2022 EC³
2022 European Conference on Computing in Construction
Ixia, Rhodes, Greece
July 24-26, 2022

SEMSPRAY: VIRTUAL REALITY AS-IS SEMANTIC INFORMATION LABELING TOOL FOR 3D SPATIAL DATA

Yiming Zhao*[1], Cyprien Fol*[1], Yuchang Jiang[1], Tianyu Wu[1], and Iro Armeni[1]
[1]ETH Zurich, Zurich, Switzerland

Abstract

Capturing the as-is status of buildings in the form of 3D spatial data (e.g., point cloud or mesh) with the use of 3D sensing technologies is becoming predominant in the Architecture, Engineering, and Construction (AEC) industry. Although the act of acquiring this data has been progressively becoming more accurate and efficient with the availability of off-the-shelf solutions in the market, the act of extracting from it as-is information has not seen similar advancements. State-of-the-art practice requires experts to manually interact with the spatial data in a laborious and time-consuming process. We propose Semantic Spray (*Semspray*), a Virtual Reality (VR) application that provides users with intuitive and handy tools to produce semantic information on as-is 3D spatial data (mesh) of buildings. The goal is to perform this task accurately and more efficiently by allowing users to experience, interact with, and immerse themselves in the data at different scales. *Semspray* is a combination of two labeling modes: *user-dynamic* and *user-static*. In the *user-dynamic* mode, the user is fully immersed in the 3D mesh and has the ability to walk and teleport themselves within the model; in the *user-static*, the user can comfortably sit on a chair and handle a small-scale version of the 3D mesh to perform the labeling, in a similar manner to hand-held painting. We evaluated *SemSpray*'s performance with a user study of ten participants and found that the combination of the different modes is able to keep the user entertained and to limit the side-effect of VR on the sensory organs, including nausea, headache, and dizziness.

Introduction

The vast commercialization of 3D sensing technology has made the 3D reconstruction of our built environment easy to acquire. This has considerable implications on the Architecture, Engineering, and Construction (AEC) industry, since the availability of accurate as-is building information can be beneficial to many industry processes: from designing for renovation to construction progress monitoring and facility management. Trends denote that an increasing number of AEC practitioners is utilizing such technologies to acquire 3D reconstructions of building as-is status, however it has also resulted in a very common question: how can one extract semantic information from the 3D geometry (e.g., 3D point cloud or mesh) that reconstruction systems produce?

Despite extensive research in the development of automatic methods for 3D semantic understanding of such data (e.g., (Tchapmi et al. 2017, Choy et al. 2019, Qi et al. 2017, Poux & Billen 2019, Bassier et al. 2020)), the results are not accurate, robust, or flexible enough for the requirements of the AEC industry. As a result, manual or assisted methods with experts operating dedicated 3D software remain the industry standard. This is a laborious, time-consuming, and error-prone process (Brilakis et al. 2010, Woo et al. 2010, Jung et al. 2014), partially due to the 2D way with which users interact with the 3D data. Virtual Reality (VR) technology has been gradually explored as a means to perform tasks in the AEC industry because of its ability to fully immerse users in a virtual setting. The ability to disconnect from the real world allows to explore different perspectives and ways to approach the task, in a way that would not be feasible in the physical world. Specifically for the task of attributing semantic meaning to 3D mesh data, VR technology can provide a platform for non-expert users to produce as-is information fast and in a gamified experience.

To this end, we develop Semantic Spray (*SemSpray*), a VR application for the task of semantic labeling of 3D mesh reconstructions of buildings. The aim is for *SemSpray* to allow non-expert users to accurately annotate the mesh of reconstructed *scenes* with intuitive VR tools in an immersive and user-friendly manner. *SemSpray* consists of two annotation modes, which act complementary and offer the user different perspectives and ways of interacting with the 3D mesh data. Specifically, the *user-dynamic* mode offers an ego-centric, first-person view of the environment; this mode is an extension and adaptation of the *Shooting Labels* work by Ramirez et al. (2019). The second mode is the *user-static* mode, which offers an allo-centric view of the data that allows users to detach themselves from the high-immersion and physical stress of the first mode. A user study was conducted with ten participants to evaluate the usability of *SemSpray* and to assess the two modes and the accuracy of the produced annotations. We performed the study on 3D mesh data of large-scale reconstructions of real-world indoor scenes that consist of cluttered office spaces, so as to better gauge the efficacy and usability of *SemSpray*.

