2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

PointCloudLab: An Environment for 3D Point Cloud Annotation with Adapted Visual Aids and Levels of Immersion

Achref Doula*, Tobias Güdelhöfer*, Andrii Matviienko**, Max Mühlhäuser* and Alejandro Sanchez Guinea*

***Abstract*— The annotation of 3D point cloud datasets is an expensive and tedious task. To optimize the annotation process, recent works have proposed the use of environments with higher levels of immersion in combination with different types of visual aids. However, two problems remain unresolved. First, the proposed environments limit the user to a unique level of immersion and a fixed hardware setup. Second, their design overlooks the interaction effects between the level of immersion and the visual aids on the quality of the annotation process. To address these issues, we propose *PointCloudLab*, an environment for 3D point cloud annotation that allows the use of different levels of immersion that work in combination with visual aids. Using *PointCloudLab*, we conducted a controlled experiment (N=20) to investigate the effects of levels of immersion and visual aids on the annotation process. Our findings reveal that higher levels of immersion combined with object-based visual aids lead to a faster and more accurate annotation. Furthermore, we found significant interaction effects between the levels of immersion and the visual aids on the accuracy of the annotation.**

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160225

I. Introduction

The annotation of datasets used for the training of machine learning (ML) models requires a considerable effort and induces huge financial costs. The annotation process becomes particularly complex and time-consuming when the dataset consists of 3D point cloud scans representing complex scenes, like urban areas. A concrete example of this is the annotation of the SemanticKitti dataset[1], which represents one of the largely used point cloud datasets for urban scene understanding. The dataset required a total of 1700 hours of annotation [1]. Therefore, the need for approaches that make the annotation process more efficient and convenient is clear.

To overcome the aforementioned challenges, two strategies have been adopted. The first strategy is based on the design of new learning paradigms that limit the need for annotated data. Techniques like unsupervised learning [2], self-supervised learning [3], and few shots learning [4] have been successfully used to train neural networks on scene understanding tasks of 3D point clouds. However, such approaches are difficult to design and require expert knowledge. Furthermore, their performance is questionable compared to the traditional approaches [5]. The second strategy focuses on designing software applications that facilitate the annotation process for human annotators. To this end, several previous works proposed software tools that supported the annotation process with visual aids [6], [7]. To enhance the process even further, some works have leveraged the recent advances in virtual reality to incorporate visual aids in immersive annotation environments [8]. Although the immersive aspect of these tools can facilitate an intuitive perception and interaction with 3D point clouds, two important limitations remain unresolved. First, such environments limit the user to one particular level of immersion, which may lead to unwanted effects during long annotation sessions, such as fatigue [9]. Second, it is unclear how to adapt the visual aids to the level of immersion to avoid perception-related issues, and ensure the quality of the annotation process. Consequently, it is crucial to investigate the influence of visual aids and the levels of immersion, and their effect on the annotation process to determine design guidelines for future annotation environments built for different levels of immersion.

In this work, we investigate the effects of levels of immersion and visual aids on different aspects of the annotation process of 3D point clouds in terms of accuracy, speed, and workload. For this, we developed *PointCloudLab*, a software platform that facilitates an annotation process under 3 levels of immersion, namely (IM-1) non-immersive, (IM-2) semi-immersive, and (IM-3) fully-immersive, and in combination with 3 different visual aids, which are (VA-1) distance-based point coloring, (VA-2) region-based point-coloring, and (VA-3) object-based point highlighting. *PointCloudLab* allows for a hardware-agnostic annotation process that provides an evaluation environment for annotation. Using this developed platform, we conducted a controlled experiment (N=20), in which we investigated how we can improve the accuracy, speed, and mental load during an annotation process given a particular level of immersion and visual aid. The results of the user study revealed a strong interaction between the factors of interest. Furthermore, our observations state that while higher levels of immersion lead to shorter annotation times and higher user involvement, visual aids have an impact on the accuracy of the annotation

II. Related Work

A number of works have proposed non-immersive tools for the annotation of 3D point cloud datasets. The work in [10] created a web-based tool that allows labeling of point clouds while viewing corresponding images recorded with a surround camera. As an additional measure to reduce the annotation efforts, they used label interpolation to propagate labels between frames in a sequence, which helped speed

*Telecooperation Lab, Technical University of Darmstadt, Germany {doula, max, sanchez}@tk.tu-darmstadt.de.
**Media Technology and Interaction Design, KTH Royal Institute of Technology, Sweden matviienko.andrii@kth.se.

