Response to past reviews

We thank all reviewers for their feedback. We feel most of the comments are about minor issues and ask us to clarify certain aspects, and we feel confident we have addressed all of those in our revision. In addition, there are some larger points that we outline below.

GENERALIZABILITY (ACs, R1)
We chose one specific scenario (plant cell division) as the entity of our study, and this question of selection in dense environments is important in many domains (e.g., assembly of mechanical parts in complex machines, multiple neurons exploration, and design integrated circuits). While it may appear that some properties of our dataset are unique, we think that other datasets from other domains have similar properties. For example, parts in complex machines or the cells in brain connectomics data are similarly tightly packed as cells in a plant embryo. So the tight packing of cells in the plant embryos that we studied is a good example of datasets in which selections of individual objects have to be achieved, despite the density. In addition, traversal tasks are common for situations that require finding relative relationships between pairs of objects in a given domain. For example, when assembling tiny parts of a machine, people need to check every component and install step by step or understand the assembly of sub-components. To better make this point of generalizability, we added a respective discussion in Section 5.3.

CONTRIBUTION (ACs)
Our most significant contribution is presenting implementations about selection interaction with dense 3D objects. We are the first one to investigate selection of densely packed datasets with traditional input modalities. Also, we contributed a separate investigation of discovery and traversal tasks, which provides insights on combining multiple advantageous techniques. Designers can use our results implementing new interaction methods for similar situations. Visualization researchers can use the take-aways from our paper in creating future exploration tools for densely packed datasets. We better make these points now by having rephrased the contributions description in our introduction.

RELATED WORK (ACs)
We agree with the reviewers that we may have used ambiguous terms which could cause confusion in describing literature such as “selection.” We now better clarify some work targeting “interaction” including “selection” rather than just “selection” in the Related Work section. For example, we refer to past work [36], [6], and [48] to explain that we chose traditional input devices like mouse and keyboard, instead of techniques for VR. To make the literature structure clearer, we re-organized the statements and the layout of Section 2.

STUDY DESIGN (ACs)
To fully explore the cell assignment process, we designed our whole study with experts, which guaranteed that biologists could use these techniques in completing the division. Nonetheless, the interaction techniques are general enough such that our layperson participant pool allowed us to study the techniques sufficiently and also in a generalizable way. We also better explain the choice of participants, see below.

- Why these techniques (R2)
We chose those three techniques based on literature and related softwares biologists are currently using. We did not foresee that participants would convert the 3D exploration tasks into 2D

searching tasks in List Selection, and that is the reason we did not include Explosion Selection with visual cues for traversing(R1). We now better explain the reasons in Section 3.

- List Technique design (AC2, R2)

We specifically did not include any property value in the list because experts do not make the decision on sister cells only based on a numeric value. Instead, they always evaluate the shapes and neighborhood of cells in a 3D view. We thus replicated this fact by means of our proxy number. We now better make this point by adding descriptions about List Selection in Section 3.1.

- Why these participants (AC2, R2)

We designed our study specifically to solve the general question of selection in dense environments (see above), and only used the plant biology use case as our example. We thus recruited non-expert participants because they could provide this general feedback, without being limited in the cell division scenario. Also, in other scenarios besides cell division, target users will come from diverse backgrounds. We still plan to conduct studies with experts in the future to see their performance with our techniques. We now better explain this in Section 3.6.

- Simulating shape comparisons with validation metric (R2)

The number we picked is not derived from the data but determined randomly, specifically because we were interested in the generic question of selection in dense environments (and did not want to limit our insights to cell development). We used a randomly chosen number only as a proxy for some selection criterion in a given domain. We now better explain this fact by adding descriptions in Section 3.2.

- Apparatus effects (R2)

Due to the COVID restrictions, we were required by our institutional review board to conduct the study remotely. Nonetheless, we minimized the effects by checking in advance whether participants could smoothly conduct the experiment with their own devices. We did not find effects from their feedback. We better explain this point in Section 3.7.

RESULTS ANALYSIS (ACs)

We used effect size rather than NHST because the latter would result in dichotomous decisions, while the actual difference of 0.049 and 0.051 could be minor. Given the fact that the Combination tasks were always presented at the end, we analyzed participants' performance of Combination independently. We indirectly analyzed the learning effect by calculating the ratios of clicking times in lists and 3D view in Combination tasks. The ratio indicates which technique they think is more intuitive to use. This analysis is common practice as we pointed out via references in Section 4.

DISCUSS (AC2)

- What is "1-neighbourhood"

We used this term to indicate cells which directly connected with the target cell, that means they are direct neighbors. We now add the clarification in Section 3.3.