The contributions of this work are three-fold:

1. We adapted features of *Shooting Labels* to indoor and cluttered environments and the specific task requirements (e.g., allowing high accuracy).
2. We developed two different modes that offer the user different ways of interacting with the data. Given how distinct the ways are, users might choose one or the other, or a combination, depending on the scene to annotate and on their physiological reaction to spending time in a virtual environment.
3. We conducted a user study to evaluate the usability, efficiency, and accuracy of using *SemSpray*.

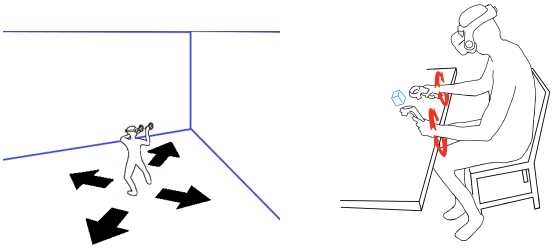
* Both authors contributed equally

Related work

Nowadays, the task of 3D semantic labeling is mostly performed manually, which causes a major bottleneck in acquiring as-is information. To avoid this problem, methods are implementing semi-automatic approaches that can speed up the annotation (e.g., (Wong et al. 2015, Armeni et al. 2019, Nguyen et al. 2021, Dai et al. 2017, Nguyen et al. 2016, Russell & Torralba 2009)). Semantic annotation tools have been implemented for augmented and virtual reality devices as well. For example, Miksik et al. (2015) introduce the labeling task during the acquisition process for an augmented reality (AR) system. Thanks to an infrared laser pointer, the AR system can annotate the object being captured in real-time. Saran et al. (2018) created an iOS application for simultaneous scanning and user-defined bounding box annotation. Furthermore, a collaborative VR system has been developed by Zingsheim et al. (2021), that enables the labeling of live-captured scenes by remote users with sparse labels. Ramirez et al. (2019) convert the tedious task of annotating into a playful first person shooter game, which we adopt as a partial basis for our system.

Semantic Spray

Semantic Spray (*SemSpray*) is a VR application for users to provide semantic labels on 3D mesh data of building scenes. Inspired by Ramirez et al. (2019), we developed a gamified experience for performing this task. The application has two modes: *user-dynamic* and *user-static*. An illustration of the two modes can be visualized in Figure 1.



(a) User-dynamic (b) User-static

Figure 1: User setup in each mode of SemSpray

User-Dynamic Mode

This mode is an extension of the Shooting Labels work by Ramirez et al. (2019). In a nutshell, Shooting Labels aims to create a gamified experience for solving the task of semantic labeling of 3D data by utilizing different weapons that paint mesh surfaces with semantic labels (each label is represented by a unique color). After thorough testing of this application, we identified that, despite having well implemented basic functionalities for semantic labeling, certain limitations are hindering the performance of the tool in cluttered indoor scenes.

Labeling Precision

The variety of different weapons - although fun and exciting - causes issues of accuracy and inefficiency when dealing with cluttered indoor scenes. The majority of the weapons does not allow for detailed and precise labeling of surfaces especially in the case of thinner and/or intricate geometric shapes found on objects and furniture in such scenes. To improve this limitation, we replaced the weapons with sprays; the nozzle size of the spray is a parameter that the user can define. Intuitively, this would suggest to the user a less harsh and more artistic approach and thus favor a more precise labeling. We also provide the user with the ability to visualize an RGB-textured mesh (*RGB mode*), in addition to a non-textured one. This can be of great assistance to the user when trying to disambiguate about mesh faces and the objects to which they belong.

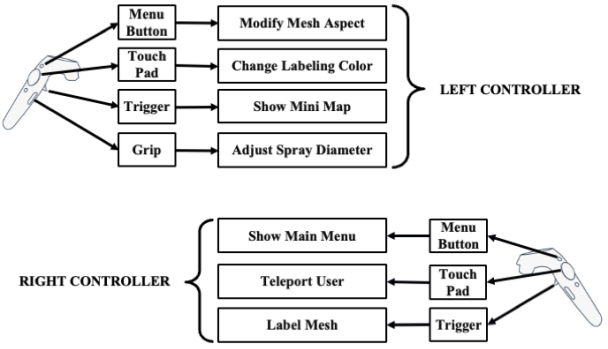


Figure 2: Diagrams of controller inputs in user-dynamic mode of SemSpray.

