https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

ImageGrains 2.0: Improved precision and generalization for grain segmentation

David Mair[1], Guillaume Witz[2], Ariel Do Prado[1,3], Philippos Garefalakis[1], Amanda Wild[4,5], Fanny Ville[6], Bennet Schuster[1], Michael Horn[2], Jürgen Österle[7, 8] Stefano C. Fabbri[9], Camille Litty[9], Stefan Achleitner[10], Sebastian Leistner[10], Clemens Hiller[10, 11], and Fritz Schlunegger[1]

[1]Institute of Geological Sciences, University of Bern, Bern, 3012, Switzerland
[2]Data Science Lab, University of Bern, Bern, 3012, Switzerland
[3]Institute of Geosciences, University of São Paulo, São Paulo, 05508-080, Brazil
[4]Institute of Physical Geography and Geoecology, RWTH-Aachen University, Aachen, 52062, Germany
[5]GFZ Helmholtz Centre for Geosciences, 14473 Potsdam, Germany
[6]Fluvial Dynamics Research Group (RIUS), University of Lleida, Lleida, 25003, Spain
[7]School of Geography, Environment and Earth Sciences, Victoria University of Wellington, Wellington , 6012, New Zealand
[8]Amt der Vorarlberger Landesregierung, Bregenz, 6901, Austria
[9]Federal Office of Topography swisstopo, Wabern, 3084, Switzerland
[10]Unit of Hydraulic Engineering, University of Innsbruck, Innsbruck, 6020, Austria
[11]Natural Hazards and Risk Management, Geoconsult ZT GmbH, Puch bei Hallein, 5412, Austria

Correspondence to: David Mair (david.mair@unibe.ch)

Abstract. Recent advances in deep-learning–based image segmentation have enabled the development of automated approaches to detect individual grains and measure them for geoscientific applications. These methods facilitate the creation of much larger and more precise datasets than traditional manual grain measurements. However, they typically perform best as specialized models trained on homogeneous, task-specific datasets, and often show reduced accuracy when used to generalize to different data types.

Here, we present an updated framework, ImageGrains 2.0 that leverages Cellpose-SAM, a recently published next-generation deep-learning model originally developed for cell segmentation in biomedical research. It currently represents the state of the art for dense segmentation in 2D and 3D biomedical datasets, and yields robust, and is capable to generalize across distinctly different image datasets. These properties allow us to re-train the model with geoscientific dataset comprising annotated images of fluvial gravel, coarse pro-glacial deposits, and X-ray computer tomography scans of glacial till and marine sand. We benchmark the segmentation performance of the method against ground-truth annotations, compare it to the performance of other segmentation methods, and we evaluate measurement accuracy. Our results indicate that this approach outperforms existing methods and confirm that the outstanding performance of Cellpose-SAM is transferable to segment sediment grains. We analyze the size and shape of these segmented grains and find that an increase in grain segmentation accuracy leads to more precise and realistic morphometric results, e.g., more accurate grain size distributions. Additionally, we introduce an interactive graphical user interface for image annotation and correction of model predictions, facilitating the use of the framework in a broader range of image settings. Furthermore, this study underscores the

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

importance of curating of more publicly available datasets, which could pave the way towards the generation of a foundation model for segmenting granular particles in geoscientific imagery.

1 Introduction

Data on the size and shape of granular particles have been used across a broad range of geoscientific research areas, and such information has provided the basis for the quantification of the physical and chemical properties of clastic materials (e.g., Sklar, 2024; Israeli and Emmanuel, 2018). Grain morphometry, for instance, is essential for studying sediment production and transport dynamics in environments such as fluvial, glacial, and hillslope systems (e.g., von Eynatten et al., 2012; DiBiase et al., 2017; Allen et al., 2017; Garefalakis et al. 2024). Traditionally, such data have been collected through laborious manual measurements of grains in the field (e.g., Bunte and Abt, 2001) or on imagery (e.g., Butler et al., 2001; Carbonneau et al., 2004; Detert & Weitbrecht, 2012; Buscombe, 2013; Purinton & Bookhagen, 2019). In this context, machine-learning tools have been developed more recently in an effort to automate grain size and shape measurements, to improve the data quality, and to allow an increased number of observations. Among these, texture-based methods predict percentile values of grain size distributions if an unambiguous correlation between an image texture and a characteristic grain sized distribution exists, and if these were included in the training data (e.g., Buscombe, 2020; Lang et al., 2021). In contrast, segmentation-based methods delineate individual grains though object detection (Chen et al., 2022; Mair et al., 2024; Miazza et al., 2024; Sylvester et al., 2025) and facilitate the creation of large datasets, which allow for size and shape analysis down to an individual grain level. However, these segmentation models work best when trained as narrow specialist models on homogenous datasets, which often require task-specific, and sometimes site-specific, training and careful curation of the corresponding datasets (e.g., Chen et al., 2023; Prieur et al., 2023; Azzam et al., 2024; Miazza et al., 2024; Zegers et al., 2025; Schuster et al., 2025).

During recent years, a new generation of deep learning models that use a transformer architecture has become widely used in the field of computer vision for tasks related to the segmentation of objects in images (e.g., Dosovitskiy et al., 2020; Li et al., 2022). This resulted in the development of foundation models, such as the Segment Anything Model (SAM; Kirillov et al., 2023; Ravi et al., 2024), which are trained on very large and general datasets. These foundation models have proven effective at generalization, i.e., being able to predict outcomes for previously unseen data, which were not used for training. Due to a strong inductive bias, they are also considered as effective at out-of-distribution detection of objects, especially when fine-tuned on smaller datasets (Hendrycks et al., 2020; Fort et al., 2021). While these models perform well at segmenting numerous different types of objects in images, they are less efficient at accurately segmenting large quantities of a narrow range of objects, especially when they are not fine-tuned to specific datasets or when no specific prompts are used (Sylvester et al., 2025; Chan et al., 2025).

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

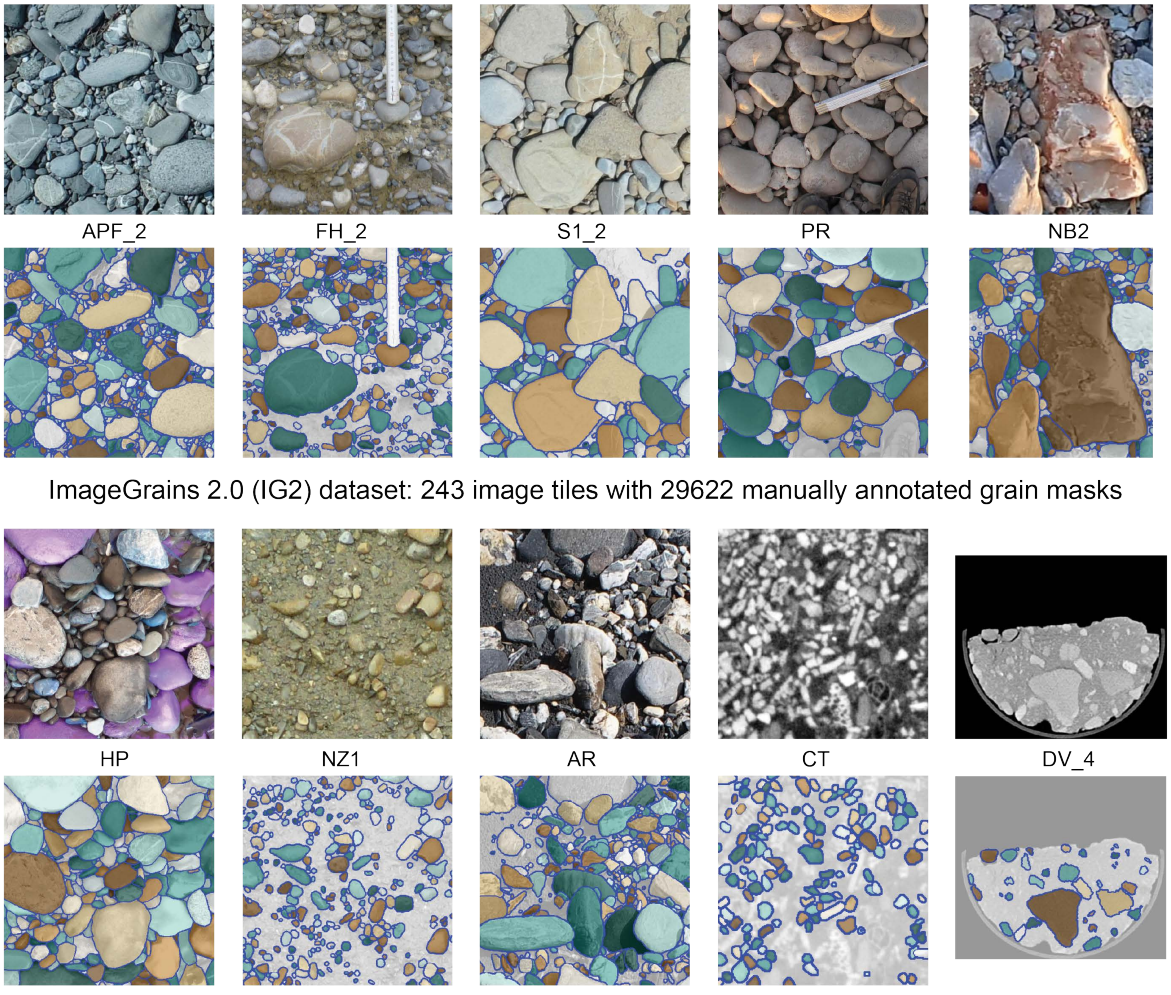


Figure 1: Example images and their manually annotated grain labels for various types of imagery and grains used in the IG2 dataset, and their respective indicated subset (Mair et al., 2025a; see Table S1 for more details). Individual clasts are shown in random colors with blue outlines.

Here, we present an updated framework building on Mair et al. (2024) that leverages the strengths of Cellpose-SAM (Pachitariu et al., 2025), a recently published next-generation deep-learning model originally developed for cell segmentation in biomedical imagery. This model eliminates weaknesses of SAM and improves the performance for dense segmentation of many instances of the same object type with high accuracy, while maintaining the outstanding generalization ability. We utilize the new Cellpose-SAM model that was trained on large datasets of predominantly biomedical imagery of cells (for details, see Pachitariu et al., 2025) and retrained it to find grains in images of clastic sediment. This transfer learning approach allows us to utilize both the representations learned by SAM that enables the generalization across widely differing datasets and data types, and the Cellpose segmentation framework (Stringer et al., 2021) that facilitates the efficient and dense segmentation of grains with high accuracy without the necessity of prompt

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

engineering. To achieve this, we curated a dataset of 243 annotated image tiles from various types of sediment grains (Fig. 1). We then compare the segmentation results of our re-trained Cellpose-SAM model with the results of other state-of-the-art approaches, and we test the models' ability to generalize by using subsets of our dataset as unseen test splits. Finally, we highlight the potential of applying our workflow to 3D datasets of stacked images retrieved by X-ray computer tomography (CT) scans. Our results indicate that the new framework outperforms existing methods both in accuracy of the resulting segmentation, and in the capability to segment grains in new types of imagery.

