the ground truth segmentation for class c at pixel i, and $\hat{s}_{c_i}$ is the predicted probability for class c at pixel i.

Additionally, the assigned class ID allows for the selective removal of one or more sets of 3D Gaussians at a large-scale, enabling targeted modifications of the 3D scene.

3.1.3 3D Gaussian Separation

We utilize our direct ID segmentation class assignment to remove/seperate 3D Gaussians from the complete set of 3D Gaussians ($\mathscr{G}_{new}$). Given the desired object class or object classes to remove ($c_{r=1..n}$), our approach enables removing one or more classes per scene. Since each Gaussian class has a direct identifier, no additional post-processing, as e.g. creating a convex hull [60], is required. Moreover, our approach allows to remove directly large-scale object using the semantic-controlled Gaussians, see Figure 2.

The assigned segmentation class ID also allows for the selective removal or modification of one or more sets of 3D Gaussians.

$$\mathscr{G}_{new}(x;\Sigma) = \mathscr{G}(x;\Sigma) \cdot \mathbf{I}(s_{ic} \neq c_{r=1..n}) \tag{6}$$

where $\mathbf{I}$ is the indicator function, ensuring only the Gaussians not belonging to the removed class ID are retained.

3.2 Large-Scale Outdoor 3D Asset Dataset

Existing semantic NVS datasets focus on indoor scenes [28, 60, 22] following a circling camera path. Our dataset provides challenging outdoor scenes containing reflective surfaces, similar features and challenging structures (trees, leaves, water). It is captured using Insta360 cameras (X1, X2 and X3). The camera is positioned in front of individuals engaged in outdoor activities, like kayaking. Employing a panoramic setup, we derive multiple camera poses from the resulting forward moving video stream. Example images can be seen in Figure 3. By combining forward-facing images with those angled $\pm 60/\pm 30^\circ$ to the left and right and $\pm 10^\circ$ up and down, we achieve comprehensive coverage of the scene. For privacy, the setup excludes the individual experiencing the activity. After extracting images from the video stream we retrieve segmentation masks using DPT [38]. Our outdoor recordings feature known classes. Therefore, we use the ADE20K labels [38]. Afterwards, the camera poses of the image set are retrieved [45].

We divide our dataset into pure NVS (images at the angle $\pm 60^\circ/\pm 30^\circ$) and a set where we add the complete 360° video. This allows to compare 360° videos and NVS in VR. We create a stacked video for the 360° images using 360° monodepth [39, 3].

4 Technical Evaluation

SCGS can extract Gaussians from large-scale scene via semantics-controlled IDs. Since SCGS aims to separate the 3D Gaussian's, we compare it on our dataset with identity encoding, namely GG [60]. On 3D-OVS, we assess segmentation quality and compare it with other state-of-the-art NVS segmentation approaches building upon language supervision [37, 22] and contrastive learning [47].

4.1 Metrics

To compare the rendering quality of the novel views, we report peak signal-to-noise ratio (PSNR), similarity index measure (SSIM) [56] and learned perceptual image patch similarity (LPIPS) [62]. For the segmentation performance, we report mIoU.

4.2 Implementation Details

We used ffmpeg to retrieve images from the video. For camera poses and sparse reconstruction we leverage COLMAP [45].

SCGS is implemented in Python with PyTorch and CUDA, trains all scenes on a single RTX4090 with 24GB VRAM, while comparison methods use an A100 with 40GB VRAM. The loss function is: $L = 0.8 * L_1(image,gt) + 0.2 * (1 - ssim(image, gt)) + 0.2 * L_{CE}$. We train for 30k iterations with an initial position learning rate of 1.6e-4 and a final one of 1.6e-5. For opacity we use 0.05, for scaling 0.05, for rotation 0.001, and for features 0.0025.

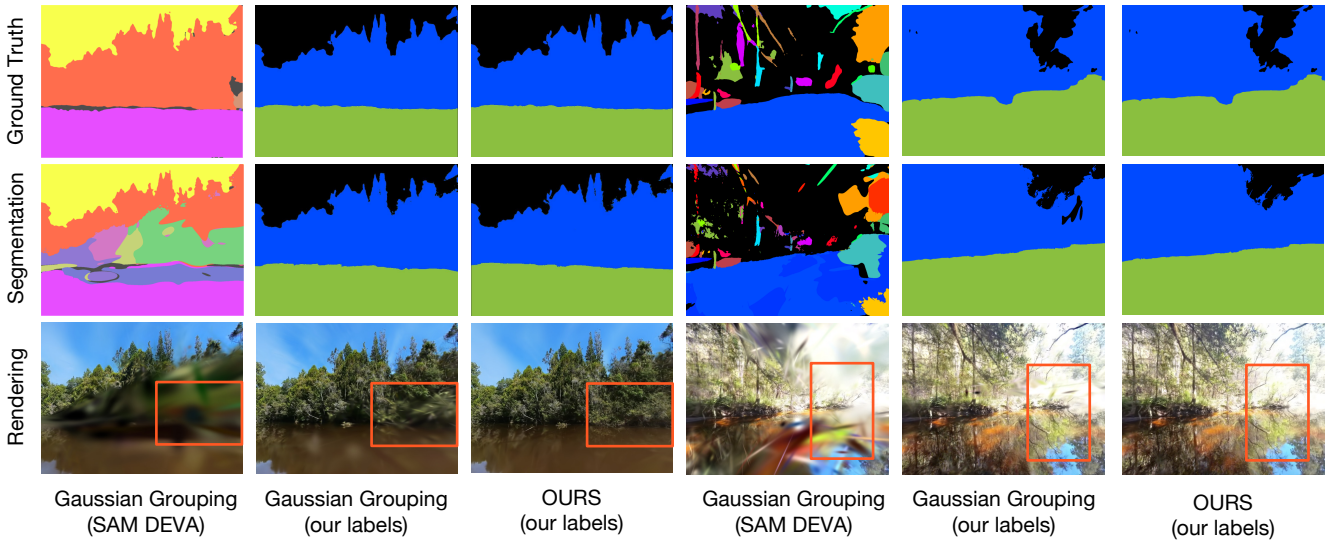

Figure 4: **Example comparison of Gaussian Grouping with SAM labels, Gaussian Grouping (with our ADE20K labels) and our approach (with our ADE20K labels).** The outback scene (right) shows the challenges of the water and the kayak scene (left) shows the challenges of the closed stacked trees.