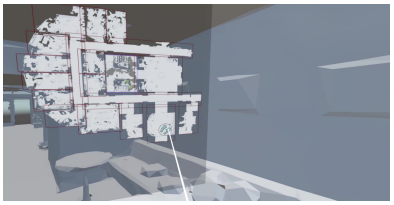
User Feedback

During labeling, the user is not given feedback on the 3D surface to which they are pointing with the weapon. They are also not made aware of the effect radius of one blast, which differs drastically for different weapons. To provide the user with feedback on where they are aiming, and with the intention of increasing the labeling accuracy, we attach a raycast tracking system to the nozzle of the spray. In addition, the user can select from a variety of different nozzle sizes and visualize how large of a radius each one can spray. The mesh triangles (faces) at which the user is aiming get highlighted in green under the influence of a specific nozzle size, hence informing the user of the final effect of their next action.

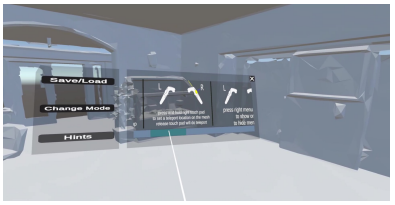
Visualization Uncertainty

When labeling the mesh with Shooting Label, a color gradient occurs systematically on the annotated surface because of the way that labels are assigned to face vertices. Specifically, the color of the face becomes solid only when all three vertices of the face obtain the same label. Apart from issues with consolidating labels on one face, this visualization is confusing to the user. We addressed the gradient effect by attributing labels to faces and not individual vertices – during labeling, if a face is within the

(a) Minimap structure in RGB mode



(b) Inspect annotation status with minimap



(c) User manual in User-dynamic mode

Figure 3: Examples of functionalities in the user-dynamic mode in SemSpray

effect of the spray it will be assigned the designated label.

Change of Perspective

Shooting Labels allows the user to change their perspective of the scene by lifting themselves from the virtual ground in the air using a jetpack functionality. This functionality is not a good fit for indoor scenes that are usually small and enclosed. More than often it generates motion-sickness if the user accidentally activates it. To facilitate the annotation of large-scale reconstructions composed of several rooms, we implemented instead a mini-map functionality, which allows users to teleport from one room to another in a single click. This reduces the discomfort from the wide motion caused by the jetpack.

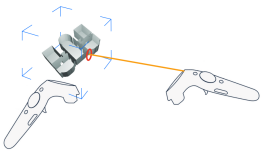
In Figure 2, we provide a detailed schematic of the framework and functionalities of the left and right controllers of the VR device for the *user-dynamic* mode in *SemSpray*. Figure 3 showcases examples of features from the user viewpoint in the dynamic mode.

User-Static Mode

Despite the user-dynamic mode being better adapted to cluttered indoor scenes, there are still some drawbacks. Semantic annotation of these scenes can last long – the user-dynamic mode cannot support a long-wear use in one session due to the following:

- **VR-induced Sickness:** VR users usually face cybersickness (LaViola 2000) or simulator sickness (Johnson 2005). These are a consequence of vergence-accommodation conflict, which results from a disconnect between the sensation of visual movement and the body's vestibular system – a collection of mechanisms in the inner ear that controls one's sense of balance and monitors spatial orientation. This could occur during the motion when using teleportation or delayed display when moving the head. Such sickness usually appears in less than 30 minutes after the user enters the VR environment.
- **Fatigue:** Performing the task while wearing a bulky VR device can easily fatigue the user. Indeed, the user-dynamic mode, which is thought as a first-person shooter application, requires the user to move the body frequently and sometimes in challenging poses so as to reach some corners or the bottom surfaces of objects in the virtual scene.
- **Physical Space Restrictions:** A large empty room dedicated to VR is a luxury for many users and a VR device set-up in offices or living rooms often encounters occlusion from objects physically present in the room, hence restraining the range of motion.