2 Methods

For ImageGrains 2.0, we employ the recently released Cellpose-SAM (Pachitariu et al., 2025) model architecture for segmenting biomedical images, which itself utilizes the ViT transformer of the Segment Anything Model (SAM; Kirillov et al., 2023) as backbone together with the gradient tracking of the original Cellpose framework (Stringer et al., 2021). Similar to the approach of Mair et al. (2024), we use a dataset consisting of images with annotated sediment grains (Fig. 1). In contrast to Mair et al. (2024), Imagegrains 2.0 (IG2) is a much larger dataset including more image types (see Section 2.1). We use the IG2 dataset to train our model and to evaluate its capability to quantify the size and shape of sediment grains (Section 2.2). In our approach, we apply transfer learning and retrain the pre-trained Cellpose-SAM foundation model to segment grains in images taken from clastic sediments (Fig. 2). In Section 2.3, we summarize key aspects of this foundation model and the adaptions we made. Next, we describe how we set up other methods and models that we use to benchmark our approach (Section 2.4). We then proceed by quantifying and comparing the segmentation performances across all methods (Section 2.5). Finally, we obtain a set of aggregated metrics, which are based on the measured sizes and shapes of individual grains, to evaluate the effect of using segmented grain masks with varying precision. We note here that we use term *method* for entire segmentation workflows, while *model* refers to a specific segmentation model. This distinction becomes important as some methods combine several models and we sometimes train several models of the same method with different datasets.

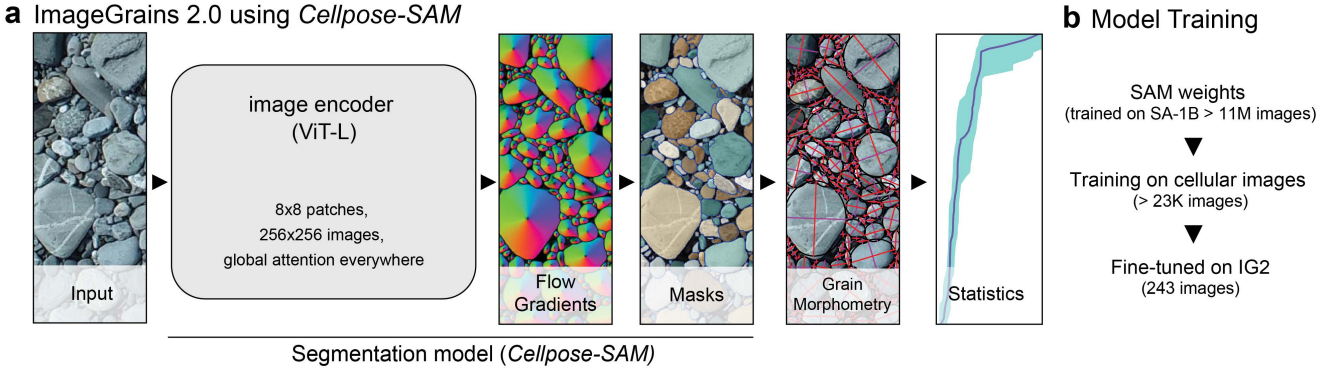


Figure 2: Overview of our workflow (a) that uses a re-trained Cellpose-SAM (Pachitariu et al., 2025) architecture for grain segmentation. The re-training was done by fine-tuning to the IG2 dataset (Mair et al., 2025a; see Table S1 for more details).

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

2.1 The ImageGrains 2.0 (IG2) dataset

To train segmentation models that are able to map a large variety of sediment grains on different image types, we complemented and expanded the dataset (IG1; Mair, 2023) of Mair et al. (2024) by adding new image data and labels. For each image that was used in the expanded IG2 dataset, we chose subset tiles of varying sizes (ranging from 50×62 to 2750×2000 pixel) that capture the full grain size variability and the complexity of the image content. Specifically, we selected tiles that contained various objects such as scales, hands and equipment; tiles that featured different types of vegetation and water bodies, and that were acquired under variable light conditions. This resulted in a large variety of tiles for each image (Mair et al., 2025; Table S1). These tiles were annotated manually using the LABKIT plugin (Arzt et al., 2022) for FIJI (Schindelin et al., 2012) and napari (v0.6; napari contributors, 2019), where each grain was labelled individually (i.e., dense labelling) as precisely as possible at the scale of individual pixels. In total, we added 162 such annotated image tiles from various sources and settings to the 81 tiles of fluvial sediment images compiled and annotated for the previous version (IG1). For all datasets, we manually generated representative subgroups of images, called stratified train and test splits (Table S1), to create balanced subsets for all imagery.

First, we proceeded by adding 5 additionally labelled image tiles to the original ImageGrains (IG1; Mair, 2023) dataset as a first task. The goal was to improve the balance between the different images and data sources (i.e., Brayshaw et al. 2012; Litty and Schlunegger, 2017; Mair et al., 2022; Chen et al. 2022; Garefalakis et al., 2023) in the respective test and training splits. In particular, we added 1 image of vertical gravel outcrops in the FH_2 subset, 1 image tile from the Swiss Sense River in the S1_2 subset, and 3 image tiles of fluvial pebbles from variable sources in the APF_2 subset. In addition, we added 6 subsets containing 103 image tiles taken from fluvial sediment from rivers in Spain (with in the field painted clasts; HP, PP, n = 15), Peru (PR, n = 7; Litty and Schlunegger, 2017), New Zealand (NZ2, n = 23), Switzerland (AR, n = 16), and Namibia (NB2, n = 42). Furthermore, we included two more subsets of image tiles taken from coarse-grained and angular proglacial sediment (JF, n = 9; Hiller et al., 2023), and images retrieved from near-vertical outcrops of lithified conglomerates (NZ1, n = 20). The images were acquired with different handheld and unscrewed aerial vehicle (UAV)-borne camera systems at varying resolutions (Mair et al., 2025a; see also Table S1). Finally, we completed the dataset by using X-ray CT (XRCT) scans taken from glacial tills (DV4, n = 19; Schuster et al., 2024; 2025), and micro-CT images of bio-clastic marine sand (CT, n = 6; Fabbri et al., 2024), which we annotated in two respective subsets. All these images were selected for variations regarding the objects displayed on the images. This includes - on purpose - the occurrence of vegetation and other objects that are not sediment particles to test the model against the possibility of false detections, in order to challenge the models beyond variabilities in the lithology, color, grain size and shape of the clasts. The combination of all subsets resulted in a total of 203 and 40 annotated image tiles that we used for training and testing, respectively.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

2.2 2D Grain morphometry

Aside from quantifying the segmentation performance, we assessed the importance of precisely segmenting grain masks for yielding accurate results using grain size and shape metrics as benchmark information. For each grain mask, or region of interest (ROI), we used standard image analysis tools implemented in scikit-image (v0.25.2; van der Walt et al., 2014) that have been successfully used to represent the morphometry of grains in geoscientific research (e.g., Szabó et al., 2015; Miller et al., 2024; Lepp et al., 2024; Benet et al., 2024; Back et al., 2025). Here we fitted ellipses to approximate the shape of the target grains for which we calculated the lengths of the minor and major axes (b- and a-axis, respectively, of an ellipse). This approach has been demonstrated to well capture grain sizes of clastic material in 2D images (e.g., Purinton and Bookhagen, 2019; Chardon et al., 2022; Garefalakis et al., 2023; Mair et al., 2024; Sklar, 2024). The uncertainties of the grain size percentile values are quantified through bootstrapping, thereby resampling any grain size distribution (GSD) a 1000 times (for details, see Section 2.4 in Mair et al., 2022). We then calculated differences between grains in the ground truth and predicted grains as difference for percentile values. Furthermore, we tested if GSDs were statistically different between predictions and ground truth with a wo-sample Kolmogorov–Smirnov test. Here, the two distributions being identical was the null hypothesis, which we consider rejected for $p > 0.05$.

We calculated the eccentricity of the same ellipse fit to approximate the grain elongation in 2D, which is the ratio of the focal distance over the length of the major axis. Similarly, we used the convexity, sometimes also called solidity (e.g., in scikit-image; van der Walt et al., 2014), which is the ratio of pixels in the ROI to pixels within the convex hull, as proxy value for the 2D roughness of each grain. Next, we obtained the isoperimetric ratio (IR) and normalized isoperimetric ratio (IR_n, Pokhrel et al., 2024; Quick et al., 2020) for each grain mask as indicator for the roundness of a grain. We note the selected approaches to compute IR and IR_n values can return values > 1 in some cases, which could be the consequence of geometrically imperfect reconstructions (see supporting information of Quick et al., 2020). Finally, we track the 2D grain orientation as azimuth angle of the b-axis of the above-described ellipse fit and the y-axis, i.e., the image height, of each image tile in degrees from 0° to 180°. We calculated all the aforementioned metrics for all ROIs in both ground truth and predicted masks that fall in the central 90% of each image tile by avoiding the outermost 5% from each image edge. We did so to avoid a bias that could be introduced by considering grains – possibly cut ones – at the border of image tiles. We then calculated differences between ground truth grains and predicted grains for all corresponding metrics. By comparing these morphometric values across datasets, we can quantify how the segmentation quality affects the morphometric results.

2.3 Cellpose-SAM: re-training and inference

The Cellpose framework (Stringer et al., 2021) used a deep-learning model, which is based on a U-Net (Ronneberger et al., 2015) type of neural network with image style transfer (Gatys et al., 2016). This framework was combined with an equation modelled on heat diffusion to predict vector flows. From these flows, individual objects are segmented through gradient tracking. In Cellpose-SAM the previous backbone model was replaced with a modified version of the SAM transformer

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

(Kirillov et al., 2023; see also Section 2.4.1 below for more details on SAM). Specifically, it used the image encoder module of SAM and replaced the decoder parts with Cellpose's vector flow representation for prediction (Pachitariu et al., 2025). Moreover, the encoder itself was modified in several ways for the Cellpose-SAM architecture. First, the dimension of the input image was reduced to 256×256 pixels (from 1024×1024), and the patch size was reduced to 8×8 (from 16×16). Accordingly, the position and patch embeddings were also down-sampled, while global attention was used for all layers. This approach differed from using global attention in only some layers in the original SAM architecture (for more details, refer to Pachitariu et al., 2025). Pachitariu et al. (2025) initialized the Cellpose-SAM model with the SAM ViT-L model weights, which itself had been trained on the SA-1B dataset (Kirillov et al., 2023). They then trained Cellpose-SAM on an updated dataset of 22826 cell and cell nuclei images with >3.3 million labelled objects. Notably, the updated architecture is much larger (> 304 million trainable parameters compared to > 6.6 million trainable parameters in the old backbone model). Furthermore, the improved model can use multi-channel, i.e., color, images, because of its training with random channel permutations. This was not possible with the previous models that converted the images to single-channel greyscale images before the segmentation. As a result, multi-channel images are now the default image input.

