Response to comments by reviewer 1 (Laure Guerit)

Reviewer comments are indicated in *blue and italic font*, while our responses are formatted as regular and black text. Our responses follow the order of the reviewer comments, which are structured by manuscript section.

In the paper, Mair et al. present a new method of image segmentation dedicated to the detection of sediments in 2D or 3D images (including CT scans) based on deep-learning algorithms. They build on previous work by their team and take advantages of up-to-date image segmentation methods (SAM, Cellpose-SAM) to provide ImageGrains 2.0. On average, this new algorithm outperforms previous methods and provides grain distributions estimated in very good agreement with the ground truth (manually annotated images). Overall, this is a very good, well-written manuscript supported by appropriate and clear figures.
However, I think the manuscript is a bit too focus on the method and the increase in precision it brings with respect to previous studies, at the cost of a general contextualisation. In the current form, I feel that it is difficult from someone that is not really into to grain-size measurements and deep-learning to really get the novelty and interest of this work. In fact, why we need such a new method is mostly addressed in a few lines at the very beginning and a bit in the discussion. Why we need to improve the precision on grain-size distribution is not really addressed (and I think this is a real question). Therefore, I think that the manuscript could benefit from some additions in particular in the introduction and in the Discussion, related to the use of grain sizes distributions, in order to reach the broad audience of Earth Surface Dynamics and to highlight the interest of this new approach (which I find really cool and promising).
We thank the reviewer for their insightful summary and for the very encouraging overall evaluation of our work. We follow the suggestion to put our work in a broader context. Therefore, we expand both the introduction and discussion to include discussions on the need for such accurate grain size and shape estimations for large datasets, the challenges that come with that, and how our approach is a step forward in addressing these challenges (for details, see responses to specific comments below and responses to related comments by reviewer 2).

Introduction
As mentioned above, the manuscript could better present the need for refined grain segmentation method, beyond the fact that better segmentation leads to higher accuracy in grain diameters and shape descriptors (this point is very well explained latter in the text, this is great). Why does it matter ? What can we do with an increase resolution / what do we miss with current methods ?
In a provocative way, my question is, if I want to get a grain-size distribution, why should I get into deep-learning rather than taking a ruler and a pen ? I think your method is really efficient and nice and the manuscript should state this point more clearly.
We expand introduction to include the need for such accurate grain size and shape estimations for large datasets.

L. 42 and following: You mentioned manual measurements and imagery but I think 3D methods should be mentioned as well here (e.g., based on roughness, Tario-Vasquez et al, 2017, on point clouds, Steer et al, 2022)

We add additional references to those methodological papers where the approaches are based on 3D point-cloud segmentation and point-cloud roughness, along with their respective references, in the suggested location.

In addition, the method presented here is based on image segmentation, ie on areal measurements that are not equivalent to a volumetric or surfacic sample (see for example Bunte and Abt on this question). Yet, all the grains are not detected on the image so that it is not a real areal distribution, nor a grid one. This is ok if you want to do relative sizes and to compare the evolution of a given site, without interested for the physical value, but this is far less correct if someone wants to do some physical calculations or numerical simulations. This is a classic limitation of such approach that must be addressed somewhere in the paper.

So far, we have only briefly discussed this inherent limitation due to the 2D nature of images in sections 4.4.1 and 4.4.3. We agree that differences in grain-size data across data types and method families should be communicated more comprehensively. Therefore, we expand section 4.4.1.

l. 83 new types of imagery: I think this is a real addition of your method, this is great and clever. You could develop this part a bit more in the results and discussion.

We expand on this part in the revised version of the manuscript, both in the results and discussion sections, in line with other comments (see response to comments below), especially on the inclusion of 3D-XRD stacks (see also response to related comments by reviewer 2). However, due to a lack of 3D labels (cf. section 4.2.2) and to keep the manuscript focused, we refrain from expanding it by comparing the approach to other 3D data types and 3D segmentation methods. This would likely warrant a dedicated study of its own.

Methods

I understand that the database has been presented in another paper and that you do not want to do it again. However, I felt it a bit confusing to have so many reference to previous papers by your team. Many readers will not have read these papers or will not remember the specifics of them. Therefore, I think that section 2.1 could be reframed in order to be more accessible (less references, less acronymes and a more general explanation of the type of rocks that are in the images - this is partly done in the discussion, it could be done here).