(a) Painting Handle



(b) Concept for mesh interaction

Figure 4: Inspirations of User-static mode

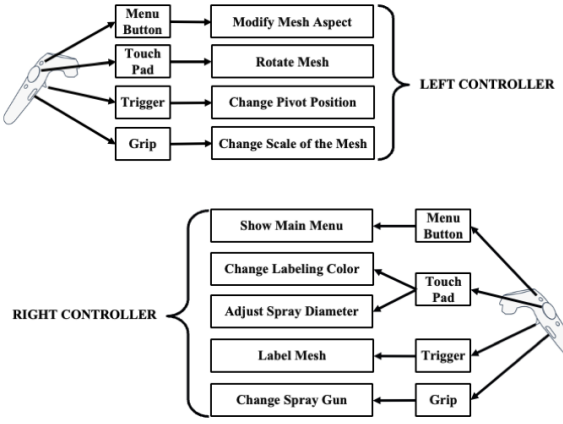
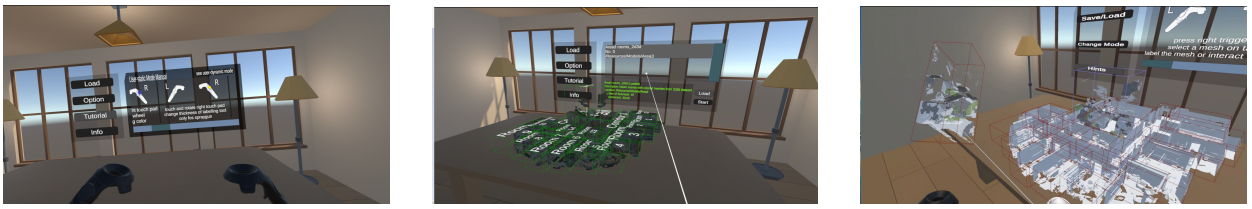


Figure 5: Diagrams of controller inputs in user-static mode of SemSpray.

To address the above, we implemented a second mode (*user-static*) that will enable the user to label the mesh of a scene only with the motion of two hand controllers, while remaining seated in a calm and neutral VR room (and subsequently in their physical space). Figure 4 shows the concept of user-static mode, which is inspired by the painting handle tool for miniature models. Essentially, we develop in the virtual world a painting handle that holds the scene to annotate, so that the user can interact with it from different perspectives and scales, while simultaneously remaining seated. First, we rescale the mesh to a smaller size. The user can move, rotate, and re-scale this



(a) GUI and user manuals (b) Initial minimap overview (c) Split dataset with bounding box

Figure 6: Examples of functionalities in the user-static mode of SemSpray

mesh with one controller, using the other one as a brush to label the mesh. As in the *user-dynamic* mode, here as well we keep functionalities such as brushes of different size, providing user with feedback on where they are aiming, allowing to view the texture RGB mesh, and using the minimap to change the space they are holding.
In Figure 5, we provide a detailed schematic of the framework and functionalities of the left and right controllers of the VR device for the *user-static* mode in our application. Figure 6 showcases examples of features from the user viewpoint in the static mode.

Fusion of the Dynamic and Static Modes

We connect the two modes using the Unity scene manager system. This permits to exchange mesh and semantic information between modes. The transition between the modes is facilitated by a Graphical User Interface (GUI) menu. A diagram of the communication between the two modes is illustrated in Figure 7.

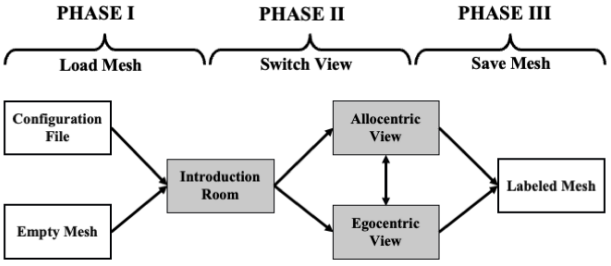


Figure 7: Fusion of two modes

User Study

To evaluate *SemSpray* and to compare the two modes (*user-dynamic* and *user-static*), we performed a usability study with 10 participants. For each mode, 5 participants are invited to test the application. The experiment is divided into three parts: pre-task questionnaire, task completion, and post-task questionnaire. The pre-task questionnaire aims at gathering the background information of participants, including any prior experience with VR devices and 3D labeling tools, while the post-task questionnaire targets at collecting the subjective opinions of participants after using *SemSpray*. Evaluation focuses on completing an annotation task, which includes the following aspects:

- Label specific objects in the designated scene (floor, table, sofa, chair).
- Adjust the size of the labeling brush/spray nozzle.
- Visualize the RGB mode.
- Use the minimap to change location.

Dataset

We performed the user study in the 3D mesh reconstructions of real-world cluttered scenes provided in the 2D-3D-Semantics dataset (Armeni et al. 2017), which includes the raw textured 3D meshes together with their semantically labeled ground truth. We use the ground truth to assess the accuracy of *SemSpray*. These meshes provide a practical scenario where *SemSpray* would be needed. We focus our study in one of the provided six areas (Area 3).