We retrained the Cellpose-SAM model on our ImageGrains 2.0 dataset using default settings for the custom re-training to obtain our new default model for ImageGrains. This included training for 500 epochs with a learning rate of 1e-5. Here, we used all 203 image tiles of the train split in every epoch. Training was accomplished during < 1.5 hours on a NVIDIA A100 GPU with 80 GB memory at the UBELIX HPC cluster maintained by the University of Bern. The image tiles from the test split were used for the validation of every 10 epochs. By default, the learning rate increased linearly from zero to 1e-5 over the first 10 epochs, and then decreased by a factor of 10 every ten epochs over the last 50 epochs. The loss function was the default Cellpose segmentation loss, which is the mean squared error between the 2D flows (in the XY plane; Pachitariu et al., 2025). This error is calculated for the ground truth and the predicted flows, to which the cross-entropy between the probabilities of the ground-truth and predicted objects is added. During the training, the images were randomly flipped, rotated and resized using a uniformly distributed scaling factor between 0.5 and 1.5 before they were randomly cropped to 224×224 pixels. By default, all image tiles were normalized to image intensity percentiles between 1 and 99 for each channel, and the AdamW optimizer (Loshchilov and Hutter, 2019) together with a weight decay factor of 0.1.

Upon evaluating the model on 2D images, we employed a block tile size of 256 pixels, a fractional tile overlap of 0.1, a threshold value of 0.0 for the object probability, and 0.4 for the flow error, respectively. Again, the image intensity was normalized to the percentile range between 1 to 99 for each input channel. All of these values were default values of the algorithm. Contrary to previous Cellpose versions, no rescaling of image tiles was applied. For 3D segmentation, we used the dedicated 3D approach of Cellpose (Stringer et al., 2021), which computes 2D flows and probabilities for slices in the XY, YZ, and XZ planes. The resulting values are then averaged to create 3D flow vectors. For the construction of 3D segmentation masks, which themselves are generated from the 3D flow vectors, we used the default computation, which considers a 3D smoothing factor of 1.0. For our 3D segmentation demonstration, we used a stack of 400 TIFF images generated with XR-CT from a drill core (from site 5068_1_C from 4-5m depth) of glacio-fluvial sediment infill in a glacially

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

over-deepened valley in southern Germany (Schuster et al., 2024). For details on the XR-CT scanning and image reconstruction, we refer to Schuster et al. (2025).

2.4 Other methods and models

We explored how well our approach compares to other methods that were publicly available and that were either used as a foundation model for general object detection or that were specifically tailored to segment grains. In particular, we first compared the outcome of our segmentation with that of the SAM (Vit-H) in its basic mask generator configuration. We viewed the performance of SAM as a baseline benchmark that every dedicated method should exceed, due to its segmentation capability on a broad range of image datasets and object types (Kirillov et al. 2023) without fine-tuning for specific data, such as sediment grains. Next, we compared our results to the results of Segmenteverygrain that uses prompt engineering for improving segmentations by SAM (Sylvester et al., 2025). Here, we used both their default prompt engineering model and one that we trained on our IG2 dataset. Finally, to evaluate the relative improvement in segmentation performance with our new default model, we compared its segmentation results with those of the best performing model of Mair et al. (2024). Note that we did not include the methods of Mörtl et al. (2022), Chen et al. (2024, or Soloy et al. (2020), because their models or code were not publicly available. Furthermore, we did not include the method of Chen et al. (2022) because of its relatively weak performance in previous studies (Mair et al., 2024). In a second step, we tested our default model's ability to generalize to data not used during training with a setup where the S1_2 and PR subsets were not used during training. We selected these two subsets for this test because for these subsets the performance of both our fine-tuned Cellpose-SAM model and most benchmark models was highest amongst all subset with heterogeneous image tiles of fluvial pebbles under natural conditions. Hence, we anticipated the largest impact on the segmentation performance if we left these out these data from the training split. Particularly, we compared the performance of our default model to that of all other models including specialized Cellpose v2 models, which were trained only on subset datasets used in this generalization. In the following section, we briefly describe how we set up all benchmark models.

2.4.1 The Segment Anything Model (SAM)

SAM is a foundation segmentation model with a vision model transformer (Dosovitskiy et al., 2020; Li et al., 2022). SAM itself was pre-trained with images from a large dataset of annotated images (11 million images with over 1 billion annotation masks; SA-1B) that was created with a custom data engine (Kirillov et al., 2023). This model can thus be used for segmenting a broad range of objects, and it is adaptable to more specific requirements related to various downstream tasks via prompt engineering, inspired by similar advances in Natural Language Processing (Brown et al., 2020). The model itself consists of an image encoder, a mask decoder for inference, and a prompt encoder that is employed for flexible and prompt handling (Kirillov et al., 2023). We used the default model checkpoint (ViT-H) together with its default mask decoder to predict grain masks. For this zero-shot instance segmentation (i.e., segmenting objects in images that had not been included upon training the model), the model generates a grid of point prompts, for which it then filters low quality and duplicate

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

masks. We used the predictions resulting from this model setup as baseline benchmark because they are based on data that is openly available and these predictions can be achieved without any fine-tuning or supervised training on images that display sediment grains.

2.4.2 Segmenteverygrain

Segmenteverygrain combines SAM (i.e., the ViT-H checkpoint; see Section 2.4.1 for details) with a U-Net style convolutional neural network for the prompt engineering upon segmenting grains in images (Sylvester et al., 2025). The default U-Net model was trained on 66 different images displaying grains. The images themselves were split into 44,533 patches of 256×256 pixels. We used both the default model and a model, which we fine-tuned with the entire IG2 dataset. Here, we employed the default train/test splits of our IG2 dataset and the test split for validation. Aside from this, we used the default configuration and followed the recommendation for fine-tuning Segmenteverygrain (Sylvester et al., 2025). This configuration included the Adam optimizer and image augmentation. We trained the refined model for 500 epochs and set the minimum object size to 15 pixels in order to match similar values of other models. We used the predictions of Segmenteverygrain as benchmark for the approach referred to as prompt-based segmentation.

2.4.3 Cellpose 2 models

To evaluate the impact of considering both expanded datasets and the new backbone architecture, we compared the segmentation results of Cellpose SAM with those of older Cellpose (v2.3) models. We started using the *IG1_full_set* model of Mair et al. (2024), which was trained on their original dataset that roughly comprised a third of the images of the IG2 dataset (i.e., subsets S1_2, APF and FH; see also Section 2.1 for details on the datasets). We then trained a Cellpose 2 model with the same architecture on the full IG2 dataset, using the same hyper-parameters and configuration as in the original publication (Mair et al., 2024). This included training any Cellpose 2 model for 1000 epochs, a learning rate of 0.2 with a step-wise reduction of the learning rate by a factor of two for every 10 epochs during the last 100 epochs and a batch size of 8 single-channel images, thereby employing the default Cellpose implementation for image augmentation. Furthermore, Cellpose models can be trained as a specialist models if fine-tuned to a specific dataset (Stringer et al., 2021; Mair et al., 2024). Therefore, we trained two more Cellpose 2 models on the IG2 data but without S1_2 and PR image tiles. For further fine-tuning, we re-trained them only on the respective subsets S1_2 and PR.

2.5 Evaluating segmentation performance

We quantified the segmentation performance by comparing the predicted grain masks to the best-matching masks in the ground truth labels using the approach of Stringer et al. (2021). This was done by calculating average precision (AP) scores, evaluated at different intersection over union (IoU) thresholds. Specifically, we calculated the IoU metric for each grain mask with its closest ground truth match. We use an IoU threshold of > 0.5 (for AP@0.5), and the increasingly stricter range of 0.5 to 0.9 (to calculate the average of AP values, i.e., mAP) to determine which grains were considered as true positives

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

(TP). Grain masks in the ground truth that were not matched by a predicted mask with an IoU value above the aforementioned thresholds were counted as false negative (FN). Likewise, predicted grains with no corresponding grain mask in the ground truth that met the IoU quality criteria were considered as false positive (FP). The average precision is then calculated as the ratio between TP, and the sum of TP, FN and FP, i.e., $TP/(TP + FN + FP)$.

We used these standard metrics in object detection (e.g., Padilla et al., 2020) to quantify the quality of the segmented grains and to compare the results with those of other methods. We chose to use this object-based metric because it combines both false negative and false positive detections in a single step, making it a more stringent metric than traditional metrics, such as precision, recall, or simple intersection over union (IoU) scores. Furthermore, the average precision and mean average precision scores are ubiquitous metrics used to evaluate object detection models in the field of computer vision, which includes models such as SAM (Kirilov et al., 2023), YOLO (Redmon et al., 2016) or Mask R-CNN (He et al., 2018) among others.

Similar to the approach of Mair et al. (2024), we excluded grains for which the minor axis of a simple ellipsoidal fit was < 8 pixels both in the ground truths and in the segmented grain masks. The reason for this is that for most image types displaying sediments, we find it difficult to consistently distinguish between grains that are smaller than those 8 pixels during image annotation, which might render any predictions of smaller grains unstable. We acknowledge that the value can vary across different image settings, e.g., it is usually easier to identify very small grains in single-channel CT images with a high contrast than in multi-channel color images taken from fluvial sediments with a coarser resolution. This is in line with similar but larger thresholds (i.e., 20 or more pixels) determined by other approaches on similar fluvial sediment imagery (e.g., Chen et al., 2022; Purinton & Bookhagen, 2019; Chan et al., 2025).

3 Results

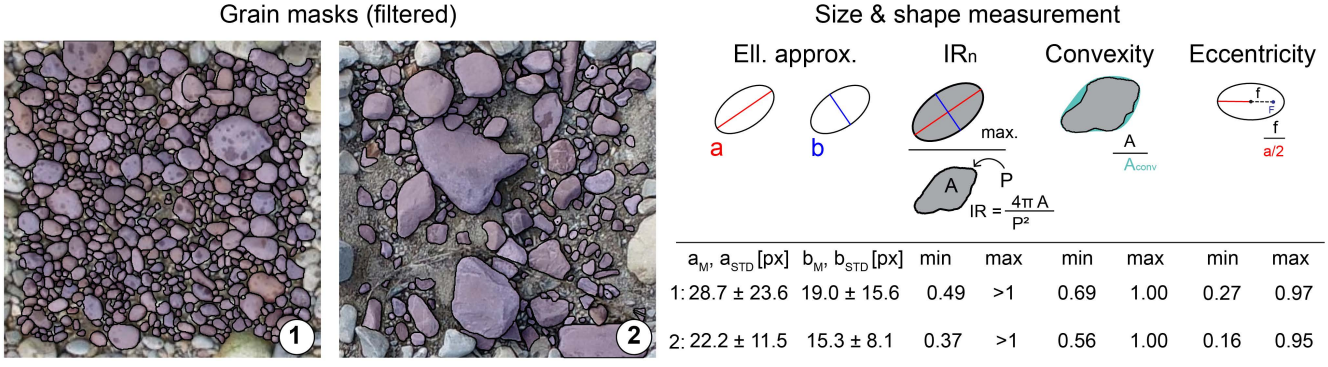
3.1 Grain size and shape in ground truth ROIs

We first calculated standard 2D grain morphometry metrics (Fig. 3) for more than 18,500 manually labeled masks (ROIs) after filtering for minimum grain size and distance to image tile edge, representing about 63% of all labeled ROIs (for train and test split combined; Table S1). The resulting grain sizes vary substantially across the dataset, with b-axis lengths ranging from 8.0 to 481.0 pixels and a-axis lengths from 9.1 to 713.2 pixels (Table S2). The mean grain sizes reflect this variation of more than one order of magnitude, with average b-axis lengths ranging from 10.8 ± 2.7 pixels (CT) to 100.3 ± 111.3 pixels (NZ2). Within each data subset, grain sizes are highly variable, and systematic differences in mean grain size are observed between subsets for both a- and b-axes (Table S2).