We want to emphasize that only parts of the images have been used in other studies, while all annotations, except those for the 81 tiles (APF, S1, FH) from Mair et al. (2024), have not been published before. They are an essential component of this work. However, we concede the need for a clearer presentation of the composition of the IG2 dataset (also in line with the comment on line 214; see also the corresponding response below). Therefore, we follow the suggestions to improve the accessibility of section 2.1 by moving the former table S1 in the previous version of the manuscript to the main text (now new table 1). This allows us to move

details from the text into the table, where we now present more information on the general image content and setting in the revised version of the manuscript.

2.2.2 I like this section and the method. Yet, in line with my comments on the Introduction, it could be a great addition to the manuscript to state more clearly why you use these metrics (beyond the description of the shape, why do they matter for the study of natural systems).
The selection of metrics is based on i) robust basic implementations for the calculations of the ROI characteristic in images that have been widely used across all scientific fields, and ii) more geoscience-specific metrics (that all are based on ROI characteristics from i). These metrics are selected as examples because they have been successfully implemented in recent studies. Still, they are by no means an exhaustive list, and – depending on the specific application – their representativeness of natural systems varies widely. Therefore, we restructure this section and add some statements about the context for which each metric has been used, but we refrain from a more in-depth characterization of each metric and their uses across different applications. This would be more suitable for a review contribution, which is not our goal. Instead, for each metric we refer the readers to other studies where the respective metrics have been presented and discussed in depth. Furthermore, we emphasize that these metrics are examples, and any custom metric that is based on ROIs can be calculated from our segmentation results. We add a statement to emphasize this point.

2.3 and 2.5 these sections are well written but might be a bit hard to follow for someone not familiar with deep-learning. I don't think this is your job to explain DL but you could explain a bit the choice of the parameters and how it could impact the segmentation, and what is a precision.
The architecture we are using (Cellpose-SAM; Pachitaru et al. 2025) is a deep model and a combination of the SAM approach (Kirillov et al, 2023) and Cellpose (Stringer et al., 2021). We consider discussing the features of the architecture in detail beyond the scope of this study, as we employ this architecture in its tested default configuration. Instead, we discuss the main elements and their functions that are relevant to evaluate and fully reproduce our workflow, while we refer readers to the original publications (and the references therein) for further details. We add several explanations and additional references to address these points, which are also raised by reviewer 2 (see responses to related comments by reviewer).

l. 214 you could explicite what king of context are in these subsets as it is not clear from the name only (and it's not ideal to move to Supplement to get a sense of it).
In line with the response to the related comment above, we move the previous table S1 from the supplement to the main manuscript. We add information on the content and image type of sets S1_2 and PR now in section 2.1.

Results
l.284 are the 18 500 masks 63% of the ROIs or do you use 6% of 18 500 ? please rephrase to avoid confusion.

We clarify that we use 63% of the entire IG2 label set to estimate the shape, which corresponds to a dataset of c. 18500 grains. Additionally, with the former Table S1 now being Table 1 in the main manuscript, this should be more clearly visible.

3.2 the accuracy of your new method is higher than previous ones, yet, it is still somehow limited (cf table 1, figure 4). I would appreciate that the authors acknowledge it more clearly in this section.

Unfortunately, there is no consensus on what level an accuracy is high enough (i.e., expressed as average precision or other scores) to avoid being limited. Such score thresholds would always depend on specific applications and datasets (e.g., Zaidi et al., 2022), as evidenced by differences in the scores across data subsets (cf. previously Table 1; now Table 2). Therefore, we strictly report performance values only in the result section 3.2. We already discuss cases where segmentation performance is still the limiting factor in the discussion in sections 4.2.1, 4.3, and 4.4. Furthermore, variations in the limitations of the segmentation performance are a main reason we critically evaluated segmentation performance across different shape metrics in section 4.3. However, we update sections 4.2.1, 4.3, and 4.4 to make it clearer that there is still room for improvement in the accuracy of the segmentation and that, depending on the application and data used, this remains the main limitation.