Metrics

Specifically, we aim to evaluate the usability, efficiency, and accuracy of each mode:

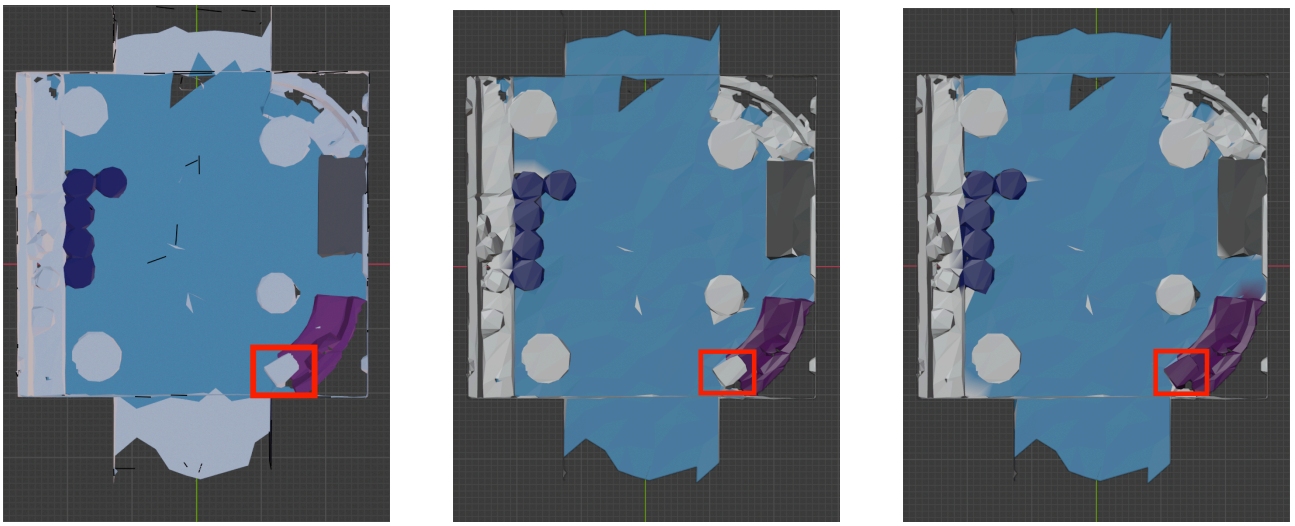
- **Usability**: We choose the System Usability Scale (SUS, (Brooke et al. 1996)) to measure the level of usability, which can convert answers like *Strongly Agree* to numerical scores. The SUS score of the application was converted to the range of 0-100. The higher the score, the higher the usability is. The average SUS score for such applications is 68.
- **Efficiency**: The time required to complete the task is used as a metric to reflect the efficiency of our application.
- **Accuracy**: To measure the accuracy of labeling, we selected the standard 3D semantic segmentation metrics - and specifically the precision (Equation (1)), recall (Equation (2)), and accuracy (Equation (3)) - to compare participants' labeling results to the ground truth.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where TP, TN, FP, and FN stand for *True Positive*, *True Negative*, *False Positive*, and *False Negative*.



(a) Ground Truth (b) User-dynamic mode (c) User-static mode

Figure 8: Examples of labelled mesh in user-dynamic and user-static mode with respect to ground truth

Results

In this section, we will address the qualitative results and then present the quantitative results of our usability study corresponding to *usability*, *efficiency*, and *accuracy* of *SemSpray*.

Qualitative analysis

Figure 8 illustrates two example labeling results from using the *user-dynamic* and *user-static* modes respectively. Qualitatively comparing with the ground truth, results show that it is feasible to achieve very good labeling output using *SemSpray*. One noticeable difference is the small cushion on the sofa marked in a red square. On the one hand, different users may have different judgement on whether this small object belongs to the label *sofa* or not. On the other hand, as the whole room is a small object in the *user-static* mode, the lack of immersive experience may prevent the user from labeling some small objects correctly. We should point out that in the ground truth data the pillow is labeled with the *sofa* label.

Usability

Figure 9 shows the histogram of the SUS scores of all ten participants. All reported SUS scores are above the average value of 68, but one. The average SUS scores for the *user-dynamic* mode and *user-static* mode are 77.5 and 83.5 respectively. This suggests a similar usability of the two modes and this result can be related to the background of users, which will be further discussed in a later section.

Efficiency

For each task, we recorded the time that users required to complete it. In Figure 10, we summarize the average time in user-dynamic mode, user-static mode, and both modes. It indicates a similar tendency: for example, labeling floor requires more time, while labeling sofa is comparatively an easy task for both modes. Floor is a harder task here because it has contact with many different objects and it requires more attention and greater precision to achieve good labeling results at object boundaries, e.g. the boundary between floor and wall. Besides labeling objects, the functionality of the minimap is also assessed. Transferring between different rooms and corridors with this function is easy-to-use and it only takes users seconds to complete.