[1]SemanticKitti dataset: http://www.semantic-kitti.org/

up the annotation of sequences significantly. Mathwork's Ground Truth Labeler [6] is a tool included in MATLAB for labeling of various sensor inputs, including point clouds. The tool allows for the automation of labeling with user-provided algorithms. Works like [11], [12] leveraged devices with higher degrees of freedom to achieve an easier manipulation of the 3D data.

While not as numerous, some works propose to visualize data via Virtual and Augmented Reality. In [13] the authors present a tool that allows for immersive segmentation labeling of scenes using a shooting metaphor. They found that users were considerably faster using their tool than the creators of the semantic KITTI dataset [1] reported their users being. The work in [13] proposes to improve the quality of annotation with the usage of an uncertainty factor calculated by combining labels from multiple users to counter this. In [7] the authors developed an immersive annotation tool, in which users can grab the scene and move it around themselves, instead of moving towards the objects. In [14] the authors proposed *Immersive labeler*, a fully immersive annotation tool, where the users can interact with the 3D point cloud scans using various techniques, such as "as a giant" locomotion. Similar to our work, they evaluated their setup with different levels of immersion. However, they only reported results for the annotation time. In this work, we additionally focus on more relevant metrics, such as the accuracy of the annotation and the mental load. Furthermore, the results of our Likert questionnaires reveal additional decisive factors for the annotation process, such as the natural scaling of objects. In [8] the authors created a tool that allows viewing point cloud sequences like a video player. The user can label points by using a controller to select points via a paintbrush metaphor. In order to ease the labeling of sequential data, they implemented two algorithms that allow for the propagation of labels into later frames of the sequence. The work in [15] created a workflow in which the user wears a Virtual Reality headset and a depth camera. The user sees a virtual representation of objects in front of them and selects areas by touching objects with their hands.

While the aforementioned works reported incorporating different techniques to improve the user experience, like changing the levels of detail in the virtual environment [7], there is no empirical evidence for a desirable impact of immersion on perception and on the annotation process. Furthermore, it is not clear whether the same visual aids

III. PointCloudLab

In this paper we introduce PointCloudLab, a platform that allows to perform annotation of point cloud data at different levels of immersion: non-immersive, semi-immersive, and fully immersive. The annotation with *PointCloudLab* can be performed using different hardware setups and adapted to the desired level of immersion by choosing one out of three display options: (1) 2D monitors, (2) projectors, and (3) HMDs. Furthermore, our annotation platform facilitates the activation or deactivation of body tracking hardware, such as motion tracking and head tracking, which allows the user to change the feeling of immersion gradually [16]. This control over the annotation environment promotes hybrid annotation teams with different and personalized hardware setups. Additionally, our annotation platform considers different visual aids, namely distance-based coloring, region-based coloring, and object-based highlighting. The role of the visual aids is to provide additional support to the annotators while interacting with the data. Furthermore, *PointCloudLab* provides a set of metrics that can be used to evaluate the interaction with the point cloud scene. These metrics are the annotation time, accuracy, and mental load based on [17]. Our annotation platform is illustrated in Figure 3.

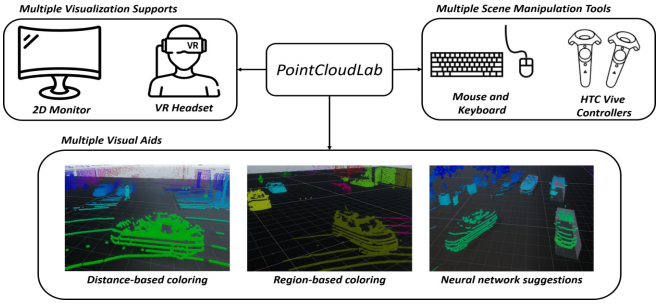

Fig. 1. PointCloudLab facilitates 3D point cloud annotation with multiple visualization supports, multiple controllers and multiple visual aids.

