

Dear Editor,

We thank the reviewer David Mair for their careful evaluation of our manuscript and for the constructive and encouraging feedback. All comments were well taken, and we have revised our manuscript accordingly. After careful consideration, we have decided to follow **option B** suggested by the reviewer. Specifically:

- We have removed the comparison with ImageGrains prediction in our manuscript, as the available manually annotated masks cover only a limited number of cropped patches. Using them would not allow a meaningful comparison with OrthoSAM, which operates on full images and addresses multi-scale object segmentation of large images.
- Accordingly, we have corrected the description of ImageGrains dataset and the related appendix figures.

We believe this approach addresses the reviewer's concerns. Please find below the detailed responses.

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

General Comments:

The authors present their revised manuscript for 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). In the revised version, they addressed all my comments from the previous review round (<https://doi.org/10.5194/egusphere-2025-4003-RC1>). They followed suggestions to include AP scores and to test their workflow with existing datasets.

However, it seems the authors did not use the recommended dataset, i.e., the manually annotated grain masks for some images in the S1 subset of the Imagegrains 1.0 (IG1 dataset, available in Mair, 2023), but instead compared their results to predictions from our workflow in Mair et al. (2024). In the newly added text (lines 287ff in the manuscript with track changes) and in Figures A10 and A11, conflicting statements are presented on whether labels or predictions were used, and the labels from the IG1 data are mischaracterized. This needs to be corrected. In particular, 1) it is unclear if and why predictions instead of labels from the ImageGrains data/workflow were used (see below). I suspect there might have been a misunderstanding about which images were explicitly meant or about the nature of the labels (see details below). Related, there are questions, 2) if using a type of images (from subset FH) is helpful, and 3) if Fig. A11 is correctly described (please see below for details). The idea behind my recommendation was to strengthen OrthoSAM's appeal by demonstrating that its predictions would be close to ground-truth annotations for images similar to those from the Ravi River (i.e., the S1 image tiles of IG1). Currently, the above-mentioned additions to the manuscript do not do that. To address this, I see, broadly speaking, three options for going forward for using (or not using) the IG1 data:

A) Keeping text and Figures A10, A11 as they are, but ensure to compare to the actual ImageGrains labels (i.e., the manually annotated masks from Mair, 2023; see below for details) for some selected S1 and/or K1 image tiles (K1 images would be part of the APF subset).

B) Not using the ImageGrains labels. In this case, the statement in lines 287ff needs to be corrected because predictions cannot be used as a substitute for labels (see below). Consequently, omit the comparison with the ImageGrains predictions, as it would require more details (see III and the points below). Instead, the authors could show OrthoSAM predictions for selected S1 and K1 images (FH might not be helpful; see below) to highlight that OrthoSAM performs well on these images (similar to how SediNet images are presented in Fig. 2 in the author response). Alternatively, they could simply not use any of these images and remove the newly added statements in lines 287ff and Figs. A10, A11 entirely.

C) Using predictions from ImageGrains as a benchmark. This would require additional information and references (see below) to help readers understand the ImageGrains (v1) workflow used. I agree with the authors that such a benchmark effort also requires ground truth for comparison, so the labels would have to be used as well. Additionally, in such a scenario, it would only make sense to include Segmenteverygrain as well. However, this benchmark approach would likely require more significant changes to the manuscript, as the comparison methods and their setup would need to be introduced, and the results would need some discussion. I do not think this approach would help the manuscript, as it was not intended in the original study design, and it would needlessly divert attention from OrthoSAM to other methods.

Nevertheless, I want to emphasize that, although important, these corrections are minor (in cases A and B) and, I hope, straightforward.

Kind regards, David Mair (Uni Bern)

Specific points:

1) Unclear if and why predictions instead of labels were used

The authors write in their response:

“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. [...] However, we refrain from treating these data as a true label dataset and do not report revised accuracy and precision statistics. [...] 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.[...].

And in line 287 (track-change version): “We explored additional validation datasets, such as: three semimanually labeled images from the ImageGrains database (Mair et al., 2024)”

Which images were used from Mair et al (2024), i.e., the photos from their Fig. 8 or some image tiles from the 81 image tiles in the train/test splits as described in their Table S1? The tiles have binary mask labels (the TIF files with “mask” in the name in the subfolders of “ML_dataset” in Mair, 2023). In contrast, the three larger images have no labels (in the dataset of Mair, 2023, the predictions used for Fig. 8 in Mair et al., 2024 are stored with “pred” in the file names in the folder “field_examples”). Could there have been confusion between these folders/images? Please specify which set of grain masks was used, and add the tile name in Figs. A10, A11, if tiles are used.

Thank you for the detailed comments and clarifications. We are aware of the TIF label masks provided in the *ML-dataset* directory and have previously inspected these data. However, in the previous revision, we compared our results to the predictions instead, as the TIF label masks are only available for a subset of the cropped patches of the original images. There are not enough labeled patches to reconstruct the full images, and it is difficult to determine precisely where these crops were taken from in the original images. Furthermore, the manually labelled patches does not cover the full grain-size spectrum and the total number of labeled grains does not allow a statistically meaningful comparison with OrthoSAM. This makes it challenging to directly compare the OrthoSAM results, which operate on the full images, with the available TIF labels. Therefore, we have used the predictions, because this provides a more complete dataset.

We also considered running OrthoSAM directly on the RGB patches. However, given the design of OrthoSAM, individual patches alone are not suitable for its evaluation and would not provide a meaningful comparison. One of the key challenges OrthoSAM aims to address is the segmentation of objects across a wide range of scales. This aspect cannot be properly evaluated using cropped patches, as larger objects cannot be fully captured within these local regions.