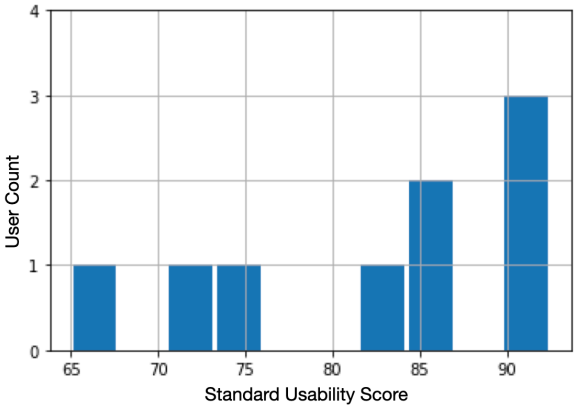


Figure 9: Results on System Usability Scale (SUS) scores

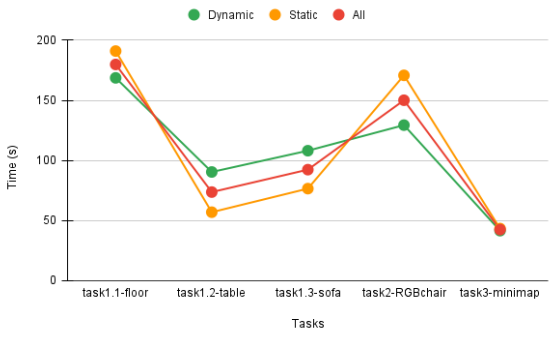


Figure 10: Result of time required to complete tasks

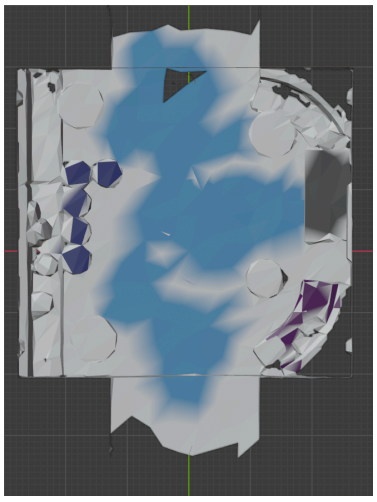


Figure 11: An example of a careless user behavior

Accuracy

Similar to the measurement of efficiency, we summarize the average precision, recall, and accuracy in dynamic mode, static mode, and both modes (as shown in Figures 13, 14, and 12). The accuracy plot (Figure 12) showcases that both modes have similar performance. On the contrary, recall when labeling *sofa* differs across modes. According to Figure 14, the dynamic mode leads to worse recall on this task. However, as more participants in the *user-static* mode evaluation had prior VR experience, our experiment could be biased.

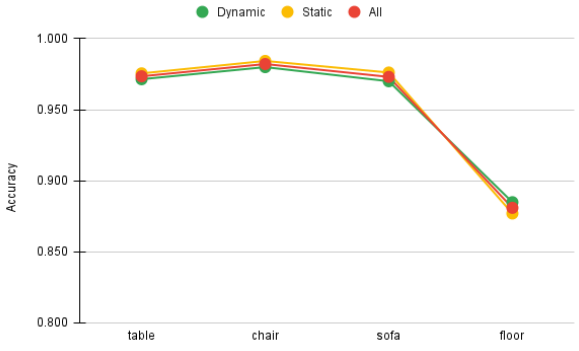


Figure 12: Evaluation of labeling accuracy.

Discussion

Comparison of two modes

According to the standard of SUS, an above-average application generally achieves more than 68 in SUS score. Based on the results of the post-study questionnaire, the average SUS scores of *user-static* and *user-dynamic* mode are around 84 and 78 respectively, suggesting that both modes demonstrate good usability to the users. A probable explanation for the slightly lower SUS score in the *user-dynamic* group is that the participants had less experience with VR devices and applications than the other

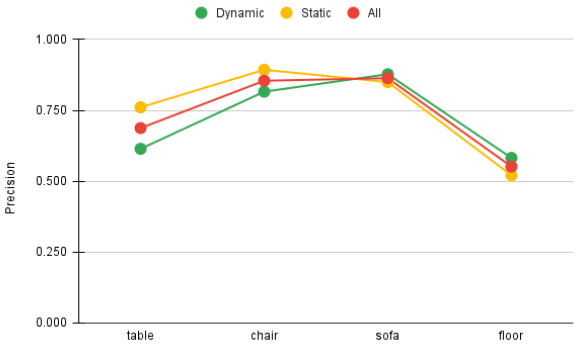


Figure 13: Evaluation of labeling precision.