IV. Study: Effects of the Level of Immersion and Visual Aids on the Annotation of 3D Point Clouds

To evaluate the influence of the level of immersion and visual aids on the annotation process of 3D point clouds, we conducted a controlled experiment leveraging *PointCloudLab* as an experimental environment.

A. Study Setup and Apparatus

To implement our proposed annotation system, we used Unity and its High Definition Render pipeline (HDRP) to allocate most rendering work onto the GPU. The HDRP shader supports the rendering of large points clouds scenes on a desktop-class computer with a dedicated GPU. For our VR setup, we utilized Steam VR[2] as the VR platform and the HTC Vive[3] as the VR-system. Further, we employed Vive controllers to perform the annotation and the teleportation in the VR environment and two lighthouses to cover the annotation space and map the user's motion to the virtual environment. The point cloud scenes were chosen to present at least one instance of the following semantic classes: 4-wheelers (e.g., cars), 2-wheelers (e.g., cyclists), pedestrians and trees. We chose these classes of objects since they belonged to the most frequently appearing classes in datasets for urban scene understanding, such as Kitti, SemanticKitti, nuScenes, and cityscapes datasets [18], [1], [19], [20].

B. Study Design

We consider, a within-subject study with two independent variables: (1) *level of immersion* and (2) *visual aids*.

[2]Steam VR: https://store.steampowered.com/app/250820/SteamVR/
[3]HTC Vive: https://www.vive.com/de/product/vive-pro-full-kit/

Fig. 2. Levels of immersion: (a) Non-immersive setup uses a 2D Monitor to interact with the scene. (b) Semi-immersive setup: the scene is visualized on a large 85" screen. (c) Fully-immersive setup: all 6 degrees of freedom are incorporated.

1) Levels of Immersion: We explore three levels of immersion: (IM-1) non-immersive, (IM-2) semi-immersive, and (IM-3) fully-immersive. The setups used in the three conditions are depicted in Figure 2.

IM-1 In the non-immersive level, the 3D point cloud scenes were visualized on a 27" 2D monitor as in [10], [21], [22], and the participants were sitting on a fixed chair with controllers in their hands during the annotation process.

IM-2 Under the semi-immersive condition, the 3D point cloud scene was visualized using a 85" monitor with 4K resolution. During the study, participants stood at a distance of 75 cm from the monitor. For our case, this distance provided a trade-off between an enhanced sensation of immersion and the possibility to clearly visualize the full 3D point cloud scene. The present semi-immersive setup was not concretely used in previous annotation platforms. However, similar setups were used for other purposes, such as virtual stress management training for soldiers [23]. The semi-immersive setup represents a compromise when a higher level of immersion is needed and an HMD is not available.

IM-3 In the fully-immersive level, we use a head-mounted display (HMD)and provide a total of 6 degrees of freedom to exactly map the motion of participants in the scene. Participants used zooming capabilities to explore the scene at different scales. The use of HMDs represents the standard when it comes to fully immersive setups [24], [8].

2) Visual Aids: We investigate three types of visual aids.

VA-1 With the *distance-based coloring,* the surrounding data points were colored based on their distance to the center of the scene, according to a predetermined color palette. In VA-1, i.e, in the distance-based coloring of the points, we aim to convey information about the position of the annotator relative to the center of the scene, which provides the scale of the visualized objects and the remaining distance to reach a particular location in the scene.

VA-2 In the *region-based coloring,* we split the 3D point cloud scene into regions with a radius of four meters each. Each region is assigned a unique color and does not account for similarities of objects within them. The intention behind the region-based coloring is to give the annotator information about points that are physically located near each other and form a reduced spatial context for a better concentration.

VA-3 Finally, for the *object-based* highlighting, we employed bounding boxes around objects in the scene to provide precise semantic indications for objects of interest. To generate the object proposals, we use a pre-trained instance of a PointPillars neural network [25]. The user is then asked to adjust, confirm or refuse the proposals generated by the network.

During the study, we only provide the Vive controllers for participants to perform the annotation, and avoid the use of a mouse and a keyboard, especially with non-immersive setup. The reason is to avoid the effects of the interaction devices that might bias the annotation performance during the study.