Table 1: **Technical evaluation.** We report PSNR, SSIM, LPIPS and mean Intersection over Union (mIoU) (only when using the same labels). The best results are highlighted in **bold**. Best results for NVS within a range of $\pm$ 0.5 dB are highlighted in light-green and above 0.5 improvement in dark-green. Results worse than 1.0 compared to our approach are highlighted in orange.

Approach	Gaussian Grouping [60] SAM DEVA (original)			Gaussian Grouping *our labels*				OURS			
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	mIoU	PSNR	SSIM	LPIPS	mIoU
Tunnel	20.54	0.654	0.379	**23.29**	**0.792**	0.403	**91.90**	23.11	0.727	**0.316**	88.11
Lake	20.80	0.703	0.329	21.59	0.734	0.306	**94.23**	**21.71**	**0.735**	**0.303**	93.10
Kayak	18.61	0.576	0.448	21.51	0.662	0.406	**91.29**	**22.02**	**0.681**	**0.386**	90.14
Open Sea	27.82	0.837	0.352	27.81	0.831	0.358	91.44	**28.06**	**0.842**	**0.352**	**96.31**
Short Ride	18.67	0.635	0.374	19.51	0.678	0.336	**82.78**	**20.02**	**0.694**	**0.324**	82.60
Outback	21.18	0.700	0.408	24.80	0.781	0.320	**85.44**	**25.13**	**0.799**	**0.299**	80.91
Picnic	23.96	0.795	0.241	24.90	**0.811**	0.225	**88.53**	**24.97**	0.805	**0.215**	84.22
Tree	23.29	0.792	0.403	25.40	**0.814**	0.358	69.84	**25.83**	0.802	**0.357**	**77.60**
Mean	21.86	0.712	0.367	23.51	**0.761**	0.34	**86.93**	**23.86**	**0.761**	**0.31**	86.62

Table 2: **Convergence example on the our outdoor dataset.** Results on our outdoor dataset for NVS quality after 1K/7K iterations. Best results are highlighted in **bold**.

	1K Iterations				7K Iterations			
	Gaussian Grouping (*our labels*)		OURS		Gaussian Grouping (*our labels*)		OURS	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Tunnel	18.92	0.561	**19.98**	**0.588**	21.24	**0.657**	**21.74**	0.665
Lake	17.63	0.535	**18.77**	**0.572**	19.93	**0.641**	**20.03**	0.378
Kayak	17.90	0.496	**18.76**	**0.555**	19.94	**0.571**	**20.25**	0.455
OpenSea	**21.82**	0.747	21.78	**0.750**	25.45	0.811	**26.50**	**0.818**
Short Ride	**17.85**	0.440	17.83	**0.488**	19.17	0.603	**19.48**	**0.642**
Outback	19.49	0.602	**19.70**	**0.613**	**23.18**	**0.715**	23.01	0.706
Picnic	21.49	0.494	**22.39**	**0.691**	23.62	0.764	**23.81**	**0.766**
Tree	20.53	0.601	**22.14**	**0.731**	24.61	0.705	**25.11**	**0.777**
Mean	19.70	0.590	**21.90**	**0.663**	20.17	0.636	**22.49**	0.651

4.3 Novel View Synthesis Quality

Our ID-based training improves convergence and enhances NVS quality, see Figure 4, Table 2 and Table 1. In our scenario, continuous labels are available a priori, allowing us to conduct a direct and fair comparison with GG [60]. As highlighted in Table 1, we outperform GG with their initial SAM DEVA [11] labels on all scenes in NVS quality. Comparing the segmentation maps of GG with ours reveals a clear quality difference, see Figure 4. As shown in Figure 4 continuous labels lead to better results for our use case.

Consequently, for a fairer comparison, we updated the segmentation maps for the GG comparison with *our labels* denotes as GG. We retrained GG using *our labels*. In Table 1, Figure 4 and Figure 6, the improved labels strongly enhance the NVS performance. Nevertheless, SCGS outperforms GG on all scenes and GG using *our labels* on seven out of eight scenes on our outdoor dataset. Moreover, we outperform it on five scenes in SSIM and on all eight scenes in LPIPS. The NVS quality improvement is also visible when comparing the images in Figure 4. Additionally, our approach outperforms GG using *our labels* early in the training process, see Table 2, potentially due to a lower training load[3].

4.4 Segmentation Performance

We distinguish segmentation performance into object-removal impact and classic segmentation performance measured by mIoU.

