Dear Editor,

We thank you and the reviewers for the careful evaluation of our manuscript and for the constructive and encouraging feedback. All comments were well taken, and we have revised our manuscript accordingly.

The major changes are as follows.

- We strengthened the quantitative evaluation by adding Average Precision (AP) metrics to better capture the performance differences observed in the synthetic experiments, particularly the sensitivity to shadows and complex color patterns. The tables have been revised accordingly.

- To address concerns regarding generalization, we expanded the experiments to include additional real-world datasets. We now include new analyses using the publicly available ImageGrains and SediNet datasets. While these datasets do not provide complete or fully reliable ground-truth labels, we added new figures and tables in the Supplementary Materials that compare predictions, size distributions, and object counts, and discuss the associated limitations. OrthoSAM predictions (binary images) have been added to the GitHub archive.

- We also revised and expanded the discussion of related approaches, in particular Segmenteverygrain, to more clearly explain conceptual differences, strengths, and limitations of the respective workflows.

- In response to reviewer comments, we improved the software packaging and installation process, added documentation and tutorials, and reported representative processing times for different image sizes and hardware configurations.

Please find below the detailed responses. For clarity, we would like to note that a section discussing recall and mean IoU based on ImageGrains in some of our earlier online responses has been removed in the finalized response. This section was included in the earlier responses by mistake, as these metrics are no longer applicable. We note that while preparing the responses, we have run several models and tested several reference datasets, and went through multiple iterations of the response letter.

We thank the reviewers for their support, and we believe that the manuscript has improved by addressing their comments. We are looking forward to hearing from you.

Sincerely,

For the three authors

Response to RC1

The authors present a novel method and proof-of-concept for pebble segmentation in orthoimages by adapting the popular and widely-used Segment Anything Model (SAM; Kirillov et al., 2023). They identify important, but often unaddressed, weaknesses of SAM, such as the reduced performance in dense segmentation tasks (where many instances of the same object class should be segmented), and its limited capability to segment objects from one class with a significant size variability. To test their approach, the authors use 1) synthetic images with circles as a proxy for pebbles and 2) ortho-mosaics of real pebbles created with handheld cameras and photogrammetric processing. In their experiment 1, they test for the effect of a variety of image perturbations on segmentation quality. Here, they find that particularly shadow effects have some negative impact on SAM's segmentation performance. In experiment 2, they apply their workflow to real-world images, showcasing the improvement of their multi-scale segmentation with SAM. In this scenario, they categorically evaluate segmentation performance through manual counting due to the lack of ground truth masks. Both experiments show that their approach is up to the task and has the potential to mitigate some of the segmentation shortcomings of SAM for such applications.

I find the method well-conceived and thought-through, the data rigorously tested and clearly reported, and the manuscript well structured. In particular, I consider the balance between technical details in the main manuscript and the appendices well struck, which makes the manuscript very readable, while not omitting relevant information. The presented results generally support the findings and conclusions. Here, I would only have two suggestions for calculating additional scores and using an additional image dataset to test the approach (see specific comments below), which might allow for a better evaluation of some aspects of the segmentation performance of SAM/OrthoSAM. However, these are just suggestions, not concerns raised. Currently, the manuscript has many small figures; maybe combining some figures into larger figures (e.g., Figures 10 and 11) would be helpful. Additionally, some minor/technical comments are included as in-line comments in the attached pdf.

In summary, I find the work of very high quality, with only a few minor points where the manuscript could be further improved. I suspect the authors will have no problems in addressing these points, and I look forward to seeing the manuscript published soon.

Kind regards,

David Mair (Uni Bern)

Thank you for your positive feedback and the helpful suggestions to improve the manuscript.

Specific comments:

Additional metrics for segmentation performance: The authors use well-established metrics to evaluate the segmentation performance. However, I would suggest additionally also calculating Average Precision (AP) scores where IoU thresholds are used (e.g., AP@0.5 IoU and/or mAP@0.5-0.9 IoU), as used in the SAM paper (Kirillov et al., 2023) or in general is widely used for instance segmentation tasks (e.g., Padilla et al., 2020). This is because I suspect that SAM segmentations are slightly worse for the colored synthetic images than for the black and white, while in both cases they score high in precision, recall, and mean IoU (see also lines 198-199). These scores could be calculated from the TP, FN, and FP values, where all TPs falling below a certain IoU threshold would count as FP. These scores might more clearly show that SAM is sensitive to shadows during segmentation (see also related comments in the pdf).