3) Experimental Conditions and Task: To create experimental conditions, we combined all three levels of immersion with 3 visual aids (3 x 3 = 9 conditions). For each condition, we selected nine unique scenes. The orders of the conditions and the scenes were counterbalanced using a Balanced Latin Square. The study was conducted with 20 participants (16 male and 4 female) aged between 18 and 32 ($M = 25.4, SD = 3.137$). The Participants' task was to annotate one object in the environment, as quickly as possible, by constructing a bounding box using VR controllers. Figure 3 provides an overview of the steps needed to annotate an object in the scene.

C. Measures

To compare different levels of immersion with visual aids, we measured the following dependent variables:

Task Completion Time (TCT) We define the task completion time as the time taken by the participants to recognize an object of a particular semantic class and to fully annotate it.

Intersection over Union (IoU) We use the Intersection over Union as defined in [26] to measure the accuracy of the annotations. Intersection over Union is calculated as a ratio of correctly selected points within a bounding box over the total number of selected points.

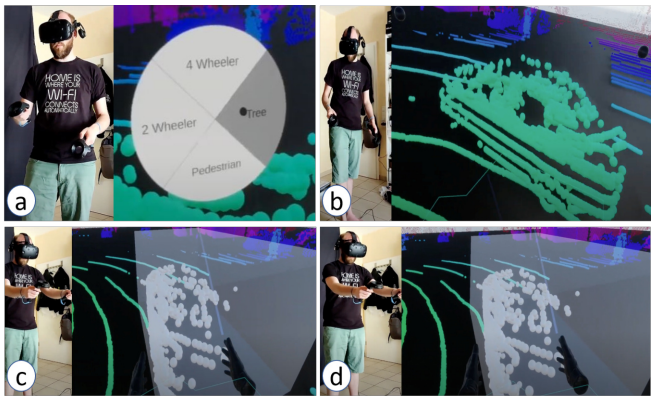

Fig. 3. The annotation process consisted of (a) semantic label selection, (b) placing four points to create a base of a bounding box, (c) auto-generation of a bounding box, and (d) a manual adjustment of a bounding box.

Mental load After each condition, participants were asked to estimate their mental load using a NASA Task Load Index (TLX) questionnaire [17].

Annotation success, convenience, and usability aAter each condition, participants were asked to estimate how successful they were in the annotation process, how convenient it was and which combination of level of immersion and visual aids they would like to use, using a 5-point Likert scale. For this, participants had to rate the following statements: *"I am confident I correctly annotated the objects"*, *"Interacting with the system was convenient"*, and *"I would like to use this combination of immersion and visual aids"*.

V. Results

The results of the experiment confirmed the effects of the level of immersion and the visual aids on the accuracy and efficiency of the annotation process. For the analysis of the results, we use two-way repeated-measures (RM) ANOVAs with the level of immersion and visual aids as the two independent factors. We tested the sphericity assumption using Mauchly's test. For the cases where the sphericity is violated, we used the Greenhouse-Geisser correction. We report eps whenever the sphericity is violated. Additionally, we used the Bonferroni corrected pairwise t-tests for post-hoc analyzes. Furthermore, we used the eta-squared η^2 to estimate the effect size and classify it as small, medium or large based on Cohen's classification [27]. We followed Searl et al. [28] and report the estimated marginal mean (EMM) to calculate the mean response of the independent factors. We analyze the non-continuous data of the Likert questionnaire using the aligned Rank Transformation [29]. We provide an overview of the results in Figure 4,where the mean and the standard errors for the TCT, IoU and TLX are represented.