The measured grain shapes vary among data subsets. For grain roughness, expressed by the convexity values, the average of the overall dataset is 0.93 ± 0.05, which is consistent across all subsets despite the subsets exhibiting a high within-subset variability (Table S2). For example, convexity shows strong variability in some individual image tiles, with values ranging between 0.62 and 1.00.In general, greater variation is observed for grain roundness values, expressed by the normalized

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

isoperimetric ratio (IR_n, or circularity). Whereas the respective values generally range from 0.38 to >1.0, the average IR_n values across data subsets range from 0.83 ± 0.07 (DV_4) to 0.97 ± 0.04 (CT), indicating systematic differences in roundness between subsets (Table S2). The average value of IR_n of 0.89 ± 0.09 calculated in image tiles basis also reflects this broad variability. The data representing the grain elongation shows the largest variability with eccentricity values for grain ellipse-approximations ranging from 0.25 to 0.97. Despite this broad range, the average eccentricity values across the data subsets are consistent showing an average of 0.73 ± 0.14, again demonstrating a strong within-subset variation (Table S2).



	a_M, a_{STD} [px]	b_M, b_{STD} [px]	min	max	min	max	min	max
1:	28.7 ± 23.6	19.0 ± 15.6	0.49	>1	0.69	1.00	0.27	0.97
2:	22.2 ± 11.5	15.3 ± 8.1	0.37	>1	0.56	1.00	0.16	0.95

Figure 3: Selected grain size (with mean and 1 sigma standard deviation values), and shape measurements based on grain mask area and outlines, displayed here for two examples of annotated images tiles from subset APF_2. Ell. approx. = Ellipse approximation, IR_n = normalized isoperimetric ratio (IR_n, or circularity; Pokhrel et al., 2024), px = pixel.

3.2 Segmentation performance

Overall, our default segmentation model achieves high accuracy in grain segmentation across all image types and most subsets (AP@0.5 > 0.6; mAP > 0.5 for train and test splits combined; Table 1; Fig. 4). Compared with alternative approaches, Cellpose-SAM consistently outperforms all other models if applied to both the full IG2 dataset and across subsets (Table 1). This advantage is maintained in both training and test splits (Table S3). Specifically, the median AP@0.5 across all test image tiles is 18% higher than that of the second-best method (0.71 vs. 0.32 for Cellpose 2 trained on IG2; Fig. 4a; Table S3), and 20% higher across all image tiles (0.72 vs. 0.52 for Cellpose 2 trained on IG2; Table 1). The performance of the fine-tuned Cellposed-SAM model remains robust even for challenging image tiles, with almost no prediction-scoring AP@0.5 values below 0.4 (Figs. 4a, S1). Upon comparing the performance of the methods other than our fine-tuned Cellpose-SAM, three observations can be made. First, the second-best model (using both test and full datasets) were trained with the IG2 dataset. Second, without fine-tuning to IG2, both Cellpose-SAM and Segmenteverygrain perform poorly (Figs. 4a, S1), which is expected since they were fine-tuned to different image data (see Section 2.4). Finally, SAM achieves moderate performance without fine-tuning, comparable to some other methodsin specific subsets (Table 1), highlighting its out-of-the-box segmentation capability. However, it does not match the performance of the best-performing model.

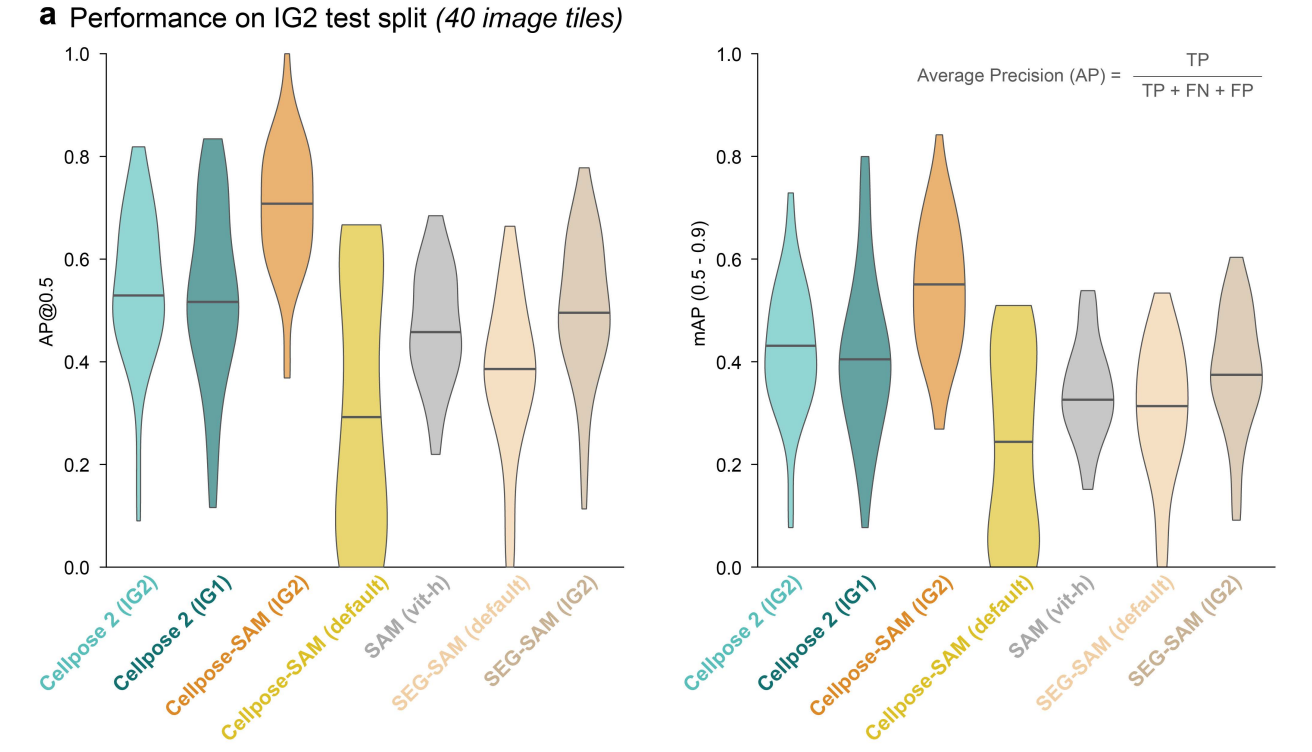
https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

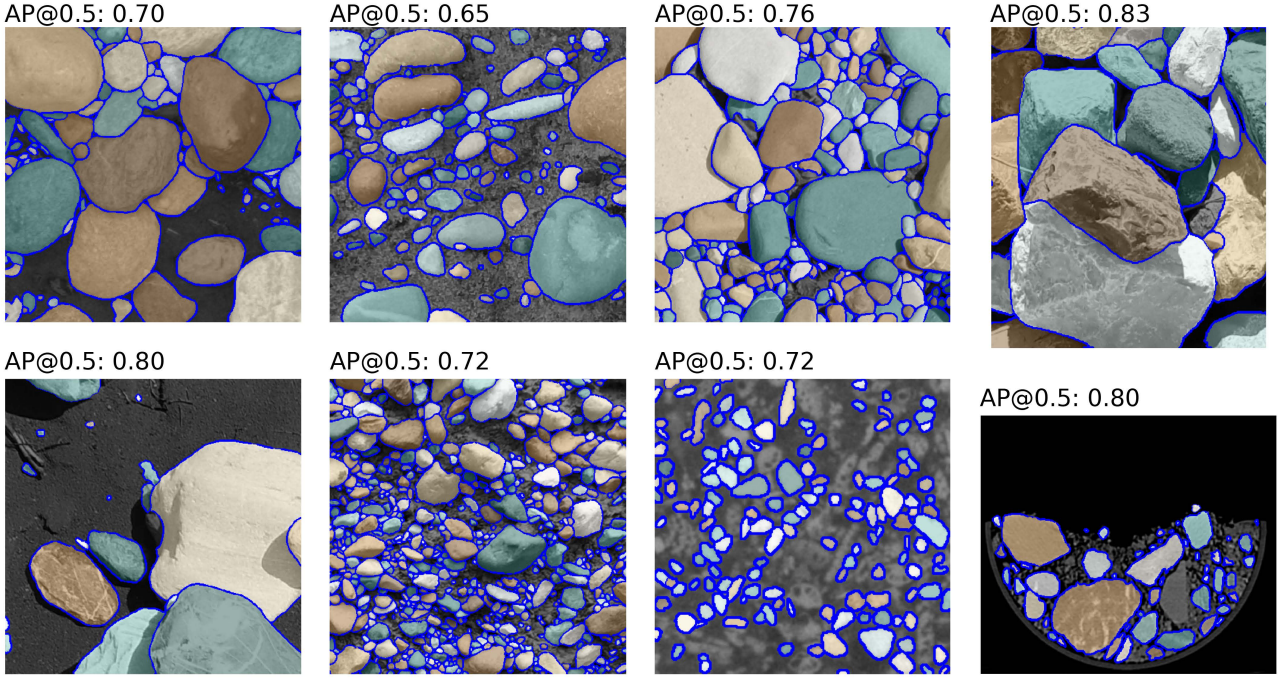
We next evaluated the generalization capability of segmenting grains on image subsets that were not included in the fine-tuning, specifically S1_2 and PR (Fig. 5). Here, a Cellpose-SAM model trained on all IG1 image tiles except S1_2 and PR outperforms all other methods that were not fine-tuned on these subsets (Figs. 5 a, b). Notably, for both subsets, this model achieves a performance that is comparable to some models that were trained with images from the respective subsets, including both versions of Segmenteverygrain and both versions of Cellpose 2 in PR (Figs. 5 a, b). Across both subsets, the overall best-performing model remains Cellpose-SAM that was trained on the full IG2 dataset (Tables 1, S2; Fig. 5). However, for S1_2, the older Cellpose 2 architecture trained as a specialist model achieves a performance close to that of the best-performing model (Fig. 5).