3.3 overall, you are able to derive grain-size distributions that match ground truth values in most cases. However, how to deal with the other situation ? is there a way to anticipate that the results are shifted ? I understand that if you work on a large collection of data, knowing that about 90% of the distributions are correct can be satisfying. But if you have only one or two images, this is more tricky. This should be addressed here (if you have materials to do so) or in the Discussion.

When our model (or a custom model) is used to measure grain size and shape from a small number of images, we recommend visual inspection of the segmentation results (which are the default outputs of our workflow across all default configurations). In case of incomplete segmentation (over- or under-segmentation, as well as missed grains), a manual post-processing, i.e., correcting the mis-segmented grains, is required. This could be done through the graphical user interface or any software tool that allows the creation and modification of mask ROIs. We add a corresponding statement to section 4.4 of the discussion.

Discussion

As mentioned previously, I miss a section dedicated to the 2D approach in terms of grain-size description (areal measurements, comparison with manual measurements, etc). In addition, I would like the authors to comment on the fact that some grains are not segmented (either because they are too small or because it is too difficult to segment them from the images), therefore, they do not retrieve a real areal distribution (Bunte and Abt 2001 would be a very classic reference to start this discussion).

We have not included such a discussion before for two reasons. First, we focus primarily on more general grain-shape applications. The classic application of determining the grain sizes of

fluvial sediments is only one among several. Second, the specific limitations of applying this approach, inherent to the 2D nature of images, are not new or different from previous work. Therefore, this discussion will necessarily reiterate arguments already presented in the literature (e.g., Purinton and Bookhagen, 2019; Steer et al., 2022; Mair et al., 2024). However, we acknowledge that it is important to clearly communicate these limitations (see also the response to the related comment above). Therefore, we add statements with references to the relevant literature that summarize the specific limitations for applying this approach to obtain grain sizes for fluvial sedimentary particles.

l 486 missing space between SAM and (Chan et al) + extra dot at the end of the line.
Corrected.

l. 503 this is a great result of your work. Do you have any idea on why this value of 68% ? Could you suggest some recommendations for image acquisition to maximize the likelihood to reach these 68% ?
This seemingly minimum threshold of 0.68 AP@0.5 IoU (or c. 0.5 mAP@50-90) for statistically representative GSDs depends on three factors: i) the specific segmentation results (i.e., whether FP, FN are caused by over/under-segmentation or missed grains), ii) the grain size sorting, and iii) the shape of the GSD itself. These factors are interdependent. For example, mis-segmented grains have a much smaller effect on the obtained size distribution for very well sorted sediment with a narrow and Gaussian-shaped GSD than for poorly sorted sediment with log-shaped or bi-modal GSDs. Therefore, we doubt that such a threshold exists as a single, generally applicable value, and it might not be the best strategy to aim solely to surpass it when evaluating custom segmentation models. We estimated this number to provide an orientation for other studies, but acknowledge that it could vary significantly depending on the underlying data. Hence, workable recommendations aim to improve segmentation performance and follow classic image acquisition guidelines, such as using suitable sensors with sufficiently high resolutions and maintaining stable image acquisition conditions (e.g., light and exposure time). We add a statement to highlight that a high-quality image acquisition (and, in the case of model training, high-quality labels) should be targeted to improve model performance and result quality.

l 537 fro -> for
Corrected.

l 557 could you elaborate a bit on what you mean by image types that are substantially different" ?
We expand on this point and add examples for clarification.