We have followed this suggestion and included the additional segmentation metric Average Precision (AP) scores. Due to the lack of confidence values in segment anything, we modified the calculation of AP to use object size as a proxy for confidence. We report AP@0.75 for both the B&W and Color with Shadow synthetic images. B&W [4,3000] has achieved an AP@0.75 of 1.00, while the Color with Shadow images have achieved an average of 0.87. For Ravi images, an overall AP of 0.97 was achieved. AP confirms that OrthoSAM's segmentation performance is slightly worse for the colored synthetic images, which aligns with the observation. We have revised the manuscript to reflect the additional metric and results.

Adding a dataset with instance labels for pebbles. In lines 99-100, it is stated that ideally the workflow should be tested on a dataset of several hundred to thousands of delineated pebbles. This is picked up in line 216, when it is correctly stated that no ground truth masks are available for the Ravi dataset, and hence no IoU scores can be calculated. Here, I would like to mention our S1 dataset (as part of the data used in Mair et al., 2024), which has > 2000 manually annotated pebble masks from orthomosaics (available here: https://zenodo.org/records/8005771). It would be interesting to see how OrthoSAM would perform here; I suspect it will perform very well, especially due to the grain size variability similar to that of the Ravi River. Using these data as an additional test could help to increase confidence in the performance of OrthoSAM.

Yes, we are aware of the limitations of a synthetic dataset, but also its benefits: a perfectly labeled dataset. In our preparation for the initial manuscript submission, we have performed segmentation for the SediNet and ImageGrains (mentioned by the reviewer) datasets. For the ImageGrains dataset, we noticed that the labels (i.e., the true ground-truth dataset) are only partially complete and do not include all pebbles. Our initial analysis and discussion among the authors of this manuscript suggested that we can not use their labeled data as a validation dataset. We have re-evaluated our initial assessment in the revisions and now include segmentation results and a comparison (see Table 1 and Figure 1 below). However, we refrain from treating these data as a true label dataset and do not report revised accuracy and precision statistics.

Instead, we compare the derived size distribution with a two-sided KS test (Figure 3). We note that according to the KS test, the distributions are not equal. We observe that OrthoSAM detects many more objects (cf. Table 1) and also many more small objects. Visual inspection reveals that OrthoSAM detects many of the smaller grains correctly (that are not in the ground-truth labeling dataset), but also identifies some objects that are not pebbles. These will need to be removed by additional filtering steps, for example through the normalized isoperimetric ratio (IRn). IRn provides a measurement for the roundness of an object, which can be used to remove irrelevant objects such as vegetation. OrthoSAM reports various statistics of segmented objects that can also be used for further filtering through more sophisticated outlier detection algorithms, such as decision trees. We note that some of the pebbles detected by OrthoSAM and not ImageGrains are low-contrast pebbles (i.e., the pebbles' color and texture are only slightly different from the background). We speculate that the gradient-based convolutional neural network in ImageGrains has not been extensively trained for low-gradient segments and thus excludes them.

We have extended the evaluation of OrthoSAM to include three images from the ImageGrains dataset (FH, K1, S1) and a selection of images from the SediNet dataset. As illustrated in Figure 1, OrthoSAM segments objects that are not pebbles or grains, and it also tends to detect finer objects. Due to its reduced sharpness and clarity, the FH image produces stronger noise in the results. The presence of very fine pebbles requires more input points, which further increases the likelihood of noise, making FH a particularly challenging case. This behavior likely explains the higher number of predicted objects and reflects the current limitations of the approach and the lack of ground-truthing data.

Table 1: Assessment of OrthoSAM prediction based on ImageGrains prediction.

	Image	ImageGrain	OrthoSAM	OrthoSAM IRn>0.7	OrthSAM (standard settings)	OrthSAM (standard settings) IRn>0.7
0	FH	1742	2819	2781	1927	1845
1	K1	1883	1358	1354	1113	1062
2	S1	4660	4316	4289	2915	2880