Metric	Method/Model	IG2	Data Subset												
		(all)	*S1_2*	*PR*	*NZ2*	*FH_2*	*NZ1*	*NB2*	*APF_2*	*AR*	*JF*	*CT*	*DV_4*	*PP*	*HP*
AP@0.5	Cellpose 2 (IG2)	0.52	0.65	0.66	0.57	0.49	0.42	0.40	0.52	0.44	0.37	0.09	0.52	0.76	0.65
	Cellpose 2 (IG1)	0.50	0.70	0.70	0.46	0.55	0.39	0.40	0.57	0.40	0.40	0.19	0.38	0.66	0.59
	Cellpose-SAM (IG2)	**0.72**	**0.80**	**0.74**	**0.73**	**0.69**	**0.58**	**0.68**	**0.69**	**0.63**	**0.72**	**0.68**	**0.84**	**0.83**	**0.76**
	Cellpose-SAM (default)	0.17	0.38	0.41	0.15	0.15	0.02	0.14	0.22	0.14	0.27	0.60	0.45	0.60	0.43
	SAM (Vit-H)	0.44	0.56	0.49	0.44	0.48	0.37	0.37	0.43	0.38	0.42	0.54	0.49	0.60	0.54
	SEG-SAM (default)	0.33	0.47	0.52	0.40	0.34	0.18	0.32	0.32	0.30	0.26	0.24	0.21	0.23	0.53
	SEG-SAM (IG2)	0.51	0.58	0.68	0.63	0.58	0.44	0.41	0.48	0.50	0.34	0.07	0.56	0.79	0.71
mAP	Cellpose 2 (IG2)	0.38	0.49	0.57	0.47	0.36	0.30	0.31	0.38	0.33	0.27	0.06	0.40	0.54	0.46
	Cellpose 2 (IG1)	0.37	0.53	0.62	0.39	0.41	0.29	0.30	0.42	0.28	0.30	0.15	0.25	0.46	0.41
	Cellpose-SAM (IG2)	**0.55**	**0.62**	**0.63**	**0.58**	**0.50**	**0.42**	**0.52**	**0.51**	**0.49**	**0.51**	**0.54**	**0.74**	**0.60**	**0.55**
	Cellpose-SAM (default)	0.12	0.27	0.35	0.13	0.12	0.01	0.10	0.17	0.09	0.16	0.41	0.34	0.42	0.31
	SAM (Vit-H)	0.31	0.41	0.41	0.35	0.34	0.25	0.25	0.30	0.26	0.29	0.42	0.36	0.45	0.38
	SEG-SAM (default)	0.26	0.38	0.47	0.35	0.27	0.13	0.25	0.26	0.24	0.20	0.18	0.19	0.18	0.39
	SEG-SAM (IG2)	0.38	0.43	0.55	0.50	0.43	0.28	0.30	0.35	0.37	0.23	0.05	0.43	0.57	0.48

Table 1: Segmentation performance of all methods and models for the IG2 dataset (test and train splits combined), and its subsets with the best performing model indicated in bold. All values are mean AP@0.5 or mAP values for all image tiles in the respective subsets, while for the entire dataset (IG2 –all), we report the median performance across all image tiles. Please note that values for IG2 (all) are calculated on an image basis and therefore they are not the average of the respective values reported for the data subsets on the right.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026





https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Figure 4: Segmentation results for the IG2 test split calculated on an image-tile-basis with performance of the tested methods methods (a), and examples of predicted grain masks (b). AP@0.5 = average precision evaluated at the intersection-over-union (IoU) threshold of 0.5; mAP = mean average precision for IoU thresholds ranging from 0.5 to 0.9; TP = true positive, FP = false positive, FN = false negative. SEG-SAM = Segmenteverygrain.

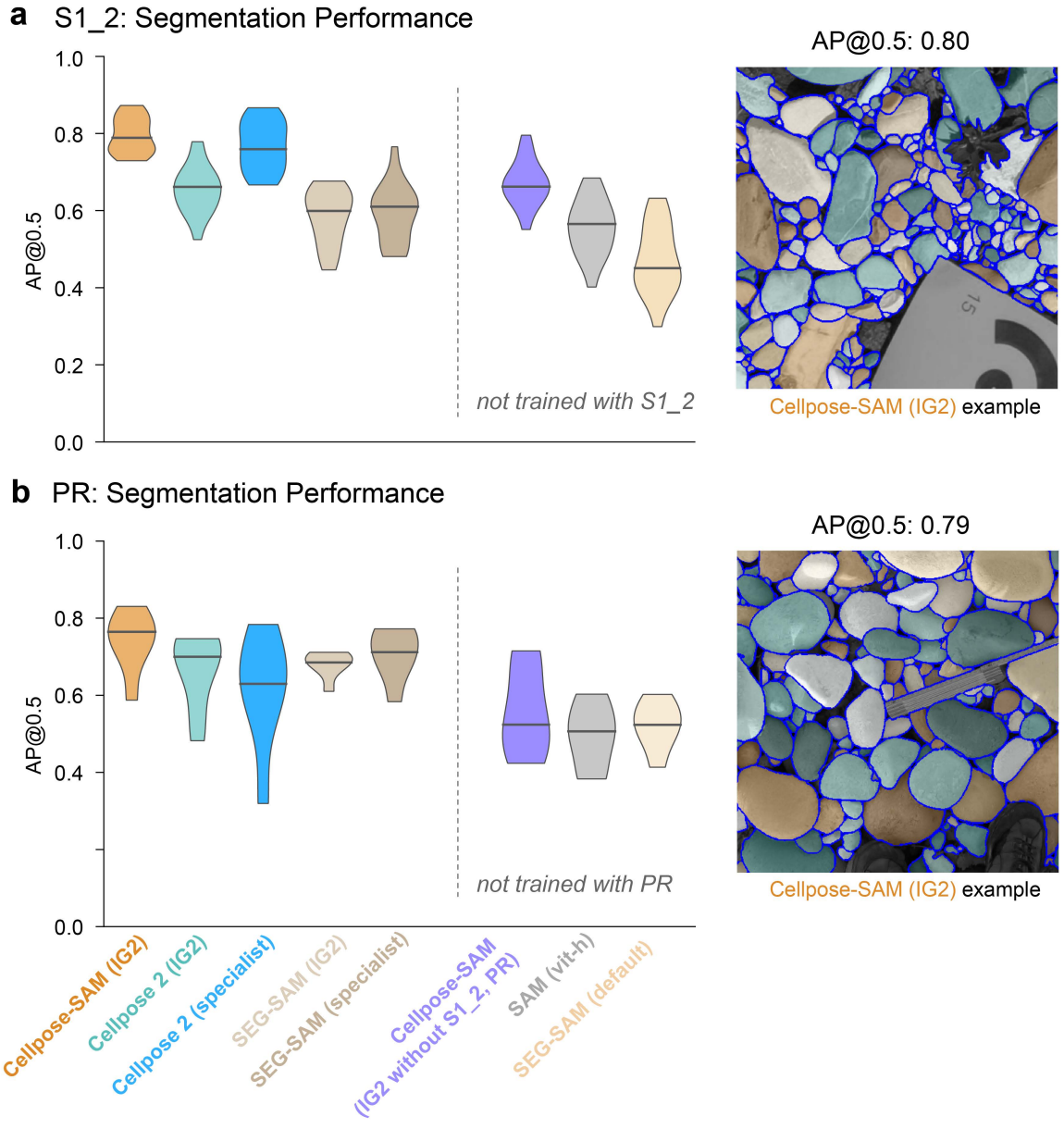


Figure 5: Segmentation results for the S1_2 **(a)** and the PR **(b)** subsets. Specialist models were fine-tuned to the respective subset (see Section 2 for details). AP@0.5 = average precision evaluated at intersection over union (IoU) threshold of 0.5; SEG-SAM = Segmenteverygrain.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

3.3 Size and shape accuracy of predicted grains

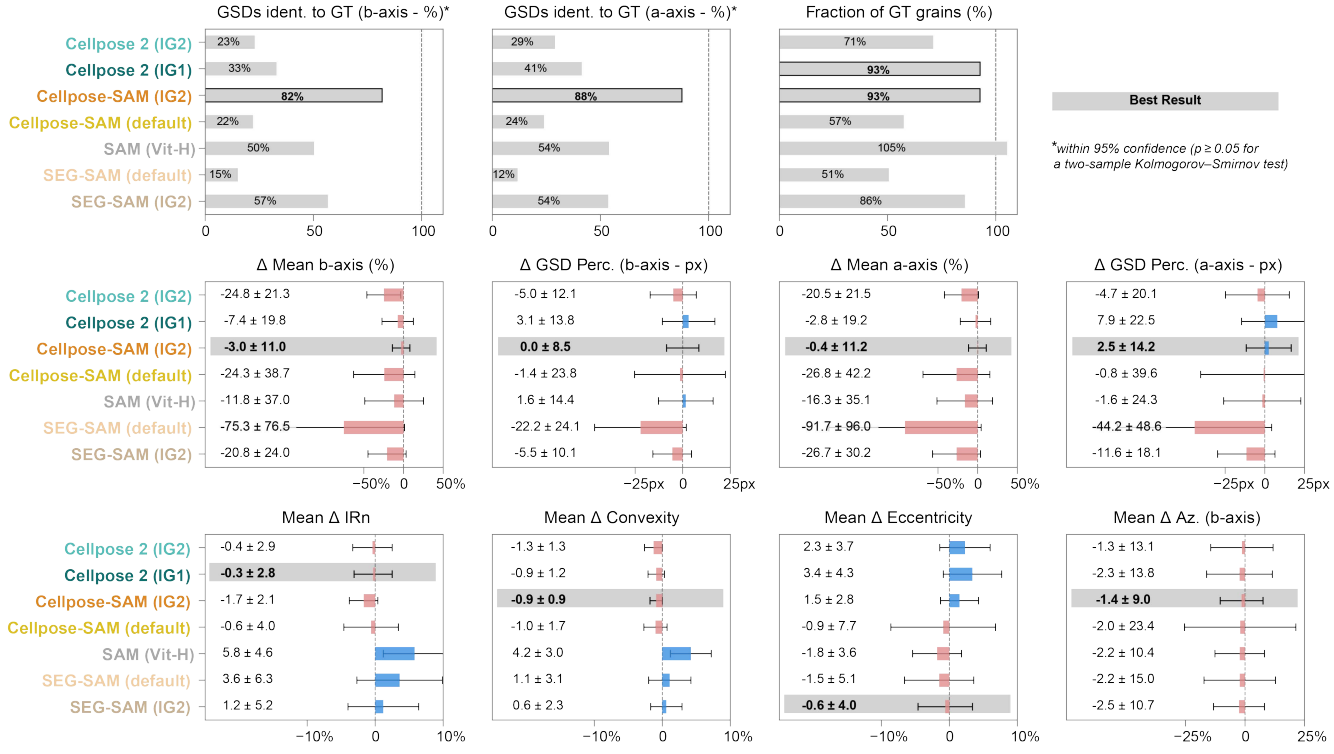
In a second step, we calculated the same 2D metrics (Fig. 3) for predictions from all tested methods and compare them to the ground truth for each image tile. We first evaluate the differences (Δ) between predicted and ground truth ROI masks for each model, aggregated across the full IG2 dataset (Fig. 6). Overall, the predictions derived from our default model (Cellpose-SAM trained on the full IG2 dataset) are the most accurate, showing the closest agreement with the ground truth for 9 out of 11 size and shape metrics (Fig. 6). For the two remaining metrics, mean ΔIR_n and mean ΔEccentricity, our model's predictions are still highly similar to the best-performing alternative models (Fig. 6).