A. Task Completion Time (TCT)

The analysis shows a significant influence of the level of immersion on the TCT with a large effect size ($F(1.37, 26.038) = 340.991, p < 0.001$, η^2 = 0.648). The post-hoc analysis backs up this finding. In fact, the TCT decreases significantly when the level of immersion increases (non-immersive: EMM μ = 1310.84 ms, σx = 140.994, semi-immersive: EMM μ = 540.583ms, σx = 100.104, fully-immersive: EMM μ = 440.126ms, σx = 60.784). The more the participant is immersed in the scene, the lower the TCT value is. Similarly, the visual aids proved to have a significant influence on the TCT with a large effect size as well ($F(1.465, 27.828) = 9.251, p = 0.002$, η^2 = 0.327). Post-hoc analysis confirmed lower TCTs for visual aids with finer context indication (distance-based: EMM μ = 920.830ms, σx = 140.104, region-based-based: EMM μ = 840.745ms, σx = 90.8, object-based: EMM μ = 520.318ms, σx = 60.945). The interaction effects between the level of immersion and the visual aids were not statistically significant ($F(1.465, 27.828) = 0.442, p = 0.642$, η^2 = 0.023).

B. Intersection over Union (IoU)

The analysis shows a significant influence of the level of immersion on the IoU with a large effect size ($F(1.607, 30.531) = 13.280, p < 0.001, \eta^2 = 0.411$). The post-hoc analysis backs up the latter observation. In general, the IoU increases when the level of immersion increases (non-immersive: EMM μ = 0.851, σx = 0.023, semi-immersive: EMM μ = 0.9, σx = 0.011, fully-immersive: EMM μ = 0.94, σx = 0.010). The visual aids proved to have a significant influence on the IoU with a large effect size as well ($F(1.646, 31.279) = 11.57, p < 0.001$, η^2 = 0.378). Post-hoc analysis confirmed higher IoUs for visual aids with finer context indication (distance-based: EMM μ = 0.878, σx = 0.014, region-based-based: EMM μ = 0.873, σx = 0.02, object-based: EMM μ = 0.941, σx = 0.009). The interaction effects between the level of immersion and the visual aids are significant, with a large effect size ($F(2.896, 55.03) = 18.235, p < 0.001$, η^2 = 0.490).

C. Mental load

The level of immersion had a significant influence on the TLX scores with large effect size ($F(2, 38) = 1038.539, p < 0.001$, η^2 = 0.982). This is confirmed by the post-hoc tests, where the non-immersive variant had significantly higher values (EMM μ = 56.833, σx = 0.417) compared to the semi-immersive (EMM μ = 35.483, σx = 0.823) and fully-immersive variants (EMM μ = 30.167, σx = 0.598). The analysis also showed a significant influence of the type of visual aids on the TLX values with large effect size ($F(1.541, 29.284) = 41.838, p < 0.001$, η^2 = 0.688). This is confirmed by post-hoc tests, where TLX scores for the distance-based visual aids (EMM μ = 42.133, σx = 0.677) are higher than in the region-based-based (EMM μ = 41.167, σx = 0.546) and object-based (EMM μ = 39.183, σx = 0.421) visual aids. The interaction effects between the two factors were not statistically significant ($F(2.683, 50.986) = 0.754$, p < 0.001, η^2 = 0.512).

D. Annotation success, convenience, and willingness to use

1) Annotation success: After each condition, we asked the participants how confident they are about the correctness and

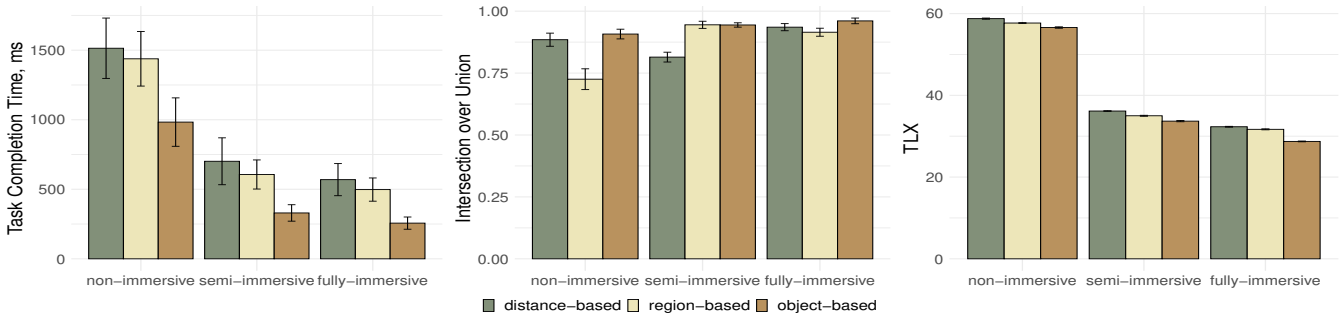

Fig. 4. Overview of results: means and standard errors for task completion time (a), Intersection over Union (b), and Task Load Index (TLX) (c).