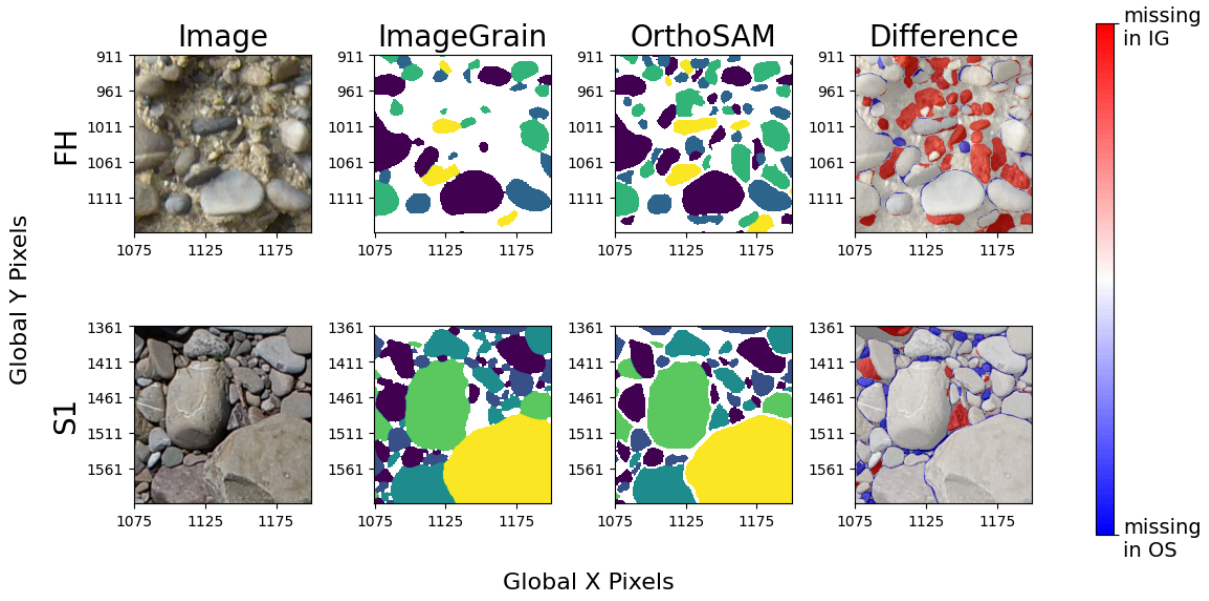

Figure 1: Comparison of OrthoSAM (OS) and ImageGrains (IG) predictions for center crops of image FH and S1. We note that the number of detected objects varies between the two models for each dataset. For example, for image S1, the OrthoSAM prediction identified 4316 objects, while ImageGrains identified 4660 objects. This ratio of 0.9 (4316/4660) is different for image FH (2819/1742=1.6). This does not allow a precision and accuracy assessment with the same training data. The difference plot visualizes the agreement and disagreement between two predictions. Red regions highlight areas where OrthoSAM identifies an object, and ImageGrains does not. While blue regions highlight areas where ImageGrains identifies an object, and OrthoSAM does not. Both examples demonstrated OrthoSAM's capability in fine object segmentation. However, due to the lack of a classification component, OrthoSAM has the inherent limitation that irrelevant objects may remain in the segmentation results. In particular, we see patches of sand that were falsely segmented in FH. Here, we see that lower resolution or blurriness in the image can exaggerate the issue, resulting in more false positives.

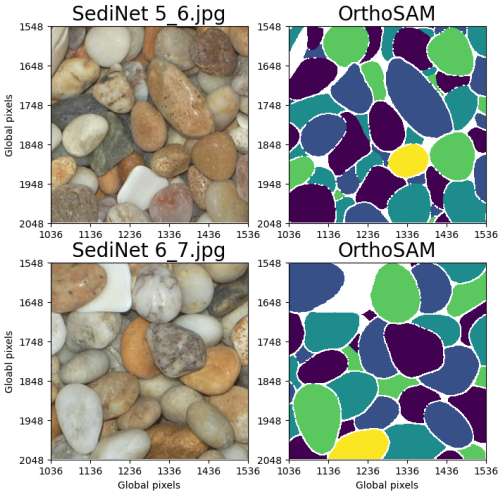

Figure 2. OrthoSAM segmentation of two SediNet images. A 500 x 500 pixel crop was taken from the lower-right corner of the full image.