When processing full images, OrthoSAM can resample the input to a coarser resolution so that larger objects fit within the tiles used during segmentation. This mechanism is not feasible when working with isolated patches. As a result, larger pebbles intersecting the patch boundaries would not be segmented correctly, while only smaller pebbles fully contained within the patches would be detected.

Furthermore, all labels in Mair et al. (2024) were entirely manually annotated for the 81 image tiles. Please correct “semi-manual” in line 287 (that is, if it actually refers to labels and

not predictions). The grains in the IG1 data were densely labelled, with all visible pebbles consistently annotated by a single annotator (which means I traced every visible pebble in each tile by hand). Of course, different annotators might annotate differently, and, despite best efforts, true pixel-precision is not always achieved. However, it is unclear why the authors arrived at “Our initial analysis and discussion [...] suggested that we can not use their labeled data as a validation dataset” and at “we refrain from treating these data as a true label dataset.”. Was there a specific reason the labels were not considered usable (beyond the potential confusion mentioned above)?

Thank you for pointing this out. The description has been corrected. We refrain from using the labeled dataset as the labeling was not available for the entire image. We found a subset of the cropped patches of the original images. There are not enough labeled patches to reconstruct the full images.

While these manual labels could be used to quantify this using AP or other scores, I do agree that too many false-positive predictions of irrelevant objects might preclude such an effort. Comparing OrthoSAM predictions to the labels based on the number of grains and size distributions (as stated in lines 287ff) is also valuable; however, the rebuttal letter and Figs. A10 and A11 state that the comparison presented uses model predictions rather than the labels (see above). These predictions cannot substitute for labels because:

ImageGrains (v1) predictions are not perfect reconstructions, as noted in Mair et al. (2024) and Mair et al. (2026). Since these labels were also used as ground truth for training ImageGrains (v1) models, a perfect model in this workflow would return masks that match the annotations. Furthermore, the number and size of predicted masks depend on the ImageGrains setup used (i.e., the minimum diameter, edge filtering, and segmentation model; see also below). Therefore, I would expect the size distributions (as shown in Fig. A11) to differ across setups. Thus, reporting predictions requires information about the ImageGrains (v1) setup and references to Cellpose (Stringer et al., 2021), the backbone architecture for the segmentation model.

If predictions from ImageGrains (v1) were to be used, please indicate this clearly (see also previous comment) and do not refer to them as labels. Also, please indicate the ImageGrains version and the model weights used at an appropriate position in the manuscript. Furthermore, please indicate which setup was used, i.e., if the default filter for min. grain size of 12 px for the b axis of fitted ellipses (i.e., “px_cutoff”) or if any edge filter was used, and if the command line interface or a custom workflow was used. The min. grain size filter could explain why fewer small pebbles were detected as reported by the authors (“As illustrated in Figure 1, OrthoSAM [...] tends to detect finer objects”).

2) The issue with FH

The authors state in the rebuttal letter that: 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. [...] 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.

The FH image (and the FH image tiles) are very different (from S1, K1, or the Ravi River) because they consist of vertically orientated images from fluvial sediment in the walls of a

gravel pit (Mair et al., 2024; Garefalakis et al., 2023). In addition, the authors' observations about reduced clarity and focus are correct. For our study, we explicitly chose these images because they differed from nadir images of gravel bars to assess how well ImageGrains (v1) could be customized. These challenges and the different image content were the reasons I explicitly suggested exploring only the S1 subset in my previous review.

In our study, we found that a specialized model (fh+) was much better for FH images than the default general model (full_set) in Mair et al. (2024). Therefore, the data for Fig. 8 in Mair et al. (2024) were obtained with different models for FH and for K1, S1 (as indicated in our paper). If the authors really want to use ImageGrains predictions for FH, this needs to be explicitly stated. However, I really think all this would be beyond the scope of the study. Therefore, I suggest either using the manual labels for comparison or omitting FH images altogether.

Thank you for explaining the complexity of the FH dataset. We were not aware of these challenges and are not using the manual labels.

3) Fig. A11

There is something off in the description of this figure. First, the authors write in the caption that OrthoSAM returns more grain masks: “[...] with OrthoSAM (OS and Oss) detecting more objects than ImageGrains (IG).”. However, in Figure A11 itself, it is stated that ImageGrains returns more grains for S1 (nIG: 4660 to nOS: 4313 and nOSs: 2915) and K1 (nIG: 1883 to nOS: 1358 and nOSs: 1113); IG only returns fewer objects for FH. Furthermore, the authors claim in the rebuttal letter and in lines 289-292 that OrthoSAM seems to detect more small objects. This seems to contradict the presented cumulative size distributions, as IG (the blue curve) yields results with objects that are relatively smaller than predicted by OrthoSAM, even beyond percentile 80 (K1, S1) or percentile ~30 (FH).

Thank you for pointing this out. The comparison with ImageGrains prediction has been removed and we have revised the text accordingly. Regarding the number of detected objects, a size filter was applied to the OrthoSAM predictions before reporting the object count. We observe that the unfiltered OrthoSAM predictions detect noticeably more objects than ImageGrains. A comparison of the grain-size distributions suggests that OrthoSAM identifies a greater number of very small objects. We attribute this difference to two potential reasons. First, it is a known limitation that OrthoSAM predictions with diameters below 20 pixels are more prone to noise and false positives, which may explain the detection of a larger number of very fine objects. Second, a minimum size filter was applied both to the training data and to the predictions of ImageGrains. To ensure a fair comparison, we therefore applied the same filtering step to the OrthoSAM results.

Other comments:

Line 288: ImageGrains refers either to the library/workflow or to the dataset. Please correct “database”. Please also include the version (v1) to avoid confusion with the newer version (Mair et al., 2026). Fig. A10: Typo: ImageGrains.

Thank you for pointing this out. We have corrected the typos and revised the text.