[3]SCGS renderings available at: https://osf.io/s9uvy/

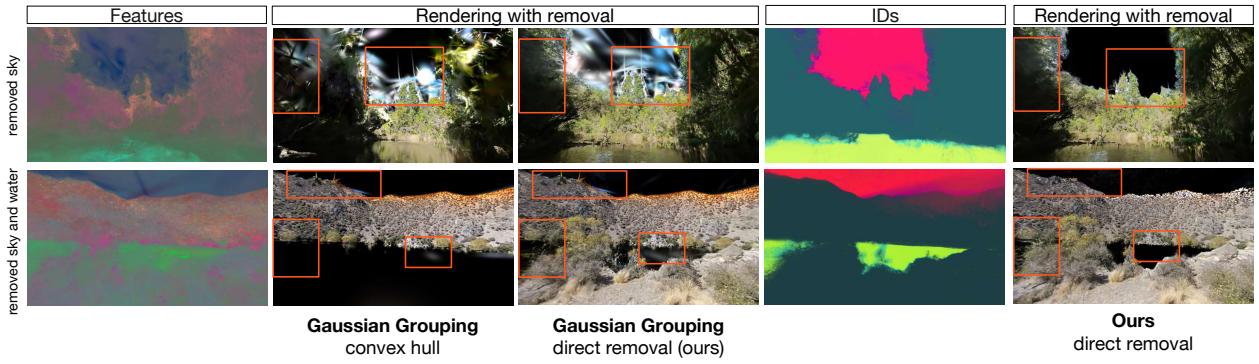

Figure 5: **Class removal on our dataset using Gaussian Grouping with *our labels*.** The convex hull removes too much of the scene (left), while direct removal with Gaussian Grouping (center) introduces more outliers compared to SCGS (right).

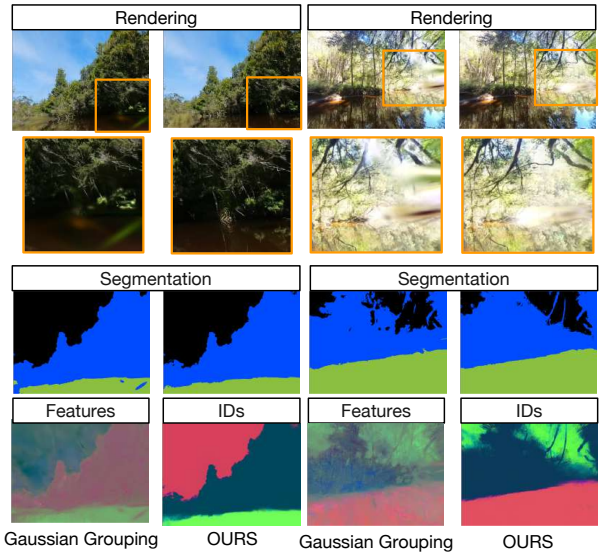

Figure 6: **Example comparison of Gaussian Grouping (with our ADE20K labels) and our approach (with our ADE20K labels).** We compare the feature space which is later classified for the segmentation and our ID-based assignment.

4.4.1 Large-Scale Object Removal

The benefit of SCGS becomes even clearer through post-processing removing Gaussians. Although our mIoU is not higher compared to GG using *our labels*, our semantics-controlled removal can better remove objects from the 3D space. As shown in Figure 6, our ID-based semantics-controlled object removal shows clearer boundaries between the objects compared to GG. While GG, which is feature-based has the tendency to assign the same features (color in Figure 6) to the water as the surrounding, their the 3D segmentation quality, lacks under these similarities as GG uses the classifier trained on 2D feature maps to remove objects from the 3D space. Therefore, the removal of objects appears more challenging. Through our ID-based assignment, the splats remain their ID consistently during training, and feature similiarty as no impact on them. Even when using *our labels*, the achieved performance in object removal is not on par with our approach using ID-based semantics-controlled object removal, see Figure 5. We even tested our direct removal in Figure 5 for the baseline using *our labels*. Still, our approach shows a better result. As shown in Figure 5, occlusion of Gaussians leads to jittering at the borders between classes. Nevertheless, an almost precise removal is possible.

SCGS can remove individual classes and shows noticeably clearer and better boundaries to other objects/classes in the outdoor scenes. This leads to a higher-quality scene which can be integrated into Game Engines, see Figure 7. As can be seen in the top line in Figure 5, the sky and the tree are too connected when using a convex hull. Since we do not use a classic circular capturing setup, a convex hull may not be the best way to remove unwanted objects. Therefore, we propose direct removal by class. As shown in the comparison in Figure 5, our approach better distinguishes the individual classes and removes large-area parts directly and accurately also when using the same direct removal technique.

Table 3: **Evaluation on the 3D-OVS dataset [28].** Following [47], we report mIoU per scene and overall.

Approach	Bed	Bench	Room	Sofa	Lawn	**Mean**
LERF [22]	73.5	53.2	46.6	27.0	73.7	54.8
GG [60]	**97.3**	73.7	79.0	68.1	**96.5**	82.9
LangSplat [37]	34.3	84.8	56.3	67.7	95.8	67.8
Contrastive Grouping [47]	95.2	**96.1**	**86.8**	67.5	91.8	87.5
Ours	94.4	89.8	73.2	**92.6**	89.0	**87.8**

4.4.2 Segmentation Quality on 3D-OVS

We compare SCGS with GG [60], LERF [22], LangSplat [37] and Contrastive GG [47] on 3D-OVS [28]. The classes per scene are not the same. We follow previous work [47] and use the same scenes. Following [47], we report mIoU per scene and on all scenes. As reported in Table 3, we outperform existing work in one of five scenes and perform competitively in all other scenes. The improvement in the "Sofa" scene shows that we can outperform the existing work in the overall mIoU. We report 93.04 for the *joy-con*, 91.88 for the *uno cards*, 97.71 for the *sofa*, 80.00 for *Gundam*, 96.45 for Pikachu and 97.63 for the *XBOX controller*. As shown by these insights we achieve a high mIoU per class on this scene.