In more detail, the total number of detected grains of our default model is also very close to the ground truth, achieving a 93% recovery rate (Fig. 6). Most notably, mean differences (including the ± one sigma standard deviation range) in grain size are below 12% for both the a- and b-axes, averaged across all detected grains. The resulting grain size distributions (GSDs) , characterized by the lengths of both axes, are statistically identical to the ground truth (within 95% confidence, $p \geq 0.05$ for a two-sample Kolmogorov–Smirnov test) in 88% and 82% of cases for the a- and b-axes, respectively - compared to 54% and 57%, respectively, for the second-best model in this comparison. Our default model also yields the lowest the average percentile differences in grain size with mean Δ values reaching values of 0 (b-axis) and 2.5 (a-axis) pixels, respectively (Fig. 6). For all shape metrics, the differences between predicted and ground truth grains are generally relatively small across all models. For our default model, mean Δ values are consistently below 2% (Fig. 6).

We next examine how the predicted grain masks from our default model compare to the ground truth ROIs across different data subsets. For grain size metrics (mean diameter and percentile differences), most subsets are close to the overall dataset average, with mean Δa- and Δb-axis differences within ±10% (Fig. S2). However, two subsetss show a larger variability: AR and NZ2 have average relative differences of –12.4% and 16.4% for the a-axis, and –14.6% and 11.5% for the b-axis, respectively (Table S4). A similar variability between subsets is also observed for the average percentile differences in some subsets (Fig. S2, Table S4). Consequently, the GSDs for both a- and b-axes are statistically identical to the ground truth in five subsets (100% of image tiles). For another three subsets, the GSDs remains identical for more than 75% of the image tiles (Table S4). For four of the remaining five subsets, the accuracy of the GSD differ particularly when the lengths of the a- and b-axes are considered separately, with 63–100% of GSDs matching the ground truth. Only subset AR shows a lower agreement where 50% of GSDs match the ground truth. Yet this is still comparable to the best results of the second-best method and consistent with the average value of the full IG2 dataset (Fig. 6).

Considering grain shape, deviations in the shape metrics of our default model from the ground truth are generally small. Here, the deviations remain below 5% (Fig. S2) even in subsets with the largest differences. The only notable exception is the mean IR_n value for DV_4, which deviates by more than 10% (Fig. S2; Table S4). Finally, the inter-image variability contributes to higher relative standard deviations for several mean difference values in both grain size and shape metrics, resulting in a broader spread in Fig. S2.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026



375 **Figure 6:** Summary of differences in 2D grain morphometry metrics calculated in relation to manually labelled grain masks across all data subsets. Mean and average standard deviation (1 sigma) values are calculated for image-averaged values. Values for best performance in each metric are indicated in bold. GT = ground truth, GSD = grain size distribution, IRn = normalized isoperimetric ratio (IRn, Pokhrel et al., 2024; Quick et al., 2020). SEG-SAM = Segmenteverygrain.

4 Discussion

380 The results show that our IG2 dataset can be used to successfully train and evaluate deep learning models for segmenting individual sediment grains in a broad variety of images and a range of depositional settings (Section 4.1). By re-training a state-of-the-art segmentation model (Cellpose-SAM) originally developed for bio-medical research with images of sediment grains, we obtain a model that significantly outperforms other current methods for the same task (Section 4.2), despite using the same backbone architecture and starting weights from the Segment Anything Model (SAM). The high-quality
385 segmentation masks generated by our approach allow us to quantify how well grain size and shape are reconstructed relative to the ground truth, which we directly relate to excellent segmentation performance (Section 4.3). Finally, we discuss the limits of our approach and avenues for future development (Section 4.4).

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

4.1 IG2 Dataset composition and characteristics

The IG2 dataset comprises over 29,000 manually annotated 2D masks of individual sediment grains captured in diverse image types, including RGB imagery taken from uncrewed aerial vehicles (UAVs), single-lens reflex cameras, compact digital cameras, and X-ray computed tomography (CT) slices (Table S1). Annotated grains cover a broad range of sedimentary contexts, including fluvial gravels in outcrops (Garefalakis et al., 2023), fluvially transported pebbles, cobbles, and boulders on gravel bars and in river channels (Mair et al., 2022; 2024; Litty and Schlunegger, 2017), bioclastic sands from marine lagoons (Fabbri et al., 2024), and glaciofluvial deposits (Schuster et al., 2025; Hiller et al., 2023).

The IG2 dataset was designed to encompass the broadest possible range of grain types (lithology, shape), bedding characteristics (imbrication, fine-material patches), image acquisition conditions (lighting, shadows, brightness), and background elements or non-grain objects (vegetation, scale objects, water bodies). Non-grain objects were excluded from annotation. Where shadows were present, only visible grain boundaries were traced across shadows. Consequently, the trained Cellpose-SAM model effectively ignores a variety of non-grain features, such as sieves, scales, shoes, ground control point (GCP) markers (e.g., Fig. 5), partial shadows (e.g., Figs. 3, 4), and vegetation that have previously hampered automated grain detection (e.g., Chan et al., 2025; Miazza et al., 2024; Mair et al., 2022). However, the robustness of the grain segmentation may only generalize to objects similar in shape, size-range, and color to those in the training data. Users should also note that grains partially obscured by other objects can introduce biases in quantifying the size and shape of grains, as their reconstructed outlines may deviate from the true boundaries. Yet, this is a problem that is associated with all image-based data collections.

The broad range of imagery and settings resulted in substantial variability in grain size and shape, with a-axis values ranging from 9 to over 700 pixels (Table S2). This order-of-magnitude range exceeds that of many object-detection tasks and has previously hindered a robust segmentation across the full size spectrum (e.g., Chan et al., 2024; Mair et al., 2024). Variations in grain roundness and elongation (Fig. 3, Table S2) surpass those observed along major terrestrial rivers (e.g., Quick et al., 2020; Pokhrel et al., 2024) and even Martian systems (e.g., Szabo et al., 2015). This large range in grain size and shape makes our dataset ideal for evaluating the capability of segmentation models for grain shapes reconstructions. We emphasize that these ground-truth sizes and shapes are not intended to represent specific geomorphic conditions, but rather to serve as a benchmark for evaluating the fidelity of model-predicted masks under highly variable conditions (see Section 3.2; Fig. 3).

Due to the high variability and the relatively small number of 243 image tiles, some dataset imbalance persists between training and test splits. Image tiles were carefully selected to minimize this, but divergent segmentation performance in certain cases (e.g., HP; Table S3) suggests some remaining imbalance in specific subsets.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

4.2 Capabilities of Cellpose-SAM

4.2.1 Segmentation performance and generalization ability

Our results demonstrate that the high segmentation accuracy achieved by the Cellpose-SAM architecture on biomedical images (Pachitariu et al., 2025) can be effectively transferred to the segmentation of sediment grains. On average, our default model that was trained on IG2 correctly segments a larger number of grains and achieves higher precision than all benchmark models, with an average improvement of $\Delta AP@0.5 = 0.18$ compared to the second-best model (Fig. 4; Table 1). It outperforms both earlier Cellpose 2 models, which employed a U-Net backbone, and workflows that also incorporated the SAM, such as SAM itself (ViT-H backbone; Kirilov et al., 2023) and Segmenteverygrain (Sylvester et al., 2025). In contrast, the not fine-tuned Cellpose-SAM model exhibits the lowest performance (Fig. 4), underscoring the effectiveness of re-training and fine-tuning even with comparatively small datasets, such as ours. Notably, training the other benchmark models on the IG2 dataset leads to moderate performance gains for Segmenteverygrain, but little to no improvement for Cellpose 2 (Fig. 4). This suggests that the default Segmenteverygrain model was originally trained on imagery substantially different from our IG2 dataset, whereas the Cellpose 2 model previously used by Mair et al. (2024) seems to have had already reached its performance limit for images of coarse-grained fluvial sediments. Across the benchmark models used in this study, the median segmentation performances of SAM, the trained Segmenteverygrain, and Cellpose 2 were broadly comparable, with differences mainly at the upper and lower end of the performance distributions. However, they did not achieve the same level of segmentation performance as the fine-tuned Cellpose-SAM model (Fig. 4). This outcome aligns with the results of Pachitariu et al. (2025), who demonstrated similar advantages of Cellpose-SAM over other SAM-based architectures (Na et al., 2024; Israel et al., 2024) in biomedical segmentation tasks.

We have evaluated the generalization capability of Cellpose-SAM by training a model that excluded the PR and S1_2 subsets. When compared with Segmenteverygrain and SAM (ViT-H), Cellpose-SAM achieves the highest segmentation performance across both datasets, though with notable differences between them (Fig. 5). For the S1_2 image tiles, the model performs on par with several models were explicitly trained on the S1_2 data. In contrast, performance for the PR images is more variable, with median AP@0.5 values similar to those of SAM and Segmenteverygrain when these models were not trained with the PR data (Fig. 4). Overall, for the PR tiles, the best-performing model remained our default ImageGrains model, which included PR images in its training data.

These results demonstrate the strong generalization capability of Cellpose-SAM, while also indicating that its performance depends, to some extent, on the type of images and how close these are, in regard to both the objects of interest and non-grain objects, to the data that were used during training. Additionally, our findings show that Cellpose 2 models trained as dataset-specific specialists, i.e., on smaller and more homogeneous image sets, can achieve a segmentation accuracy, which is comparable to, or even exceeding that of the more generalist SAM-based architectures. This was particularly evident for the S1_2 data subset (Fig. 5).