MINOR

- We defined "sisters" in the Introduction. (AC2)
- Edit the size of the figures. (AC2)
- "Access" was referred from [20]. (AC2)
- Make the techniques term consistent. (AC2)
- Figure captions displacement. (AC2)
- Discuss other 3D environments. (R1)

Reviews Details

1AC review
score 2/5

Expertise
Knowledgeable
Recommendation
Possibly Reject: The submission is weak and probably shouldn't be accepted, but there is some chance it should get in; 2.0

1AC: The Meta-Review

SUMMARY

This paper compares several selection techniques for tightly packed 3D objects. The techniques compared included "list", "explosion", and "hybrid" (i.e. list + explosion). The comparison was done on the task of cell division history marking as required by plant biologists who study
the development of embryos. The biologists need to mark cells as "sisters" according to their shape. In this scenario, the 3D objects (cells) are tightly packed, and occlusions are common. The user evaluation experiments indicated that the use of list was faster and more accurate than the explosion technique.
Overall, the participants preferred the hybrid approach.

STRENGTHS

The major strengths of the approach are summarized as follows:

- the paper is well-written, easy to follow, and is well organized (2AC, reviewer 1, reviewer 2). It is an enjoyable read. The video demo is nice; adding voice-over would make it more convenient to view (1AC)

- the submission explores an interesting problem. The motivation is clear. (2AC, reviewer 1)

- the experimental setup is solid and the analysis is valid (reviewer 1)

- the project follows an open science approach. All the experiment results are available (reviewer 2)

WEAKNESSES

The major weaknesses of the approach are summarized as follows:

1) Generalizability Concerns. It is unclear whether the findings can be generalized to the target population (expert users) and to HCI in general (2AC, reviewer 1).

It is uncertain how HCI/visualization researchers can generalize and apply the findings to other HCI problems because the experiment setup and tasks studied in this paper are quite unique to cell division history annotation (reviewer 1). More discussion on extension and generalizability is recommended.

The authors should include other dense 3D environments and expand the traversal task to gain more insights on the generalizability of the approach (reviewer 1).

2) Analysis Concerns. The hybrid approach was not analyzed in comparison to the list and explosion techniques (2AC). It is difficult to follow section 4 (results) and section 5 (discussion) because each factor is analyzed independently, not together (2AC).

Also, some of the claims made do not seem to be entirely supported by all of the data (2AC).

Some statements (e.g., 3D techniques were best for truly 3D tasks) could be misleading and need to be tempered to reflect the specific conditions evaluated (e.g., the properties of the target) that lead to the findings (2AC)

3) Related Work. The cited works in Section 2.1 do not refer to the selection of targets in dense datasets and are not very related to the paper (2AC).

4) User Evaluation Settings.

It is unclear why the authors focused on the 3 interaction techniques (list, explosion, hybrid) and whether these choices of techniques are valid with respect to satisfying expert users' needs. (reviewer 2)

On the other hand, the user study was run on non-expert participants, not expert users. It is unclear if the findings from the study can be transferred to predict the results of expert users. It is recommended that the authors evaluated their approach with a few expert participants at least (reviewer 2).

The average score for this paper is negative. The authors may provide a rebuttal to address any misunderstanding or factual errors in the reviews and to respond to the main weaknesses summarized above.

Rebuttal response

(blank)

2AC review
score 2/5

Expertise

Knowledgeable

Recommendation

Possibly Reject: The submission is weak and probably shouldn't be accepted, but there is some chance it should get in; 2.0

Review

This submission explores the usability of three selection techniques (list, explosion, list+explosion) within environments where possible sections are densely packed (e.g., botany cell development reconstruction). The results seem to indicate that while the use of lists was faster and more accurate than the explosion technique, however participants preferred the combination technique.

This submission was an interesting read as I am unfamiliar with the domain of botany. I found that the submission was reasonable easy to read and the submission was well organized. I do think it was submitted to the wrong sub-committee because EICS is typically focused on the engineering of systems rather than a user study about the evaluation of interaction techniques.

I am not an expert in estimation techniques for data analysis so I am not 100% sure if all of the analysis is correct so I will defer to the other reviewers (or possibly a 3AC) to ensure that the analysis was conducted correctly. While estimation analysis becoming more common, it is not as popular as NHST so it would be useful to clarify which effect sizes were presented (i.e., "simple effect sizes"). Although this submission has three main metrics (and then questionnaire data) I found that it was quite difficult to read section 4 and 5 and come to a concrete outcome because each factor is analyzed independently rather than together, so no claims can be made about potential interactions between technique, task subtype (discovery/traversal), and target type (inside/outside)(e.g., perhaps outside cells when performing discovery with a list end up being slower than inside cells while performing discovery with explosion?). I also found the analysis of the combined technique to be disappointing because there really wasn't any comparison to the list and explosion techniques. I know that it was always presented at the end but learning effects could have been analyzed. The current comparison simply divides the technique into distinct sections, but this really seems inappropriate because there is an interleaved element here that is not captured (that the participant comments / questionnaire seems to allude to).