4.5 Use Cases

SCGS has broad applicability in Game Engines. The primary objective, extracting large-scale scene parts via semantic-controlled Gaussians, addresses the needs of diverse VR environments. This can be particularly valuable for games or virtual experiences in fields like virtual tourism, where specific assets, such as nature, sports fields, or famous statues, need to be seamlessly integrated into virtual worlds. As shown in Figure 7, single classes (e.g., sky) or multiple classes (e.g., sky and water, or sky and buildings) can

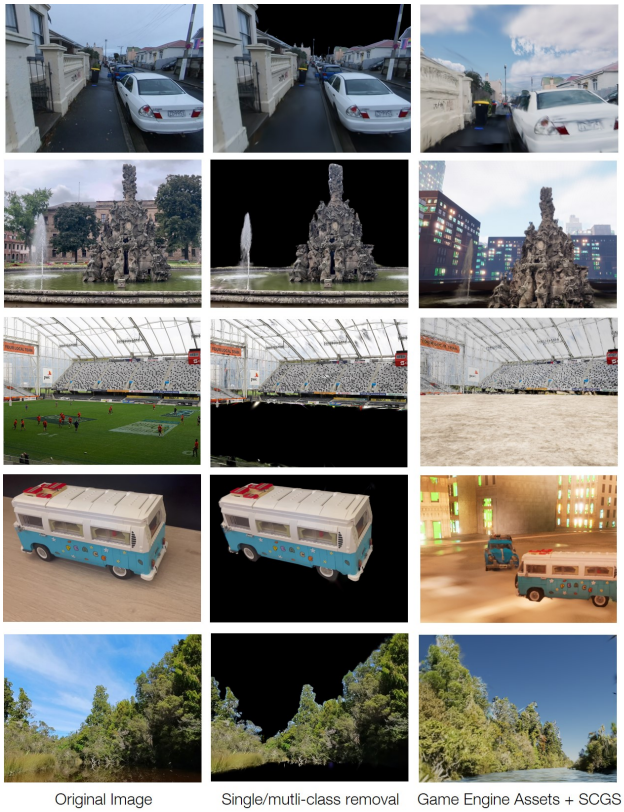

Figure 7: **Use Cases.** Our approach enables large-scale scene removal across various cases: sky replacement (first, second, last row), scenes outside our dataset like sports fields or fountains (second and third row), and smaller objects like brick cars (fourth row).

selectively be removed via semantics-controlled Gaussians. SCGS enables the incorporation of new assets from Game Engines, allowing novel viewpoints and more dynamic scene rendering.

5 User Study

To investigate user perceptions of plain GS and SCGS (SCGS combined with 3D assets) in VR, we conducted an exploratory and a main user study using a within-subject (repeated measures) design. The ethical approval of the participating institutions was granted.

5.1 Apparatus

We used an Oculus Quest 3 head-mounted display (HMD) connected via Oculus Link to a workstation powered by an NVIDIA RTX 4090. Rendering was done on the workstation in Unreal Engine using the Lumalab plugin [29] for GS and custom scene setups.

5.2 Procedure

After welcoming participants and obtaining consent, they completed a questionnaire on demographics and VR experience. Followed by familiarizing them with the HMD. Then they experienced the conditions in a randomized, balanced order, filling out a questionnaire after each one. At the end, they ranked the conditions.

5.3 Analysis Strategy

All analyses of the user studies were performed using RStudio Version 4.4.1. We evaluated the study using one-way repeated measures ANOVA where suitable (three conditions), a paired samples t-test (two conditions), and Tukey's post-hoc analysis with Bonferroni correction where suitable. Our significance level is set to 0.05. We applied the Shapiro-Wilk test to test for normal distribution.

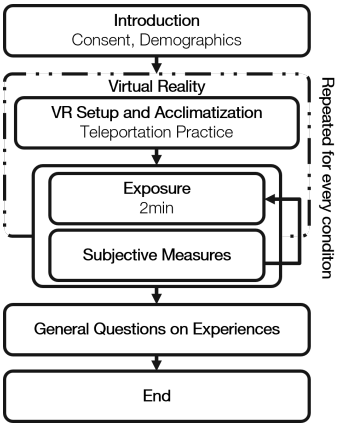

Figure 8: **User Study Procedure Diagram.**

5.4 Explorative Study

According to the literature [51], 360° panorama images/videos enhance users' sense of presence, but research on perceived realism and presence in GS is limited. In our preliminary study, we focused on these aspects using 360° RGB-D video as a baseline.

As existing work on GS provides image metrics or VR examples without specific user feedback, the perceived presence in a GS environment is so far unknown. The goal of this explorative study is to establish a frame of reference within which we will operate in our main study in which we give the user more freedom.

Conditions The original 360° video was recorded in a seated kayak scenario, which we replicated by adding a virtual kayak to both GS scenes. Our baseline was a 360° panorama video with generated depth (condition 1) [39, 3]. The other two conditions were: plain GS without dynamics (condition 2) and SCGS with added water dynamics (condition 3). Throughout the experience, the user followed the camera path at the center of a river while seated in a (virtual) kayak.

Measures To measure perceived presence in VR, we used the igroup presence questionnaire (IPQ) [44, 53]. Participant preferences were assessed through environment ratings and a final condition preference question.

Participants We recruited 24 participants (14 male, 9 female, 1 non-binary) through announcements, notice boards and word-to-mouth. The participants had an average age of 22.42±4.93).

Results and Discussion We found a significant difference in "realism" when comparing video with plain GS and video with SCGS, see Table 4. Applying Tukey's post-hoc analysis and pairwise t-tests with Bonferroni correction, we reveal a significant difference between video and GS ($p < 0.028$) and video compared to SCGS ($p < 0.016$).