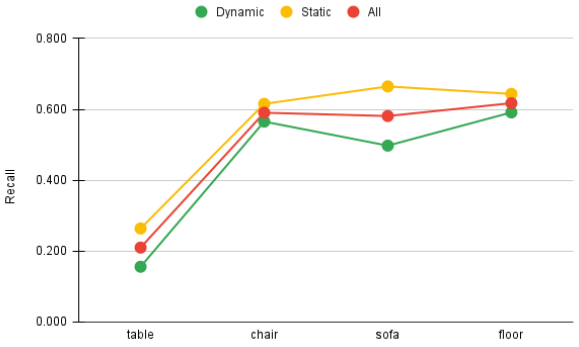
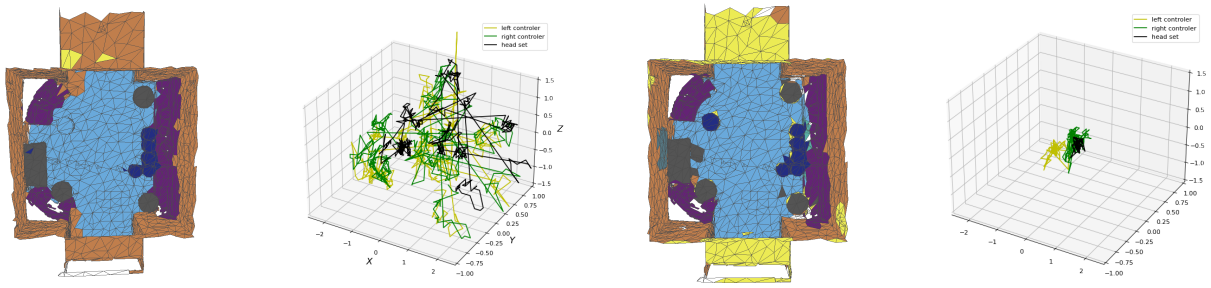


Figure 14: Evaluation of labeling recall.

group, so more efforts may be needed to familiarize with the technology initially. With respect to the evaluation of efficiency, Figure 10 shows the average task completion time of the two modes on different objects and there seems to be no clear distinction on which mode outperforms the other on all types of objects. Regarding labeling accuracy, *user-static* mode demonstrates a slightly better performance than *user-dynamic* mode. Nevertheless, as discussed in the previous section, this discrepancy could be attributed to the different levels of background and experiences of the participants.

We also performed an experiment to assess the physical space that each mode occupies when used. The results can be seen in Figure 15. For the labeling the same 3D scene, the user-dynamic mode requires almost as big of a space as the mesh to label – in this example it requires a free of obstacle space of approximate size $5m \times 2m$ (Figure 15(b)). In contrast, the user-static mode requires a substantially smaller space, which in this example is less than $1m \times 1m$ (Figure 15(d)). However, as can be seen in Figure 15(a) and (c), the labeling results of user-static mode are less accurate than those of the user-dynamic mode. These findings can be of particular practical importance; the user cab begin by providing all labels in user-static mode and then strategically enter user-dynamic to address any inaccuracies.

In summary, though preliminary, the user study shows that each mode has its own benefits; hence a combination



(a) Labeled mesh in dynamic mode

(b) Trajectory in dynamic mode

(c) Labeled mesh in static mode

(d) Trajectory in static mode

Figure 15: Trajectory of hands and head in 3D space during the two modes.

of both modes is beneficial for completing the task of semantic labeling.

Limitations of User Study

During the study, we found that different users could behave very differently even on the same task. For example, when the users were asked to label an object, some labeled it very fast and carelessly (e.g. Figure 11), while others tried to label every piece and corner of the object (e.g. Figure 8b). These diverse behaviors have a significant effect on the task completion time and accuracy. Furthermore, the participants were split randomly into two groups to test different modes regardless of their VR background or experiences. However, during the analysis of the user study, we found that all three users who have previous experience with labeling tools or 3D projects were assigned to the *user-static* mode, while none of the participants in the *user-dynamic* have background knowledge in related field. Therefore, it is important to bear in mind the possible bias in their performances.

Conclusion

In conclusion, we have developed a VR 3D labeling application that is user-friendly, accurate, and efficient. In this application, we built upon *Shooting Labels* – an existing VR tool, advancing its functionalities while proposing a new labeling mode. We demonstrated the usability of *SemSpray* in a user study with 10 participants. The user study showed that both labeling modes have their own distinctive advantages; their combined use can maximize their benefits for a well-performed labeling task.