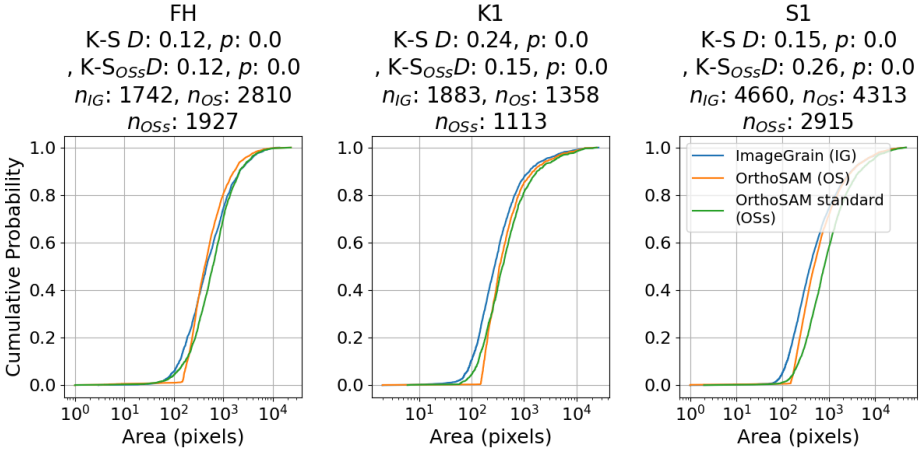

Figure 3. Cumulative size distribution of ImageGrains predictions and OrthoSAM predictions for images FH, K1, and S1. OrthoSAM predictions were made with two different settings: custom parameters for the respective image (OS) and standard parameters for large images (OSs). The number of identified objects varies between the methods, with OrthoSAM (OS and OSs) detecting more objects than ImageGrains (IG). A two-sided K-S test was performed to compare the similarity of the size distributions. For all images, the null hypothesis that the samples come from the same distribution was rejected ($p < 0.05$), suggesting that the segment size distributions produced by ImageGrains and OrthoSAM differ significantly. We partly explain this discrepancy by the different number of detected objects (more small objects detected by OrthoSAM).

Response to RC2

The manuscript by Chan et al. focuses on the description and validation of an open-source Python machine learning model called ‘OrthoSAM’, which relies on the Segment Anything Model (SAM) to generate instance segmentations of images of coarse-grained fluvial sediment. As someone who has also done some work on using SAM for grain segmentation, I think this is a promising approach and a having access to a variety of techniques and implementations at this stage are overall an advantage. The paper is well written and nicely illustrated, it includes a number of novel approaches that have not been implemented before, and the authors have clearly put a significant amount of thoughtful and careful work into the software and into validating the results with synthetic and field data. In addition, they have made the code open-source and available as a GitHub repository, which makes it a lot easier for these methods to be adopted and tested on other datasets.

Thank you for your thoughtful feedback and the helpful suggestions to improve the manuscript.

I do have a number of comments that I think should be addressed by the authors before publication; these are as follows.

The SAM-based approach and the tiling of large images are features of OrthoSAM that our Python module called ‘Segmenteverygrain’ also relies on. Although Segmenteverygrain is mentioned in the manuscript, I think there should be a bit more detailed discussion of what are the differences between the two techniques - not just the fact that OrthoSAM only relies on SAM, without the need for the U-Net pass, but also aspects like how broadly is the model applicable, how is it possible to improve the model outputs, is it possible to fine-tune the model. I do think that there is room for a variety of approaches to taking advantage of SAM (and of other similar) models in sedimentology and geomorphology, but it will be useful for the reader to get a brief a overview of the differences between the existing tools.

OrthoSAM is specifically designed as a workflow to assist SAM in delineating densely packed objects in large, high-resolution images. It achieves this by focusing on three main components: a tiling scheme, improved input point generation, and a multi-scale resampling scheme (resolution passes).

Segmenteverygrain, conversely, benefits from its initial U-Net pass because it restricts SAM's operation to areas already classified as grains, thereby effectively filtering out irrelevant objects. It might be a good idea to combine efforts in the future on this.

OrthoSAM segments all objects in an image and may delineate objects that are not pebbles. These will need to be removed by additional filtering steps or

manually. Segmenteverygrain instead will only segment pebbles that have been identified as pebbles using a convolutional neural network approach. This will ensure that only pebbles are delineated, but may also miss pebbles that were not initially detected by the neural network.

In the first version of the manuscript, we had a paragraph dedicated to Segmenteverygrain in the introduction. We have now elaborated on these points in the discussion in the revised manuscript and included additional points of OrthoSAM's performance on both ImageGrains and Sedinet datasets (see Figures 1 and 2 below). While precise metrics cannot be computed due to the lack of ground truth data, the results were visually assessed and are available through our GitHub repository. These examples demonstrate that OrthoSAM generalizes well across images with different grain characteristics and scene complexities.