SCGS ranked second for first preference and highest for second preference. The video condition ranked first, while plain GS scored lowest in user preference. We found a significant difference when comparing the video condition with both GS and SCGS. This is reasonable as the video, where users follow the original camera path, naturally looks more realistic in terms of image quality. SCGS performed similarly to plain GS, which is supported by similar median and stndard deviation (SD). Our approach ranked second in preference, following the video condition, while plain GS ranked last.

Table 4: **Results of the IPQ for the explorative study.**

IPQ	M_{video}	SD_{video}	M_{GS}	SD_{GS}	M_{SCGS}	SD_{SCGS}	F	p
General Presence	3.83	1.37	3.38	1.70	3.42	1.63	0.595	0.554
Spatial Presence	3.68	1.14	3.42	1.11	3.39	1.14	0.189	0.828
Involvement	3.35	1.32	3.27	1.24	3.26	1.29	0.009	0.991
Realism	2.86	0.92	2.05	0.87	2.02	0.95	5.145	**0.008**
Overall	3.37	0.95	2.98	0.85	2.96	0.92	0.921	0.403

Table 5: **Results of the IPQ from the main study.**

IPQ	M_{GS}	SD_{GS}	M_{SCGS}	SD_{SCGS}	$t(df)$	p
General	4.00	1.40	4.50	1.03	-3.378(29)	**0.002**
Spatial	3.50	0.94	4.00	0.82	-3.062(29)	**0.005**
Involvement	2.75	0.75	3.38	1.70	-1.586(29)	0.120
Realism	1.63	0.90	2.75	0.84	-6.755(29)	<**0.001**
Overall	2.64	0.90	3.47	0.86	-4.015(29)	**0.007**

The IPQ does not reflect all feedback, as participants expressed a desire for free movement and described the 360° video as flat and resembling 2D content. Plain GS was criticized for lacking immersion. In contrast, comments like *"The moving water in the river had a huge impact, it felt so realistic."* suggest positive feedback for SCGS, particularly regarding the added dynamic assets like flowing water and reflections. These findings highlight the benefits of SCGS and suggest exploring it further in the main study, where participants can move freely instead of sitting in a virtual kayak.

5.5 Main Study

Our main study investigates the effect of SCGS in combination with 3D assets compared to plain GS. In our preliminary study, we received feedback that self-directed movements would be appreciated ($N = 10$). Thus, the user could now move freely in the virtual world by teleportation. We explored whether SCGS enhances the presence with free user movement. Our hypotheses based on previous indications and literature [48, 50] are:

HM1: The addition of 3D assets into GS using SCGS will induce significantly higher spatial presence in users than plain GS. Given that the quality of GS in terms of accurate reflections is decreasing with varying viewpoints, we assume that the 3D assets, i.e. water, can improve realism and sense of presence, as the reflections adapt to the viewpoint of the user.

HM2: We hypothesize that SCGS is more graphically pleasing and visually coherent than plain GS. Considering, the relevance of the captured camera trajectory for GS and NVS in general, we expect a higher rate for visual coherence in SCGS, as 3D assets have the potential to enrich the consistency of the overall 3D scene.

5.5.1 Study Setup

Measures We measured presence with IPQ [44]. Participants again rated their favorite experience, commenting if wanted.

Building on the explorative study, we added questions on graphical appeal, visual coherence, presence, and environment behavior, from 1-strong disagreement to 10-strong agreement, see Table 6 for details. Inspired by Mal et al. [30], we created these questions to analyze the perceived quality of the 3D environment.

Participants We recruited 30 participants (16 male, 14 female, 0 non-binary) with no overlap to the participants from the explorative study. The participants had an average age of 26.97±3.37).

Design We used the same setup as the exploratory study, enhancing the experience with free movement within a predefined space. To ensure comparability across participants, we selected three locations marked by rocks and teleportation indicators, see Figure 9. With free user movement, we excluded the 360° video and compared only GS and SCGS. The participants explored the VR scene freely before teleporting to designated rocks, where they looked around for 20 seconds and answered questions.

In contrast to the exploratory study, each user received a teleportation tutorial before the actual conditions started. We then followed the same procedure as in the exploratory study.

Figure 9: **Plain GS (left) and our SCGS (left) on VR.** In the main study, we examine the effect of adding 3D assets with dynamic features (e.g., water, water currents) on user experience when moving beyond the camera path.

Table 6: **Preference rating.** Result of median M and standard deviation SD reported from the three locations in the VR environment.

	M_{GS}	SD_{GS}	M_{SCGS}	SD_{SCGS}	$t(df)$	p
How present do you feel in the environment?	5.33	1.66	6.58	1.42	-4.016(29)	<**0.001**
How ... is this location ... graphically pleasing ...	5.50	1.59	6.11	1.55	-3.610(29)	**0.001**
... visually coherent ...	4.67	1.67	6.11	1.55	-5.587(29)	<**0.001**
The water was a plausible part.	5.50	2.25	8.18	1.39	-6.965 (29)	<**0.001**
The reflection in the water matched.	5.50	2.40	7.98	1.40	-6.781 (29)	<**0.001**

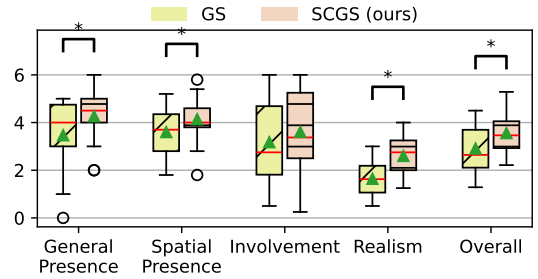

Figure 10: **IPQ results of the main study.**

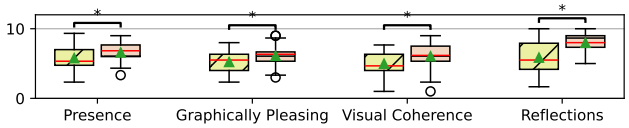

Figure 11: **Results of the individual preference questions from the selected spots.**

5.5.2 Results and Discussion

In our main study, participants experienced the same VR scene with the ability to move freely. This new interaction revealed significant improvements in General Presence, Spatial Presence, Realism, and Overall Presence, indicating higher participant engagement in the SCGS-generated scene, see Table 5 and Figure 10.