Aside from this, after reading the entire submission, I was confused as to the generalizability of the findings to the target population and to HCI in general. The submission noted that experts make use of the shape and size of targets and the surface they share. The results suggest that a list-based method was faster and more accurate (albeit for specific conditions): "We found evidence that list led to more efficient (faster and fewer clicks) and more precise input than explosion overall". The list used in the study appears to just use text that (i) doesn't visually show these properties and (ii) doesn't textually encode these properties. So there is a mismatch here between the skills of the target audience and the utility of the findings. The submission then presented claims that generalized the results from this task (which normally requires expert knowledge and techniques) to all selection tasks (e.g., Section 5.3 and 6; "3D interaction techniques work best for truly three-dimensional tasks", "We have advanced our understanding of interaction techniques for the selection of objects in dense 3D environments"). I would argue that the results didn't really show that 3D techniques were best for "truly three-dimensional tasks" because the entire task was a 3D task and the results showed that the 2D list was faster and more accurate, but there are caveats to this depending on the type of target (inside vs outside). These statements are currently misleading and need to be tempered to reflect the specific conditions that were evaluated and influenced the findings.

I also found the related work to be a bit confusing because some of the references in Section 2.1 don't actually refer to the selection of targets in dense datasets (Section 2,1) nor do they seem to be related to the present paper. For example [6] and [45] weren't selection tasks but rather tasks requiring the placement of objects at certain locations or the movement of a controller to match a specific orientation. References [33] and [47] are report on the results of user-defined gesture experiments and don't involve selection at all, nor did they compare different input methods. It also seems like the last paragraph of 2.2 should be moved into section 2.1 because it actually describes selection techniques (no idea why the word "access" is used). Within 2.2, I was confused as to why some of these techniques were discounted because there wasn't space between the targets. It seems like RayCursor (and other techniques) would work and could highlight occluded targets, regardless of the space between them. I was also surprised that BubbleRay (Lu, Y., Yu, C., & Shi, Y. (2020, March). Investigating Bubble Mechanism for Ray-Casting to Improve 3D Target Acquisition in Virtual Reality. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 35-43). IEEE.) and Bubble Gaze were not mentioned (Myungguen Choi, Daisuke Sakamoto, and Tetsuo Ono. 2020. Bubble Gaze Cursor + Bubble Gaze Lens: Applying Area Cursor Technique to Eye-Gaze Interface. In ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers). Association for Computing Machinery, New York, NY, USA, Article 11, 1–10. DOI:https://doi.org/10.1145/3379155.3391322) as I could imagine that they could work as well. As written, I was unconvinced about the arguments against existing techniques (and implicitly the need to not evaluate them).

In summary, this submission does explore an interesting domain, however I was not convinced that the results generalize to the target population (and thus make a contribution), there is a lack of analysis between the various factors and all three conditions, and some of the claims made don't seem to be entirely supported by all of the data. As such, I am reluctant to recommend this submission for acceptance at this time.

Other:

- "sisters" should be defined
- Both occluded/surface and inner/outer were used to describe the different cell locations, which got confusing
- The study itself is not a contribution, but rather the findings from the study.
- I didn't understand what the discussion of selection techniques (contribution section in the Intro) referred to and was uncertain as to what contribution could follow from this, unless, for example, a new classification of the techniques arose.
- All figure captions should appear under the figure, otherwise the reader

is looking around to find the figure description.

- The violin plots are also difficult to read because the confidence intervals are so small and hard to see. Removing the segments of the plots that aren't ever evaluated would allow for the remaining plots to be larger (e.g., Figure 11 a; Figure 13 b and c, etc.)
- What is '1-neighbourhood'?
- "neighbourhoodin" -> "neighbourhood in"
- "though people select objects easier due to the increased depth perception due to stereoscopic projection," this is a cumbersome phrase that is difficult to understand

Rebuttal response

(blank)

----

reviewer 1 review
score 1.5/5

Expertise

Knowledgeable

Recommendation

. . . Between reject and possibly reject; 1.5

Review

The authors present a controlled user study on the comparison of three selection techniques for dense 3D visualization that informs the design of an annotation tool for plant-based cell embryos.

Overall, the paper is well-written, the motivation is clear, and the presentation (in particular the violin plots) is very well executed. The experimental setup seems to be solid (which I acknowledge can be challenging during these times) and the analysis appears to be correct. Hence, I strongly believe that the presented results are true. Although the video figure could be improved to show an actual walk through of a trial from the experiment, I appreciate the authors submitted a video as it clarifies the experimental setup and tasks.