fullness of their annotation. The analysis showed a significant effect of the level of immersion on the confidence of the participants ($F(2, 171)= 5.375$, $p<0.001$). This was confirmed by the post-hoc tests, showing more confidence in the semi-immersive ($p<0.001$) and fully immersive ($p<0.001$) set-ups compared to the non-immersive one. The visual aids we provided did not however show a significant influence on the confidence of the annotators ($F(2, 171)=0.646$, $p=0.525$) and the interaction effects between the level of immersion and the visual aids showed no significance ($F(4, 171)=0.318$, $p=0.866$).

2) Convenience: We asked the participants about how convenient they found the combination between the immersion level and the visual aids. The level of immersion showed a significant effect on the convenience aspect of the presented variants ($F(2, 171)=82.504$, $p<0.001$). Post-hoc tests showed a leaning towards semi-immersive and fully-immersive variants (both $p<0.001$) compared to the non-immersive variant. We also found a significant influence of the visual aids on the convenience ($F(2,171)=3,513$, $p=0,032$). Post-hoc tests revealed a higher approval for the object-based type of visual aid ($p<0,001$) compared to the distance-based and region-based-based variants. No interaction effects were found ($F(4, 171)=1.277$, $p>0.05$).

3) Willingness to Use: We further asked the participants how strong they agree to have this combination of level of immersion and visual aids. The analysis revealed a strong significance of the level of immersion on the participants' rantings ($F(2, 171)=58.188$, $p<0.001$). Post-hoc tests showed a significant leaning towards more immersion ($p<0.001$ for semi/fully immersive variants compared to the non-immersive variant). The visual aids showed a strong significant effect on the willingness to use ($F(2, 171)=14.580$, $p<0.001$). Post-hoc tests showed that the object-based variant had the highest approval rating compared to distance-based aids and region-based coloring($p < 0.01$). No significant interaction effects were detected ($F(4, 171)=0.294$, $p>0.05$).

VI. Discussion

In general, the level of immersion and visual aids show an impact on the annotation process. More specifically, higher levels of immersion and object-based visual aids lead to a faster annotation process and a lower mental load. In addition, we observe a significant interaction effect between both factors on the annotation accuracy.

A. Assisted Immersion for Annotation

Our analysis reveals a strong influence of the level of immersion on all the measured metrics. For the TCT, higher levels of immersion lead to lower values, implying a faster annotation process. Similarly, the TLX values decrease when the level of immersion increases, indicating a lower mental load. Consequently, a more immersive annotation experience leads to a fast and effortless annotation process. The results of the Likert questionnaire and the qualitative assessment confirm this observation. Participants reported significantly more comfort and convenience while interacting with the semi-immersive and fully-immersive environments. The observed speed and ease of annotation that come along with a higher level of immersion can be attributed to the intuitiveness that accompanies the perception and manipulation of 3D data in a fully immersive environment. Compared to the usual 2D monitors, projectors and HMDs facilitate a better perception of depth and scale, which assist the annotator in quickly understanding the scene and spotting the objects of interest.

As for the Intersection Over Union (IoU), the strong effects of the level of immersion were accompanied by significant interaction effects with the visual aids.Increasing the immersion level for a given visual aid does not necessarily lead to higher IoU values. Therefore, both independent variables need to be adjusted simultaneously to ensure the accuracy of the annotation, as it is vital for the future use of the dataset.

B. Influence of the Visual Aids

The type of visual aids significantly affected all the measures, except for the confidence about the correctness of the annotations. A closer look at the metrics reveals a strong influence of the visual aids on the TCT values. For a given level of immersion, TCT values decreased with visual aids providing more semantics about objects in the scene. For example, the region-based visual aid makes it easier for the annotator to spot constellations of points that might contain an object of interest than distance-based point coloring. This shifts the attention of the annotator to a particular region of the scene, which makes the recognition and annotation

task faster. In the case of the object-based visual aid, the annotator's task is further reduced to spotting the highlighted objects. Thus, the time needed to perceive particular objects in the scene decreases when the visual aids target specific zones or objects.