The overall ranking indicates a strong preference for SCGS, with 28 out of 30 participants favoring it. Across individual criteria (visual coherence, graphical appeal, and reflection) SCGS con-

sistently received higher ratings. Moreover, SCGS significantly scored higher for presence, which is consistent with the IPQ. Furthermore, visual coherence and realism in terms of reflections also achieved a higher scoring when using SCGS.

Several participants commented positively on SCGS and the 3D asset (flowing water): *"The movement of the water made the experience more realistic"* and *"Water effects and spatial layout were very presence-provoking."*.

In terms of criticism, the water current and depth were mentioned by $N = 4$: *"I feel more realistic, but it would be better if the ground of the water gets deeper."*. A comment possibly pointing to future work was: *"It is an environment where sounds are expected, that felt like a reminder that it was not real."*. This is consistent with feedback from the exploratory study. While the focus of our current work was on the visuals, spatial sound could be part of future work.

With regards to our two main hypotheses, a higher sense of presence is measured using IPQ when comparing plain GS and SCGS. We found significant differences in general presence, spatial presence, realism, and overall presence, confirming **HM1**, see Figure 10 and Table 5. At the individual spots participants were asked a general presence question related to the current location. There, we found no statistical significance, see Figure 11. However, the mean and median show a higher indication as well as less standard deviation when using SCGS. According to the evaluation on visual coherence and graphical pleasing, we can confirm **HM2**.

6 General Discussion

We present SCGS, semantics-controlled GS for 3D scene editing in large-scale environments. Our method is showcased on our novel outdoor dataset, additional captures, and 3D-OVS, complemented by two user experience evaluations.

6.1 Technical Aspects

SCGS enables the segmentation and removal of large scene parts, outperforming the state-of-the-art in image quality and, as shown in Figure 5, in object removal.

Existing semantic 3D GS approaches typically focus on scenes where a camera circles around a single object [24] or multiple objects [60, 28] but show limitations in their ability to remove large parts of the scene. As shown in Figure 5 and Figure 7, our approach can not only handle smaller and larger objects, it is additionally capable of successfully removing large scene parts via semantics-controlled Gaussians. This is enabled by directly assigning the class IDs to the Gaussians. With the adapted rasterization process, our approach can handle more diverse datasets.

To validate SCGS, we propose a rather complex dataset, capturing large outdoor scenes with a forward motion. The dataset is captured in a different setup compared to existing work [60] posing new challenges to separable GS and NVS. The forward motion of the camera in our dataset results in a few frames per spot, challenging both GS approaches as well as preprocessing. As depicted in Figure 5, SCGS can better handle this new dataset and is able to remove parts of the scene without affecting remaining parts. Moreover, as shown in Table 1, our approach leads to improvements in NVS quality on this dataset.

6.2 User Evaluation

Participants generally responded positively to SCGS, particularly when they were allowed to move freely. When tied to the camera-path users preferred the original panorama video which is conclusive with previous research [51] on other 3D reconstruction approaches. In our main study, we found significant differences for enhanced realism and presence in the scene generated with SCGS compared to plain GS. Criticism of the SCGS-generated scene focused on the water's depth and current, with $N = 4$ participants suggesting improvements. Our findings support our hypotheses:

HM1: *The addition of 3D assets into GS using SCGS will induce significantly higher spatial presence in users than GS alone..*

HM2: *We hypothesize that SCGS is more graphically pleasing and visually coherent than plain GS.*

The questions at individual locations and the IPQ confirmed that free user interaction enhanced realism, visual quality, coherence, and presence in SCGS compared to plain GS. Preference ratings reinforced this: 28 of 30 participants favored SCGS, while the few preferring plain GS favored the stillness of the scene.

6.3 Limitations

Our approach depends heavily on predefined labels, posing challenges with inconsistent labeling in new scenes. Furthermore, it is limited to static scenes and does not support dynamic GS.

Our study investigates the advantage of using our large-scale scene parts together with 3D assets from a Game Engine. A large, regularly dynamic part is replaced by a 3D asset. We assume that when parts of a 3D scene that are less influenced by the environment are replaced, e.g., a car or concrete of the street, the effects in presence or preference could be lower.

6.4 Future Work

From a technical and user perspective, future work could look at floating splats far outside the camera path where the 3D position is not accurately learned. Removing these could be beneficial for users who move freely. Moreover, our dataset offers potential for further improvements, for example, object removal and NVS quality for large outdoor scenes.

Future work in VR could incorporate user interactions, such as rowing [16, 46, 18] in water environments, or walk-in-place for hiking areas [15, 2]. Participants noted that sound would enhance realism, but for comparability, we focused solely on visuals and intentionally omitted sound, as it can influence presence [19, 25].

7 Conclusion

Overall, we present a novel approach for 3D asset generation based on semantics-controlled GS, alongside a new dataset featuring challenging outdoor scenes that pose various difficulties for NVS.