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

4.2.2 3D Segmentation

Data about grain size and shape achieved through image-based measurements in 2D have been successfully used to investigate sedimentary systems across a broad range of research (e.g., Garefalakis et al., 2024; Allen et al., 2017; Williams et al., 2013; Marchetti et al., 2022). However, 2D data do not always provide an accurate representation of the size and morphology of grains in 3D (e.g., Garefalakis et al., 2023; Steer et al., 2022; Bunte and Abt, 2001). Consequently, the segmentation of sedimentary grains in 3D remains a critical objective, even though several methodological advancements have been made in recent years to address this challenge (e.g., Rheinwalt et al., 2025; Steer et al., 2022; Walicka and Pfeifer, 2022; Domokos et al., 2024; Kettler et al., 2023). Most of these approaches are designed for segmenting grains from datasets approximating the grains' surfaces in 3D, such as topographic point clouds or meshed grids derived from LiDAR (e.g., Brodu and Lague, 2012) or structure-from-motion (SfM) photogrammetry (e.g., Eltner et al., 2016; Woodget et al., 2018). Grains from such datasets are typically only partially visible due to occlusion (Rheinwalt et al., 2025), which often necessitates the fitting of predefined geometric models during segmentation (e.g., Steer et al., 2022).

In contrast, XR-CT scans provide complete volumetric representations of entire grains (e.g., Cnudde and Boone, 2013). These XR-CT data are analogous to the 3D image stacks of microscopic samples used to train and evaluate the 3D-segmentation capabilities of Cellpose (Stringer et al., 2021) and Cellpose-SAM (Pachitariu et al., 2025). Schuster et al. (2025) demonstrated that models with a Cellpose 2 architecture can be successfully trained to segment coarse grains in 3D XR-CT stacks of images taken from glacio-fluvially transported sediment. In our study, we incorporated the annotated 2D images from Schuster et al. (2025) as subset DV_4, along with annotated micro-XR-CT imagery from Fabbri et al. (2024), to leverage the dedicated 3D capabilities of Cellpose-SAM.

The 3D segmentation of the used example of stacked images yielded 4,647 visually well-defined coarse grains from glaciofluvial diamictic sediment (Fig. 7). Among all IG2 data subsets, the segmentation performance on the 2D image tiles was highest for DV_4, with a mean AP@0.5 of 0.84, whereas the CT dataset achieved an intermediate performance (Table 1). Notably, for the DV_4 images, the new Cellpose-SAM-based approach produced a net increase in mean AP@0.5 of more than 0.2 compared to the Cellpose 2 model of Schuster et al. (2025). These results indicate that the new default model implemented in ImageGrains 2.0 is well suited for such datasets, with its 3D segmentation capability being particularly promising. It should be noted that, ideally, 3D ground-truth labels would be required to rigorously benchmark the segmentation performance in 3D. However, to our knowledge, the manual effort required to annotate large numbers of image slices in 3D XR-CT stacks has so far impeded the creation of such a reference dataset.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

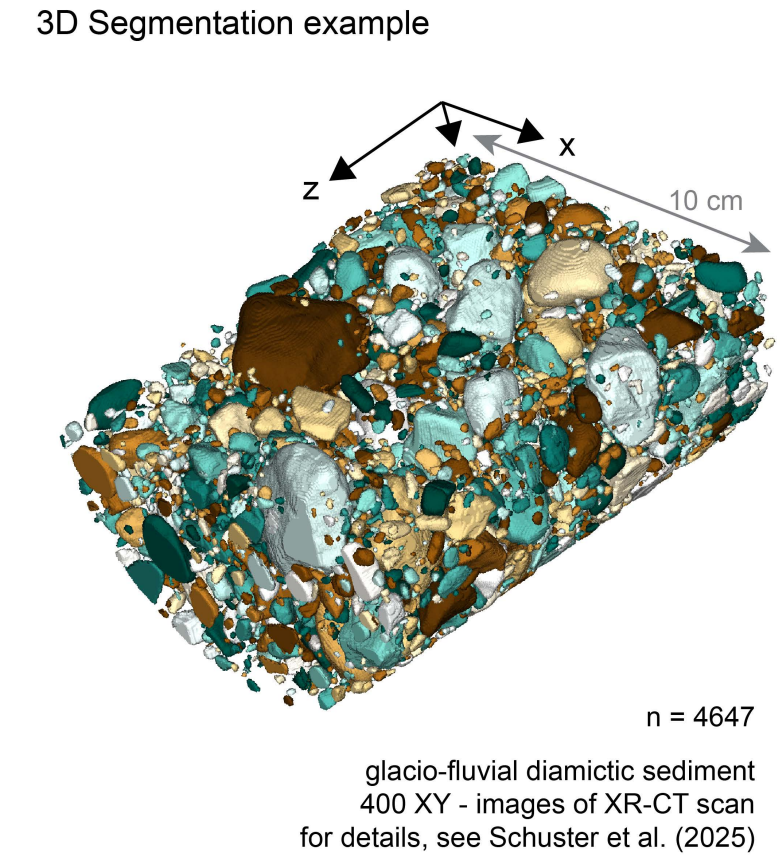


Figure 7: Example where grains were segmented with the default Cellpose-SAM-based segmentation model of ImageGrains 2.0 in 3D from a stack of 400 XR-CT scans taken from of a drill core (drill site 5068_1_C from 4-5m depth; Schuster et al., 2024) made up of coarse-grained glacio-fluvial sediment.

4.3 Relating size and shape accuracy to segmentation performance

Segmentation-based approaches for measuring grain size and shape have long been limited by inaccuracies arising from over-segmentation, under-segmentation, and imprecise grain boundaries (Chardon et al., 2022; Mair et al., 2022; Steer et al., 2022). The introduction of deep-learning models has substantially improved the accuracy and precision of automated grain segmentation in 2D imagery (Mair et al., 2024; Miazza et al., 2024), with recent developments additionally leveraging the capability of SAM(Chan et al., 2025; Sylvester et al., 2025)..

Our new default segmentation model within ImageGrains - a fine-tuned Cellpose-SAM model - further improves both the accuracy and the precision across the entire IG2 dataset, achieving up to 20% improvement in AP@0.5 and mean average precision (mAP) metrics compared to previous models (Section 3.2; Fig. 4). Moreover, the reconstructed grain masks produced by this default model most closely match the ground-truth regions of interest (ROIs) in both size and shape measurements among all benchmark models (Section 3.3; Fig. 6). These results confirm the inferences (e.g., Mair et al.,

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

2024) where an improvement in the segmentation performance results in a more accurate quantification of the size and shape of individual grains.

Despite this strong overall performance, the model yields results with variable quality across data subsets and individual images (Tables 1, S4; Figs. 8, S2). This variability provides an opportunity to assess how informative the model predictions are for individual size and shape metrics (Fig. 8). In this context, we observe the proportion of grain size distributions (GSDs) of predicted grains that are statistically indistinguishable from the ground truth ($p \geq 0.05$; two-sample Kolmogorov–Smirnov test) to increase significantly with segmentation performance for all data subsets (Fig. 8). We observe similar trends in the results of other tested methods, with an even higher statistical significance due to their generally lower and more variable segmentation performance (Fig. S3). Although the relatively large number of statistically indistinguishable GSDs prevents the definition of a AP threshold for perfect correspondence, a 100% match between corresponding grain size distributions in the ground truth and the results, is only achieved for cases with average AP@0.5 values exceeding 0.68 (Figs. 8, S4). For some datasets, higher average AP values yield not necessarily perfect GSD matches, underscoring the importance of the effects related to dataset-specific variabilities; however, no perfect match is achieved below this average AP@0.5 value of 0.68.

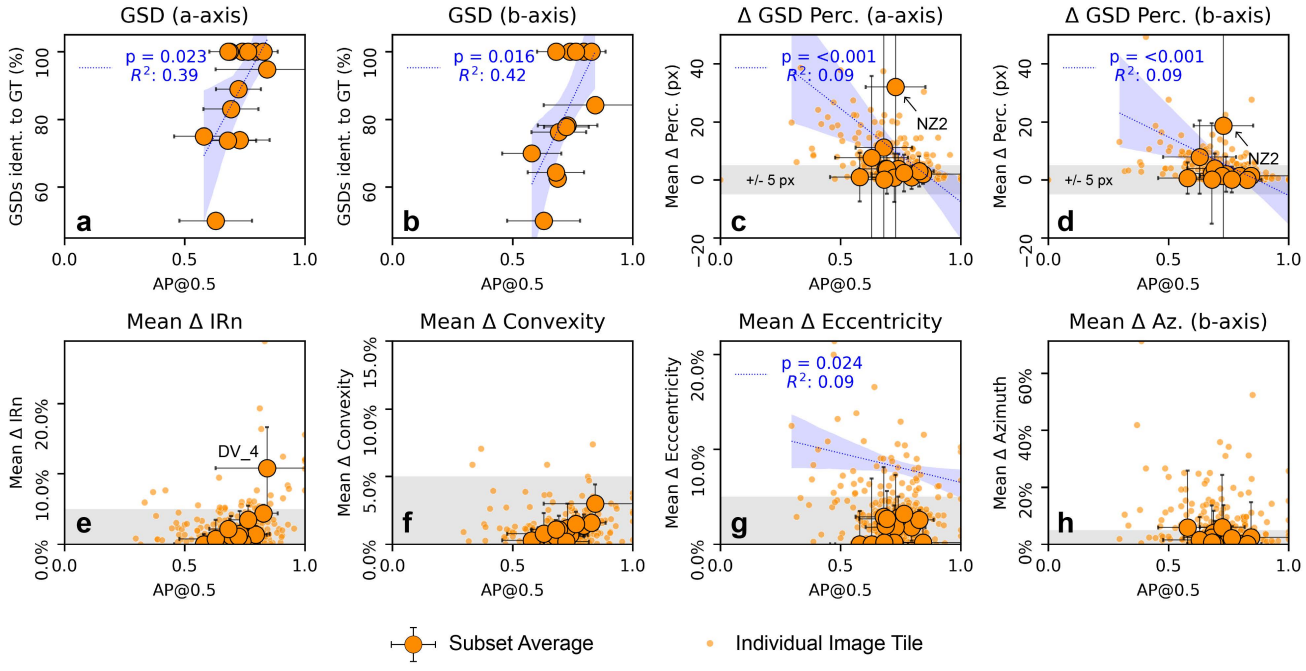


Figure 8: Comparison of segmentation performance metric (AP@0.5) with relative differences in grain size (**a-d**) and shape (**e-h**) between predicted grains and the ground truth ROIs for our default model (Cellpose-SAM). Grey areas indicate very low differences between predicted grain masks and ground truth ROIs, with differences within ± 5 pixels (**c, d**) and < 5% (**e-h**), respectively. Only statistically significant correlations ($p \leq 0.05$, $R^2 \geq 0.05$) for individual images are indicated. For shape metrics (**e-h**) only values with Δ values > 5% were considered for correlation. R^2 = coefficient of determination. Please note that the y-axes in panels **c**, **d** are cropped for a better visualization of the bulk of results.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

The average differences between percentile values of GSDs is generally small, typically <5 pixels (both a-, and b-axes; Table S4) for a majority of the data subsets and for most individual images (Fig. 8). In general, correlations - where significant - indicate that higher AP@0.5 values are associated with smaller differences in percentile-based GSD metrics relative to the ground truth (Figs. 8, S3).