The analysis of the IoU values revealed lower values in the region-based variant, while the object-based variant had the best values. We attribute the IoU decrease in this variant to the fact that we cluster the points according to the distance that separates them. In this case, points from different objects and classes may be colored with the same color if they are in the same region in the scene. This can induce confusion about the physical boundaries of the objects of interest, leading the annotator to discard some points that originally belonged to the object of interest. We, furthermore, observed a strong influence of the level of immersion on the chosen visual aids and the resulting accuracy. Therefore, it is necessary to adjust the visual aids according to the level of immersion, to achieve a more accurate perception experience. The TLX results confirm our findings. The object-based option leads to the lowest mental workload, while distance- and region-based coloring methods lead to slightly higher mental load. The Likert questionnaire results have further indicated that more refined visual aids (particularly object-based) received better ratings regarding confidence, convenience, and willingness to use. Finally, our participants mentioned the effectiveness of "cooperating" with the pre-trained neural network.

C. Interaction Effects between Levels of Immersion and Visual Aids

The analysis revealed no interaction effects between the level of immersion and the visual aids on the TCT and TLX. When considered separately, we can say that higher levels of immersion lead to a faster and easier scene understanding process. The same applies to visual aids targeting smaller or more specific regions and objects. However, we observed interaction effects on the IoU and, consequently, on the accuracy of the perception. Interestingly, the region-level visual aids yielded worse IoU values, except for the semi-immersive variant. We attribute this to the confusion that the visual aids bring to the user if the level of immersion is not suitable for it. In our case, the combination of semi-immersive and region-based options offers the trade-off between immersion and augmented scene content. The TLX results confirm the observations. Participants reported more mental load for region-based coloring when combined with full immersion or no immersion. In the qualitative feedback, the users stated that it is sometimes more difficult to annotate the full object for certain colors in the non-immersive and fully immersive variants. In summary, despite the strong effects of visual aids and the level of immersion on annotation, when considered separately, the interaction between the two independent variables affected only IoU, and, consequently, the accuracy of the annotation.

The results of our user study shed light on a considerably important yet previously neglected factor for the design of immersive annotation environments. The analysis of the interaction effects between the level of immersion and visual aids shows that one cannot assume that keeping the same visual aids while varying the level of immersion will not have undesirable effects on the annotation accuracy however small they may be. Adapting the visual aids to the desired level of immersion is needed to achieve a higher annotation accuracy.

VII. Limitations and Future Work

We are convinced that the presented results and analysis provide valuable insights for the design of enhanced annotation environments. However, the implementation and evaluation methodology revealed some limitations, which we leverage to set directions for future works.

The annotation sessions were sufficiently long to measure the metrics of interest (TCT, IoU, TLX). However, several other metrics like VR fatigue could not be measured as they required longer annotation sessions. A future direction of this work would focus on adapting the level of immersion and the visual aids to the fatigue level of the annotator.

In our experiment, we limited the annotation process to be performed with only one single type of controllers (HTC Vive). Although our annotation environment supported several types of inputs, this choice was made to isolate the visual effects of levels of immersion and visual aids from those of the annotation devices. However, the HTC Vive controllers have many buttons on them as well as thumb pads for continuous input, which could be used in the future for various interaction techniques for point clouds.

As hardware tools for recording 3D point cloud scenes, such as Lidars, are getting less expensive, 3D point cloud scans are becoming more dense and incorporate other features like colors. Enlarging the scope of our investigation to cover these aspects is important to cope with the current technological progress.

VIII. Conclusion

In this paper, we present *PointCloudLab*, a software platform that facilitates the annotation under different three levels of immersion combined with three different visual aids with increasing contextual granularity. The results of the controlled experiment confirmed the positive impact of immersion and object-based visual aids on the time and workload of the perception process. Furthermore, the results indicate that for more accurate perception experience, it is necessary to adjust the visual aids to the levels of immersion to reduce the visual confusions induced by the interaction effects of both factors.