In summary, SCGS introduces an enhanced GS approach for generating large-scale 3D assets for VR. We evaluated SCGS from both a technical and user perspective. In the user study, we set a baseline for presence on our scenes. Therein, SCGS was compared to plain GS, with results demonstrating that SCGS significantly outperforms plain GS in terms of presence and perceived quality when users move freely within the environment. From a technical perspective, we outperform the state-of-the-art in object removal and scene editing on our new dataset. For segmentation quality we provide state-of-the-art results demonstrating that SCGS can handle a variety of different scenes. Additionally, we showcased our approach for other use cases outside of its purposed dataset, showing promising results fostering VR research.

Ethics Statement

The study involving Human Participants conducted in New Zealand was approved by the Ethics Board of the University of Otago (24/0281 - UOHEC). The user study conducted in Germany was approved by the Ethics Board of Technical University of Munich. The participants gave their written consent.

Acknowledgments

This work is partially supported by an MBIE research project (Contract UOOX2308). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

REFERENCES

[1] R. Aguina-Kang, M. Gumin, D. H. Han, S. Morris, S. J. Yoo, A. Ganeshan, R. K. Jones, Q. A. Wei, K. Fu, and D. Ritchie. Open-universe indoor scene generation using LLM program synthesis and uncurated object databases, 2024. 2

[2] E. Alvarado, O. Argudo, D. Rohmer, M.-P. Cani, and N. Pelechano. TRAIL: Simulating the impact of human locomotion on natural landscapes. *The Visual Computer*, pages 1–13, 2024. 9

[3] S. Asija, E. Du, N. Nguyen, S. Zollmann, and J. Ventura. 3d pano inpainting: Building a vr environment from a single input panorama. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 1019–1020. IEEE, 2024. 2, 3, 4, 7

[4] L. Baker, S. Mills, S. Zollmann, and J. Ventura. CasualStereo: Casual capture of stereo panoramas with spherical structure-from-motion. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 782–790, 2020. 2

[5] T. Bertel, M. Mühlhausen, M. Kappel, P. M. Bittner, C. Richardt, and M. Magnor. Depth augmented omnidirectional stereo for 6-DoF VR photography. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 660–661, 2020. 2

[6] T. Bertel, M. Yuan, R. Lindroos, and C. Richardt. OmniPhotos: Casual 360° VR photography. *ACM Trans. Graph.*, 39(6), 2020. 2

[7] Boming Zhao and Bangbang Yang, Z. Li, Z. Li, G. Zhang, J. Zhao, D. Yin, Z. Cui, and H. Bao. Factorized and controllable neural re-rendering of outdoor scene for photo extrapolation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 2

[8] J. P. A. Campos and R. Rieder. Procedural content generation using artificial intelligence for unique virtual reality game experiences. In *2019 21st Symposium on Virtual and Augmented Reality (SVR)*, pages 147–151, 2019. 2

[9] S. Chen, X. Chen, A. Pang, X. Zeng, W. Cheng, Y. Fu, F. Yin, Y. Wang, Z. Wang, C. Zhang, and others. MeshXL: Neural coordinate field for generative 3d foundation models, 2024. 2

[10] Z. Chen, J. Yang, J. Huang, R. d. Lutio, J. M. Esturo, B. Ivanovic, O. Litany, Z. Gojcic, S. Fidler, M. Pavone, L. Song, and Y. Wang. OmniRe: Omni urban scene reconstruction, 2024. 2

[11] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee. Tracking anything with decoupled video segmentation. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 5

[12] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun. FoV-NeRF: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022. 2

[13] A. Dickson, J. Shanks, J. Ventura, A. Knott, and S. Zollmann. VRVideos: A flexible pipeline for virtual reality video creation. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 199–202, 2022. 2

[14] J. Freer, K. M. Yi, W. Jiang, J. Choi, and H. J. Chang. Novel-view synthesis of human tourist photos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3069–3076, 2022. 2

[15] L. Haliburton, B. Pirker, P. Holinski, A. Schmidt, P. W. Wozniak, and M. Hoppe. VR-hiking: Physical exertion benefits mindfulness and positive emotions in virtual reality. In *Proc. ACM Hum.-Comput. Interact.*, volume 7, 2023. 9

[16] M. Hedlund, C. Bogdan, G. Meixner, and A. Matviienko. Rowing beyond: Investigating steering methods for rowing-based locomotion in virtual environments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, 2024. 9

[17] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, and C. Jiang. VR-GS: A physical dynamics-aware interactive gaussian splatting system in virtual reality, 2024. arXiv preprint arXiv:2401.16663. 2

[18] N. Keller, N. McHenry, C. Duncan, A. Johnston, R. S. Whittle, E. Koock, S. S. Bhattacharya, G. De La Torre, L. Ploutz-Snyder, M. Sheffield-Moore, G. Chamitoff, and A. Diaz-Artiles. Augmenting exercise protocols with interactive virtual reality environments. In *2021 IEEE Aerospace Conference (50100)*, pages 1–13, 2021. 9

[19] B. Kenwright. There's more to sound than meets the ear: Sound in interactive environments. *IEEE Computer Graphics and Applications*, 40(4):62–70, 2020. 9

[20] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023-07. 1, 2, 3

[21] B. Kerbl, A. Meuleman, G. Kopanas, M. Wimmer, A. Lanvin, and G. Drettakis. A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics*, 43(4), 2024. 2

[22] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 1, 4, 6

[23] C. Kleinbeck, H. Zhang, B. D. Killeen, D. Roth, and M. Unberath. Neural digital twins: reconstructing complex medical environments for spatial planning in virtual reality. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–12, 2024. 3

[24] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3, 9

[25] P. Kurucz, N. Baghaei, S. Serafin, and E. Klein. Enhancing auditory immersion in interactive virtual reality environments. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 789–792, 2023. 9