One the novel aspects of the work presented by Chan et al. is the generation of synthetic data that is then used for validation. While I totally see the value of this in increasing the community's confidence in the model, one of the important questions about ML models is their ability to generalize. Although SAM has been trained on a wide variety of images and is good at generalization, I think it is less clear how well OrthoSAM would perform on real images of coarse-grained sediment that are quite different from the examples used in the paper. Although the authors are right that "manual validation is inevitably prone to subjectivity and human error, leading to potential biases and inconsistencies", I would argue that a carefully QC-d segmentation of real datasets is potentially more valuable for validating a machine learning model than a synthetic dataset that does not fully reproduce the complexity and variety of actual datasets. So I concur with the other reviewer that applying OrthoSAM to other datasets would be a valuable addition to the paper. It should not take too long to run it on some other publicly available datasets.

We agree that a high-quality training dataset for pebble segmentation will be useful for several machine-learning applications. However, these data do not (yet) exist, and it would be an important community effort to produce such a dataset - similar to reference datasets that have been generated for the lidar classification community.

We have used the synthetic training dataset to identify SAM's sensitivity to grain size and color. The identification of a lower detection size limit was achieved with the synthetic training dataset. We identified that SAM is not very sensitive to color variation and color noise - until a very high level of noise is added. These were important findings of the synthetic analysis, and real-world imagery provides additional challenges.

We elaborated in the revised manuscript on the validation of real-world datasets. We included additional discussion of OrthoSAM's performance on both ImageGrains and Sedinet datasets, although they do not contain a high-quality ground truth dataset. While precise metrics cannot be computed due to the lack of ground truth data, the results were visually assessed and are available through our GitHub repository. These examples demonstrate that OrthoSAM generalizes well across images with different grain characteristics and scene complexities.

We provide Jupyter Labs that guide through the processes of segmenting SediNet Images on our Github repository. This contains the parameters used for their segmentation in OrthoSAM. The parameters will need to be adjusted for different imagery, because grain packing varies. We note that the SediNet images are only somewhat useful in this regard, because many of the images are not scaled, and the sizes are only relative to the pixel areas.

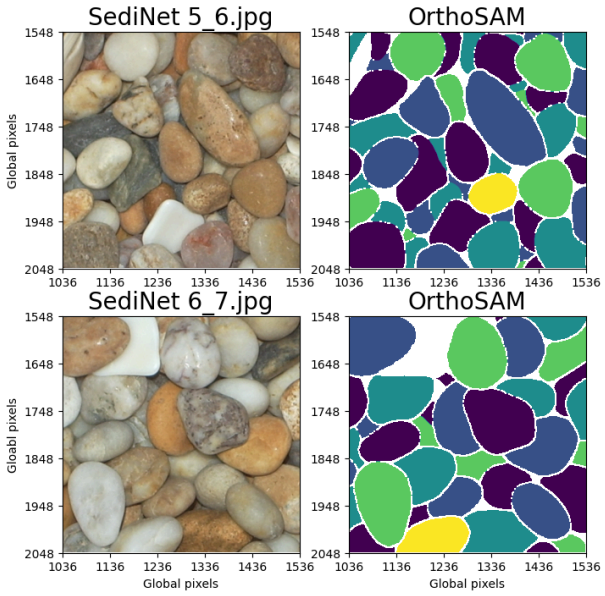

Figure 1: Example of an OrthoSAM segmentation of two SediNet images. We provide a few examples and their code and parameters for selected SediNet images on our GitHub repository.