My main concern with the paper is twofold:

First, while the use case is well-motivated, I do not see how the findings generalize and what HCI/visualization researchers can learn in general. I do not agree that "We have advanced our understanding of interaction techniques for the selection of objects in dense 3D environments.". As a reader, I have learned that the 3D explosion technique is more suitable for tasks that require an understanding of the 3D view while the abstract list interface is useful for an abstract search task. However, it's unclear if the abstract list interface would still be useful if the 3D view would actually inform the traversal task. Moreover, I do not believe that the findings generalize to other use cases as the setup and tasks are unique to cell division history annotation. For instance, the explosion view appears to work great in the given scenario as the embryo and cells all have a roughly equal shape and size. In other dense 3D views, as found, for example, in connectomics or 3D reconstructions of indoor environments, the shapes, sizes, and tasks can be entirely different and it's unclear how the explosion technique would perform.

Second, I am concerned that the experiment does not actually inform us well about the performance of selecting 3D objects in a dense 3D space as it appears that the cell morphology does not correlate with the "likelihood" score. As the authors state, the traversal task was converted into an abstract search task where it does not seem to be necessary to inspect the cell shape to complete the task. However, I believe this to be in contrast to how biologists would find sister cells. While I understand and acknowledge why the authors introduced this task abstraction (the "likelihood" score), I would have liked to see some correlation between the 3D shape and the likelihood such that the 3D shapes provide useful visual cues for the traversal task. As it is right now, it's unclear whether the list view would truly perform better in a real scenario.

In summary, unfortunately, I do not see how that the presented results are interesting and impactful enough to the broader HCI and visualization community to justify a presentation at CHI. I would suggest that the authors include other dense 3D environments and expand the traversal task such that it covers other use cases as well.

Rebuttal response

(blank)

----

reviewer 2 review
score 2.5/5

Expertise

Passing Knowledge

Recommendation

. . . Between possibly reject and neutral; 2.5

Review

This paper presents an empirical comparison of three interactions techniques to visualize and select 3D densely packed objects. These techniques rely either on a 1D list of tags, 3D models densely packed, or a mix of the two with the possibility to space out the 3D models. The purpose of these techniques is to help plant biologist experts to deal with cell division history marking, a very specific task in which the experts need to mark cells as “sisters” based on their shape. The study results show that the list technique was preferred by the participants and led to better user performance than the two others.

First, I would like to point out the open science approach of the authors. All the results of the experiments are already available and the authors pre-registered the experiment. These are methods we need to see more often in the HCI community.

The paper is well written, which makes the reading very smooth and enjoyable. Overall, I have two major concerns about the current version of the paper. First, on the set of participants used for the experiment and the metric used to recognize cells as being sisters, and second, on the comparability of the techniques used.

Set of participants and validation metric

The authors present the study as a comparison of visualization techniques that can benefit plant biologist experts. Yet, they decided to run the study on non-expert participants using a number to represent the alikeness of cells. Although I guess the reasons behind this choice, i.e., more statistical power and better availability of the participants, I think this created a bias in the experiment that sheds doubt on the transferability of the results to expert users. As the authors highlighted, one of the primary advantages of the list technique is to be simple to navigate using the tag labels. Doing so, however, the participants do not seem to look at the cells’ shapes in the process. To my understanding, an expert, in contrast, must focus exclusively on the shapes of the cells to find “sisters”. Therefore, the visual tasks are different, and I wonder whether the list technique is applicable in a real scenario and whether experts would really benefit from it. The authors could have focused on a set of a few expert participants, in which case the transferability of the results is less questionable.

In other words, how confident are the authors of the transferability of the results to a real scenario? I did not see the authors discussing this aspect in the paper. Moreover, how relevant is the validation metric (a number) compared to shape comparisons used in a real scenario?

Comparability of the techniques

As pointed out above, the list technique seems inadequate for the task at hand. It is using labels instead of 3D models to represent the cells (my understanding is that users can still see the 3D models while using the list, but they prefer navigating the labels), while expert users would compare shapes of cells, thus would rely mostly on shape recognition. The 3D techniques, in contrast, are prone to occlusion, which makes the task harder. Being able to space out the 3D models lightens this problem, so it is not surprising to find better performances in this case. Although the list technique provides a solution to the occlusion problem, it also hides the shapes of the cells, thus making the task harder for expert users in a real scenario. Overall, it is a bit unclear why the authors decided to focus on these three interaction techniques and how their choice is valid considering expert users' needs.

Did the authors design interaction techniques to solely minimize the occlusion problem? Some clarification would help the reader better understand the technique comparisons.

Other remarks

- Apparatus was different for almost all participants due to the remote setup. How can the authors ensure this had no impact on the study results?

- It is unclear if the names of the cells are important for completing the task. Would an expert user rely on these names, hence benefit from the list in other ways than explained in the text?