4.4.2 I would encourage the authors, if possible, to mention the environmental costs of DL methods

We thank the reviewer for raising this important point (e.g., Strubell et al., 2019; van Wynsberghe, 2021). It is difficult to estimate the environmental costs of such models completely due to the question of how to factor in the initial training of SAM (which is reported to have cost c. 6963 kWh or equivalent to 2.8 metric tons of CO_2 release; Kirilov et al. 2023), and the open question of how to account for secondary effects beyond the computational cost and energy consumption (e.g., Yu et al., 2024; Bouza et al., 2024). Even quantifying the environmental impact of energy consumption for training or inference requires dedicated tests and often-unavailable information, such as where a specific piece of hardware was used and what energy source powered it (e.g., Lacoste et al., 2019; Patterson et al., 2021; Jay et al., 2023).
However, we can discuss the computational cost (and therefore the energy cost) of our approach, based on the results of Pachitariu et al. (2025; cf. Tables S2 and S3 therein). We note that the Cellpose-SAM architecture runs on local desktop computers with mid-range consumer hardware configurations, both for inference and for retraining smaller models (not more than a few hundred images). Therefore, despite the rather costly initial training and release of SAM (Kirilov et al., 2023), using or fine-tuning our model has a comparatively small computational cost and thus a low environmental impact, comparable to running any other software for similar durations on a desktop PC. Following the suggestion of both reviewers 1 and 2, we add further statements in section 4.4.2 that raise awareness of the issue and that reflect the arguments made here.

4.4.3 in line with my comment on the Introduction, I think that the manuscript could do a better job at presenting why grain size distributions and metrics are of interest, and what can be learn from such data set.

In line with the related comments, both above and by reviewer 2, we expand the section by discussing grain size applications where we see improvements through our approach, while referring to existing literature for a more in-depth discussion.

References

Bouza, L., Bugeau, A., and Lannelongue, L.: How to estimate carbon footprint when training deep learning models? A guide and review, Environ. Res. Commun., 11, https://doi.org/10.1088/2515-7620/acf81b, 2023.

Jay, M., Ostapenco, V., Lefevre, L., Trystram, D., Orgerie, A.-C., and Fichel, B.: An experimental comparison of software-based power meters: focus on CPU and GPU, in: 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 106–118, https://doi.org/10.1109/CCGrid57682.2023.00020, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R.: Segment Anything, https://arxiv.org/abs/2304.02643, 2023.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T.: Quantifying the Carbon Emissions of Machine Learning, https://arxiv.org/abs/1910.09700, 2019.

Mair, D., Do Prado, A. H., Garefalakis, P., Lechmann, A., Whittaker, A., and Schlunegger, F.: Grain size of fluvial gravel bars from close-range UAV imagery – uncertainty in segmentation-based data, https://doi.org/10.5194/esurf-2022-19, 23 May 2022.

Mair, D., Witz, G., Do Prado, A. H., Garefalakis, P., and Schlunegger, F.: Automated detecting, segmenting and measuring of grains in images of fluvial sediments: The potential for large and precise data from specialist deep learning models and transfer learning, Earth Surf. Process. Landf., 49, 1099–1116, https://doi.org/10.1002/esp.5755, 2024.

Pachitariu, M., Rariden, M., and Stringer, C.: Cellpose-SAM: superhuman generalization for cellular segmentation, https://doi.org/10.1101/2025.04.28.651001, 1 May 2025.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J.: Carbon Emissions and Large Neural Network Training, https://doi.org/10.48550/arXiv.2104.10350, 2021.

Purinton, B. and Bookhagen, B.: Introducing PebbleCounts: A grain-sizing tool for photo surveys of dynamic gravel-bed rivers, https://doi.org/10.5194/esurf-7-859-2019, 17 September 2019.

Steer, P., Guerit, L., Lague, D., Crave, A., and Gourdon, A.: Size, shape and orientation matter: Fast and semi-automatic measurement of grain geometries from 3D point clouds, Earth Surface Dynamics, 10, 1211–1232, https://doi.org/10.5194/esurf-10-1211-2022, 2022.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation, Nat. Methods, 18, 100–106, https://doi.org/10.1038/s41592-020-01018-x, 2021.

Strubell, E., Ganesh, A., and McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP, 2019.

van Wynsberghe, A.: Sustainable AI: AI for sustainability and the sustainability of AI, AI and Ethics, 1, 213–218, https://doi.org/10.1007/s43681-021-00043-6, 2021.

Yu, Y., Wang, J., Liu, Y., Yu, P., Wang, D., Zheng, P., and Zhang, M.: Revisit the environmental impact of artificial intelligence: the overlooked carbon emission source?, Front. Environ. Sci. Eng., 18, https://doi.org/10.1007/s11783-024-1918-y, 2024.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., and Lee, B.: A survey of modern deep learning based object detection models, https://doi.org/10.1016/j.dsp.2022.103514, 30 June 2022.