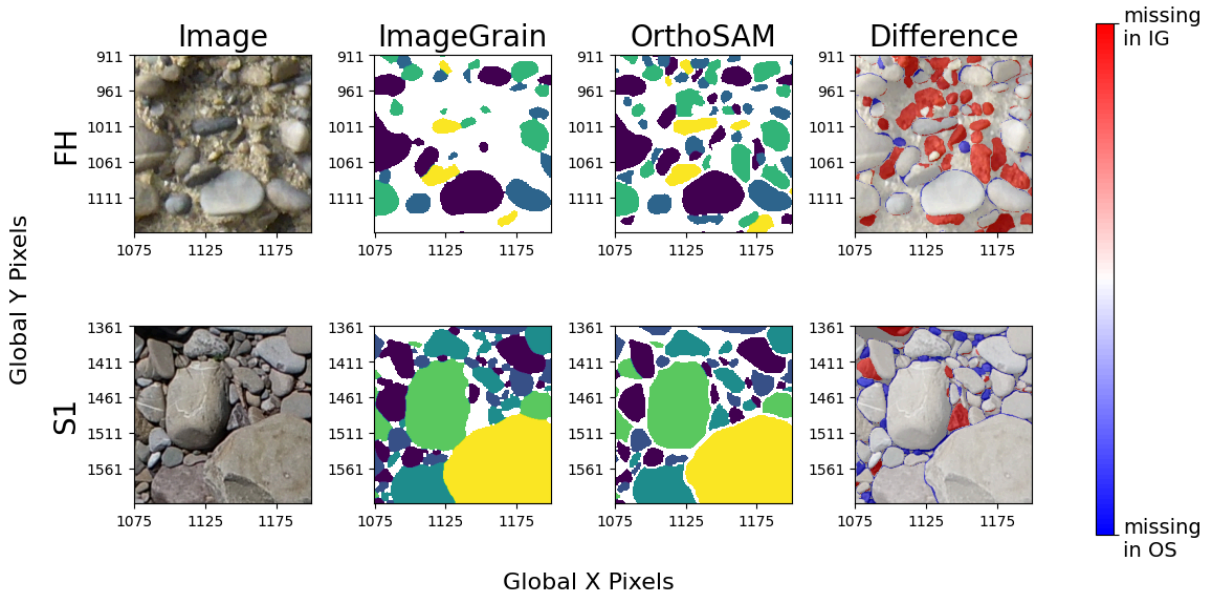

Figure 2: Comparison of OrthoSAM (OS) and ImageGrains (IG) predictions for center crops of image FH and S1. We note that the number of detected objects varies between the two models for each dataset. For example, for image S1, the OrthoSAM prediction identified 4316 objects, while ImageGrains identified 4660 objects. This ratio of 0.9 (4316/4660) is different from image FH (2819/1742=1.6). This does not allow a precision and accuracy assessment with the same training data. The difference plot visualizes the agreement and disagreement between two predictions. Red regions highlight areas where OrthoSAM identifies an object, and ImageGrains does not. While blue regions highlight areas where ImageGrains identifies an object, and OrthoSAM does not. Both examples demonstrated OrthoSAM's capability in fine object segmentation. However, due to the lack of a classification component, OrthoSAM has the inherent limitation that irrelevant objects may remain in the segmentation results. In particular, we see patches of sand that were falsely segmented in FH. Here, we see that lower resolution or blurriness in the image can exaggerate the issue, resulting in more false positives.

I do not think this is a major issue, certainly not for this manuscript, but: I have tried to install OrthoSAM on my computer and to run one of the notebooks but I gave up without getting to a result because I got a number of errors early on. Making it easier for a broad range of users to install and run the code will ensure a broader adoption of OrthoSAM.

We note the reviewer's comments and have modified the packaging and installation routine of our setup. OrthoSAM is now properly packaged and included in the requirements.txt. It can be installed using pip install -e . as well, provided that the repository directory is set as the working directory. Once

installed, OrthoSAM can be imported system-wide within the active virtual environment. These updates streamline the installation process and has improved the overall usability of the software.

We provide video material that guides through the installation and processing steps (in addition to the tutorials on the GitHub webpage).

We note that the processing can also be done within a Google Colab Environment, and we provide an example of this: https://www.youtube.com/watch?v=bLU6dbQ3vt0

An example of a video guiding through the analysis: https://www.youtube.com/watch?v=vu67RpeNHO4

The 'hardware requirements' section is quite useful, but it could be improved if typical compute times were added, e.g., how long does it take to create a segmentation result for an image with ~1000 grains? Is it possible/feasible at all to run the segmentation on a CPU?

In the revised manuscript, we have added duration for average processing times. The segmentation of a synthetic image with 10,000 × 10,000 pixels requires approximately 4 hours, whereas an image with 2,048 × 2,048 pixels requires about 5 minutes. Both were run on Quadro RTX 5000 GPU with 16 GB RAM and Intel Xeon W-1290 10-core processor. The processing of the similarly sized ImageGrains K1 image with 1,350 x 1,200 pixels took 10 minutes on the same system due to the use of an upscaled layer, whereas image S1 with 3,062 x 2,722 pixels took 50 minutes on the same system with the same settings.

We note that this may differ with different hardware setup. OrthoSAM is aimed at GPU processing, but will work with CPU processing (processing with CPU can be one order of magnitude longer).

We have included these numbers in the revised manuscript.

I hope the authors will find these comments / suggestions somewhat useful.

Sincerely,

Zoltan Sylvester

Citation: https://doi.org/10.5194/egusphere-2025-4003-RC2