ACKNOWLEDGMENT

This work has been funded by National Research Center for Applied Cybersecurity ATHENE and LOEWE initiative (Hesse, Germany) within the emergenCITY center.

REFERENCES

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9297–9307.

[2] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2624–2632.

[3] H. Jiang, G. Larsson, M. M. G. Shakhnarovich, and E. Learned-Miller, "Self-supervised relative depth learning for urban scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–35.

[4] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4013–4022.

[5] A. Nguyen and B. Le, "3d point cloud segmentation: A survey," in *2013 6th IEEE conference on robotics, automation and mechatronics (RAM)*. IEEE, 2013, pp. 225–230.

[6] Mathworks, "Ground truth labeler," https://de.mathworks.com/help/driving/ug/get-started-with-the-ground-truth-labeler.html, 2021, [Online; last-accessed January 2022].

[7] F. Wirth, J. Quehl, J. Ota, and C. Stiller, "Pointatme: efficient 3d point cloud labeling in virtual reality," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1693–1698.

[8] J. D. Stets, Y. Sun, W. Corning, and S. W. Greenwald, "Visualization and labeling of point clouds in virtual reality," in *SIGGRAPH Asia 2017 Posters*, 2017, pp. 1–2.

[9] E. Chang, H. T. Kim, and B. Yoo, "Virtual reality sickness: a review of causes and measurements," *International Journal of Human–Computer Interaction*, vol. 36, no. 17, pp. 1658–1682, 2020.

[10] W. Zimmer, A. Rangesh, and M. Trivedi, "3d bat: A semi-automatic, web-based 3d annotation toolbox for full-surround, multi-modal data streams," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1816–1821.

[11] M. Veit and A. Capobianco, "Go'then'tag: A 3-d point cloud annotation technique," in *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2014, pp. 193–194.

[12] F. Bacim, M. Nabiyouni, and D. A. Bowman, "Slice-n-swipe: A free-hand gesture user interface for 3d point cloud annotation," in *2014 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 2014, pp. 185–186.

[13] P. Z. Ramirez, C. Paternesi, L. De Luigi, L. Lella, D. De Gregorio, and L. Di Stefano, "Shooting labels: 3d semantic labeling by virtual reality," in *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2020, pp. 99–106.

[14] A. Doula, T. Güdelhöfer, A. Mativiienko, M. Mühlhäuser, and A. Sanchez Guinea, "Immersive-labeler: Immersive annotation of large-scale 3d point clouds in virtual reality," in *ACM SIGGRAPH 2022 Posters*, ser. SIGGRAPH '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3532719.3543249

[15] J. Valentin, V. Vineet, M.-M. Cheng, D. Kim, J. Shotton, P. Kohli, M. Nießner, A. Criminisi, S. Izadi, and P. Torr, "Semanticpaint: Interactive 3d labeling and learning at your fingertips," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 5, pp. 1–17, 2015.

[16] M. Slater, B. Lotto, M. M. Arnold, and M. V. Sánchez-Vives, "How we experience immersive virtual environments: the concept of presence and its measurement," *Anuario de Psicología, 2009, vol. 40, p. 193-210*, 2009.

[17] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[18] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[21] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[22] Y.-S. Wong, H.-K. Chu, and N. J. Mitra, "Smartannotator an interactive tool for annotating indoor rgbd images," *Computer Graphics Forum*, vol. 34, no. 2, pp. 447–457, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12574

[23] S. Bouchard, F. Bernier, É. Boivin, T. Guitard, M. Laforest, S. Dumoulin, and G. Robillard, "Modes of immersion and stress induced by commercial (off-the-shelf) 3d games," *The Journal of Defense Modeling and Simulation*, vol. 11, no. 4, pp. 339–352, 2014.

[24] J. D. Stets, Y. Sun, W. Corning, and S. W. Greenwald, "Visualization and labeling of point clouds in virtual reality," in *SIGGRAPH Asia 2017 Posters*, 2017, pp. 1–2.

[25] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[26] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.

[27] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[28] S. R. Searle, F. M. Speed, and G. A. Milliken, "Population marginal means in the linear model: an alternative to least squares means," *The American Statistician*, vol. 34, no. 4, pp. 216–221, 1980.

[29] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 143–146.