References

Armeni, I., He, Z.-Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J. & Savarese, S. (2019), 3d scene graph: A structure for unified semantics, 3d space, and camera, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 5664–5673.

Armeni, I., Sax, S., Zamir, A. R. & Savarese, S. (2017), 'Joint 2d-3d-semantic data for indoor scene understanding', *arXiv preprint arXiv:1702.01105* .

Bassier, M., Vergauwen, M. & Poux, F. (2020), 'Point cloud vs. mesh features for building interior classification', *Remote Sensing* **12**(14).
URL: *https://www.mdpi.com/2072-4292/12/14/2224*

Brilakis, I., Lourakis, M., Sacks, R., Savarese, S., Christodoulou, S., Teizer, J. & Makhmalbaf, A. (2010), 'Toward automated generation of parametric bims based on hybrid video and laser scanning data', *Advanced Engineering Informatics* **24**(4), 456–465.

Brooke, J. et al. (1996), 'Sus-a quick and dirty usability scale', *Usability evaluation in industry* **189**(194), 4–7.

Choy, C., Gwak, J. & Savarese, S. (2019), 4d spatio-temporal convnets: Minkowski convolutional neural networks, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 3075–3084.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T. & Nießner, M. (2017), Scannet: Richly-annotated 3d reconstructions of indoor scenes, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5828–5839.

Johnson, D. (2005), 'Introduction to and review of simulator sickness research'.

Jung, J., Hong, S., Jeong, S., Kim, S., Cho, H., Hong, S. & Heo, J. (2014), 'Productive modeling for development of as-built bim of existing indoor structures', *Automation in Construction* **42**, 68–77.

LaViola, J. J. (2000), 'A discussion of cybersickness in virtual environments', *SIGCHI Bull.* **32**(1), 4756.
URL: *https://doi.org/10.1145/333329.333344*

Miksik, O., Vineet, V., Lidegaard, M., Prasaath, R., NieSSner, M., Golodetz, S., Hicks, S. L., Pérez, P., Izadi, S. & Torr, P. H. (2015), The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces, Vol. 2015-April, Association for Computing Machinery, pp. 3317–3326.

Nguyen, D. T., Hua, B.-S., Yu, L.-F. & Yeung, S.-K. (2016), 'A robust 3d-2d interactive tool for scene segmentation and annotation'.

Nguyen, T., Hua, B.-S., Nguyen, D. T. & Phung, D. (2021), 'Single-click 3d object annotation on lidar point clouds'.

Poux, F. & Billen, R. (2019), 'Voxel-based 3d point cloud semantic segmentation: Unsupervised geometric and relationship featuring vs deep learning methods', *ISPRS International Journal of Geo-Information* **8**(5).
URL: *https://www.mdpi.com/2220-9964/8/5/213*

Qi, C. R., Su, H., Mo, K. & Guibas, L. J. (2017), 'Pointnet: Deep learning on point sets for 3d classification and segmentation'.

Ramirez, P. Z., Paternesi, C., Luigi, L. D., Lella, L., Gregorio, D. D. & Stefano, L. D. (2019), 'Shooting labels: 3d semantic labeling by virtual reality'.
URL: *http://arxiv.org/abs/1910.05021*

Russell, B. C. & Torralba, A. (2009), Building a database of 3d scenes from user annotations, *in* '2009 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 2711–2718.

Saran, V., Lin, J. & Zakhor, A. (2018), Augmented annotations: Indoor dataset generation with augmented reality.

Tchapmi, L., Choy, C., Armeni, I., Gwak, J. & Savarese, S. (2017), Segcloud: Semantic segmentation of 3d point clouds, *in* '2017 international conference on 3D vision (3DV)', IEEE, pp. 537–547.

Wong, Y.-S., Chu, H.-K. & Mitra, N. J. (2015), 'Smartannotator an interactive tool for annotating indoor rgbd images', **34**.

Woo, J., Wilsmann, J. & Kang, D. (2010), Use of as-built building information modeling, *in* 'Construction Research Congress 2010: Innovation for Reshaping Construction Practice', pp. 538–548.

Zingsheim, D., Stotko, P., Krumpen, S., Weinmann, M. & Klein, R. (2021), 'Collaborative vr-based 3d labeling of live-captured scenes by remote users', *IEEE Computer Graphics and Applications* **41**, 90–98.