[26] C. Li, S. Li, Y. Zhao, W. Zhu, and Y. Lin. RT-NeRF: Real-time on-device neural radiance fields towards immersive AR/VR rendering. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, ICCAD '22, 2022. 2

[27] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 3

[28] K. Liu, F. Zhan, J. Zhang, M. Xu, Y. Yu, A. El Saddik, C. Theobalt, E. Xing, and S. Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 4, 6, 9

[29] Luma. Unreal Marketplace, LumaAI, 2024. https://www.unrealengine.com/marketplace/en-US/product/luma-ai. 7

[30] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. Virtual human coherence and plausibility – towards a validated scale. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 788–789, 2022. 8

[31] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. 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*, pages 7210–7219, 2021. 2

[32] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[34] P. Mohr, S. Mori, T. Langlotz, B. H. Thomas, D. Schmalstieg, and D. Kalkofen. Mixed reality light fields for interactive remote assistance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, 2020. 3

[35] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1

[36] G. Pintore, A. Jaspe-Villanueva, M. Hadwiger, E. Gobbetti, J. Schneider, and M. Agus. PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for metaverse applications. In *Proceedings of the 28th International ACM Conference on 3D Web Technology*, Web3D '23, 2023. 2

[37] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister. LangSplat: 3d lan-

guage gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, 2024. 2, 3, 4, 6

[38] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 4

[39] M. Rey-Area, M. Yuan, and C. Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3772, 2022. 4, 7

[40] T. Rolff, K. Li, J. Hertel, S. Schmidt, S. Frintrop, and F. Steinicke. Interactive VRS-NeRF: Lightning fast neural radiance field rendering for virtual reality. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, SUI '23, 2023. 2

[41] M. Sakashita, B. Thoravi Kumaravel, N. Marquardt, and A. D. Wilson. SharedNeRF: Leveraging photorealistic and view-dependent rendering for real-time and remote collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2024. 3

[42] H. Schieber, F. Deuser, B. Egger, N. Oswald, and D. Roth. Nerftrinsic four: An end-to-end trainable nerf jointly optimizing diverse intrinsic and extrinsic camera parameters. *arXiv preprint arXiv:2303.09412*, 2023. 2

[43] H. Schieber, F. Schmid, U.-H. Mubashir, S. Zollmann, and D. Roth. A modular approach for 3d reconstruction with point cloud overlay. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 609–610, 2023. 2

[44] T. W. Schubert. The sense of presence in virtual environments:. *Zeitschrift für Medienpsychologie*, 15(2):69–71, 2003. 7, 8

[45] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[46] N. A. Shoib, M. S. Sunar, N. N. M. Nor, A. Azman, M. N. Jamaludin, and H. F. M. Latip. Rowing simulation using rower machine in virtual reality. In *2020 6th International Conference on Interactive Digital Media (ICIDM)*, pages 1–6, 2020. 9

[47] M. C. Silva, M. Dahaghin, M. Toso, and A. Del Bue. Contrastive gaussian clustering: Weakly supervised 3d scene segmentation. *arXiv preprint arXiv:2404.12784*, 2024. 1, 3, 4, 6

[48] M. Slater, A. Steed, J. McCarthy, and F. Maringelli. The influence of body movement on subjective presence in virtual environments. *Human factors*, 40(3):469–477, 1998. 8

[49] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846, 2006. 2

[50] K. Szita, P. Gander, and D. Wallstén. The effects of cinematic virtual reality on viewing experience and the recollection of narrative elements. *PRESENCE: Virtual and Augmented Reality*, 27(4):410–425, 2018. 8

[51] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock. Mixed reality remote collaboration combining 360 video and 3d reconstruction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–14, 2019. 3, 7, 9

[52] L. Tong, R. W. Lindeman, H. Lukosch, R. Clifford, and H. Regenbrecht. Applying cinematic virtual reality with adaptability to indigenous storytelling. *J. Comput. Cult. Herit.*, 17(2), 2024. Place: New York, NY, USA Publisher: Association for Computing Machinery. 2, 3

[53] T. Q. Tran, T. Langlotz, J. Young, T. W. Schubert, and H. Regenbrecht. Classifying presence scores: Insights and analysis from two decades of the igroup presence questionnaire (ipq). *ACM Transactions on Computer-Human Interaction*, 2024. 7

[54] J. Waidhofer, R. Gadgil, A. Dickson, S. Zollmann, and J. Ventura. PanoSynthVR: Toward light-weight 360-degree view synthesis from a single panoramic input. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 584–592, 2022. 2

[55] C. Wang, M. Chai, M. He, D. Chen, and J. 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*, pages 3835–3844, 2022. 2

[56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[57] L. Xu, V. Agrawal, W. Laney, T. Garcia, A. Bansal, C. Kim, S. Rota Bulò, L. Porzi, P. Kontschieder, A. Božič, D. Lin, M. Zollhöfer, and C. Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Proceedings*, 2023. 2

[58] B. Yang, Y. Zhang, Y. Xu, Y. Li, H. Zhou, H. Bao, G. Zhang, and Z. Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[59] J. Yang, X. Chen, S. Qian, N. Madaan, M. Iyengar, D. F. Fouhey, and J. Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701, 2024. 2

[60] M. Ye, M. Danelljan, F. Yu, and L. Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, 2024. 1, 2, 3, 4, 5, 6, 9

[61] Z. Yin, Y. Wang, T. Papatheodorou, and P. Hui. Text2vrscene: Exploring the framework of automated text-driven generation system for VR experience. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 701–711, 2024. 2

[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[63] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *In Proceedings CVF/IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2024*, 2024. 2, 4

[64] S. Zollmann, A. Dickson, and J. Ventura. CasualVRVideos: VR videos from casual stationary videos. In *Proceedings of the 26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20, 2020. 2