For most grain shape metrics, which include roughness (mean IR_n), roundness (mean convexity), elongation (mean eccentricity), and orientation (mean azimuth), the corresponding differences between predicted masks and the ground truths are minor (<5% on average) for our default model (Fig. 8; see also Fig. 6 and Table S4), both at the data-subset and individual-image levels. We therefore conclude that segmentation performances with AP@0.5 ≥ 0.6 enable sufficiently accurate and precise reconstructions of the grains' shapes for nearly all images, with only a few outliers (Fig. 8). The other tested methods exhibit larger differences between predicted masks and the ground truths, and correlations between segmentation performance and reduced deviations are more frequently detectable (Fig. S3). Overall, the models that were fine-tuned with our dataset tend to yield results that are more reliable in representing the shape than in representing the size of grains.

In summary, the segmentation performance, expressed by AP scores combining false-negative and false-positive detections, correlates with the degree of agreement between predicted and ground-truth grain properties. Therefore, improving segmentation performance indeed reduces the differences between predicted grain masks and ground truths until they are statistically identical. On our data, our default model reaches that performance threshold for some data subsets in all metrics (Fig. 8).

4.4 Limitations, Applicability, and outlook

4.4.1 Limitations of image-based segmentation

Segmentation-derived grain size and shape data from 2D imagery are widely used across geoscientific disciplines, but the 2D nature of images itself imposes some limits on the applicability of segmentation-based approaches. First, image data have size limits for grain detection that are controlled by image resolution and image content. In the IG2 dataset, our default segmentation model applies a minimum size filter of 8 pixels for the minor axis of the fitted ellipses. This threshold is close to the technical lower limit of the Cellpose-SAM backbone, which can theoretically detect circular objects as small as 5 pixels in diameter (Stringer et al., 2021). The reported lower detection limits fro grain segmentation vary considerably among studies, depending on image type and resolution: 6 pixels (Schuster et al., 2025), 12 pixels (Mair et al., 2024), 20 pixels (Chen et al., 2022; Purinton and Bookhagen, 2019), and up to 30 pixels (Chan et al., 2025). At the upper end of the size spectrum, the size of detectable objects is limited by the image extent. By default, Cellpose-SAM detects only objects that are not larger than 40% of the entire image area (Pachitariu et al., 2025). In addition, although the SAM-based architecture can handle a greater range of size variability than previous models (Pachitariu et al., 2025), an extremely large

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

variability in grain size, i.e., spanning more than an order of magnitude, may require combining masks predicted from multiple segmentation runs using differently rescaled images (Chan et al., 2025).

Second, our approach is well suited for segmenting objects in 3D in stacked imagery data (see section 4.2.2 above). However, it is not well suited for segmenting 3D data of surfaces, which are routinely obtained from topographic point clouds. Usually, such data do not contain the complete 3D representation of grains, and therefore require a geometric extrapolation to successfully segment those grains (Steer et al., 2022; Rheinwalt et al., 2025).

Third, some limitations arise from image type and content. While modern deep-learning models, particularly Cellpose-SAM, can be applied to a broad variety of imagery because of the strong capability for generalization (Pachitariu et al., 2025; see also Section 4.2.1), segmentation can be hampered, and, thus, the performance be limited by complex contents in the imagery. Examples include unrelated objects or vegetation, challenging lighting conditions (e.g., shadows, reflections, or glare from water bodies), motion blur, and color imbalance. To counter these effects, we composed the IG2 dataset with images encompassing a broad range of image conditions, thereby enhancing the model's robustness. Consequently, our default model is able to effectively handle a variety of adverse imagery conditions and un-related objects. However, due to the finite dataset size, the performance may decline when applied to image types that are substantially different from those represented in IG2. In such cases, fine-tuning with a small number of additional training tiles (as few as seven or fewer) can yield substantial improvements, as demonstrated for the CT, HP and FH_2 data splits (Table S1). Here, the combination of a deep transformer architecture with SAM's generalist encoder weights effectively reduces the need for extensive dataset balancing, which was essential for earlier, shallower architectures (e.g., Mair et al., 2024).

4.4.2 Hardware requirements

Cellpose-SAM and similar transformer-based architectures require considerably more computational resources and dedicated GPU support than earlier, shallower segmentation models, both during training and inference. Nonetheless, training a Cellpose-SAM model is feasible on a standard desktop equipped with a mid-range GPU, such as an NVIDIA GeForce RTX 3070 with 8 GB of RAM. Under this configuration, training with our dataset required more than 40 hours, compared to under 1.5 hours on an NVIDIA A100 GPU with 80 GB of memory (see Section 2.3 for details). For inference, dedicated GPUs (e.g., NVIDIA or Apple M2 and newer chips) with at least 3 GB of RAM are required to segment large images (> 1000×1000 pixels) within a few seconds. Detailed performance benchmarks are provided in Pachitariu et al. (2025; Tables S2, S3). Measurements of grain size and shape in ImageGrains operate at comparable speeds, enabling the automated analysis of thousands of grains within minutes.

The ImageGrains library is distributed as an installable Python package, allowing both local and cloud-based deployment across platforms (see code availability section below). This enables efficient execution even on free online platforms, such as Google Colab, for users without access to suitable local hardware.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

4.4.3 Applicability

The introduction of a fine-tuned Cellpose-SAM model as the default segmentation engine in ImageGrains 2.0 significantly enhances the performance of segmenting grains relative to previous versions (Mair et al., 2024). This improvement directly benefits from a broad range of applications that rely on the segmentation of grains in 2D, such as conducted in studies of coarse-grained fluvial sediments (e.g., Patel et al., 2025; Rezwan et al., 2025; Zegers et al., 2025). The expanded IG2 dataset enables the application of our default model to additional sedimentary contexts, such as XR-CT imagery of glaciofluvial clasts, bioclastic marine sands and proglacial angular sediments. For CT-based image stacks, the segmentation can be extended to full 3D. The high precision of the predicted grain masks allows for a robust analysis of the shape of grains, with the metrics used in this study provided as default outputs in ImageGrains. Furthermore, the availability of individual 2D grain masks as output enables the analysis of shapes tailored and customized to specific research requirements.

More generally, the high segmentation performance and generalization capability enable robust and precise segmentation of sediment grains in a broad range of image datasets and applications. These outputs could then be used as inputs for other machine learning tasks, e.g., for classification tasks. Furthermore, our publicly available dataset can be used in combination with our own labels for fine-tuning to other image types and settings. Finally, the manually annotated masks for individual grains can be used to train and test other segmentation approaches.

4.4.4 Future directions

This study demonstrates that state-of-the-art deep learning models originally developed for biomedical image segmentation can be successfully adapted for segmenting sediment grains. Similar to the segmentation of biomedical images, domain-specific architectures optimized for grain imagery outperform generic computer vision approaches, despite sharing foundational components such as SAM encoders and pre-trained weights (see Section 4.2.1). This suggests that the same underlying principles apply for these tasks.

However, a notable difference between the two application domains lies in absolute performance of the segmentation. The median AP@0.5 for our full IG2 dataset (0.72) is lower than that reported for Cellpose-SAM on the Cellpose biomedical dataset (> 0.85; Pachitariu et al., 2025). We attribute this difference primarily to the smaller size of the IG2 dataset (243 image tiles compared to over 1000 microscopy images in the Cellpose nuclei dataset), the greater variability of the imagery regarding the size distribution and texture of the grain displayed in this imagery, and the conditions at which they were taken in the field. This interpretation is supported by the variable model performance across IG2 subsets, which correlates with grain size accuracy - and to a lesser extent - with grain shape accuracy (Fig. 8; Table S4). Thus, future improvements in measuring the size and shape of grains will depend strongly on enhancing the performance of segmenting grains.

Progress in this direction is likely to come from the creation of larger and more variable annotated datasets and from the adoption of standardized image acquisition protocols that reduce the variability in image content. Moreover, obtaining annotations from multiple experts for the same images would enable calculation of values reflecting an inter-annotator

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

consensus, a common benchmark for assessing absolute segmentation quality in other fields (e.g., Braylan et al., 2022; Yang et al., 2023; Zhou et al., 2025). Current SAM-based segmentation models appear capable of achieving such inter-annotator consensus–level accuracy for 2D images when trained on suitable datasets (Pachitariu et al., 2025; Na et al., 2025; Israel et al., 2025). This suggests that future advances in grain size and shape measurement will be driven less by architectural refinements but more by the development of larger, high-quality and representative training datasets.

5 Conclusions

We present a collection of 243 manually annotated images of sediment, the IG2 dataset, designed to enable systematic assessment and training of deep-learning architectures for segmenting sediment grains. Furthermore, we introduce ImageGrains 2.0, an updated open-source framework for automated measurement of grain size and shape that integrates a fine-tuned Cellpose-SAM as default segmentation model for this task. Across the IG2 dataset, the model improves segmentation performance by up to 20% in AP@0.5 and mAP compared to previous workflows, yielding grain masks that most closely match the ground-truth regions of interest for both size and shape metrics. The model's strong generalization capability enables accurate segmentation across multiple image types and settings, including XR-CT imagery, and across variable grain textures and imaging conditions. The segmentation performance correlates directly with the accuracy of the derived grain size and shape metrics, confirming that improved segmentation performance translates to more robust geomorphic measurements. Additionally, the model can be fine-tuned with only a few additional image tiles to adapt to new sediment types or imaging conditions, making it applicable for a broad range of applications.

Code availability

All code is available as open-source code in the ImageGrains library (https://github.com/dmair1989/imagegrains), which is also installable as Python package (https://pypi.org/project/imagegrains). A graphical user interface and Jupyter notebooks are provided, enabling the use of ImageGrains 2.0 without the need to write custom code.

Data availability

The image and annotations of the IG2 dataset are available in the dedicated Zenodo repository (Mair et al., 2025a; https://doi.org/10.5281/zenodo.17866827). The model weights for the fine-tuned Cellpose-SAM default segmentation model, and the other Cellpose-2-based models, are available in a separate Zenodo repository (Mair et al., 2025b; https://doi.org/10.5281/zenodo.15309323).

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Author contribution

DM conceptualized the research and developed the code together with GW. The data were curated by DM, with image annotations performed by DM, AdP, AW, BS, and FV, and images contributed by AdP, AW, PG, FV, BS, JÖ, SF, CL, SA, SL, CH, and FS. DM interpreted the results with scientific inputs from GW, MH, and FS. DM prepared the manuscript and figures with contributions from all authors.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The images for NB2 are part of a larger dataset, not yet published in full, whose collection was funded by EU Horizon 2020 grant No. 860383. JÖ acknowledges funding from the MBIE Science Whitinga Fellowship (21-VUW-029), which supported fieldwork in Canterbury, New Zealand. SA, SL and CH acknowledge the funding by the Austrian Academy of Sciences, Earth System Sciences research initiative (ESS), Hidden.ice project in which data for subset JF (JamtalFerner) was obtained. We acknowledge access to high-performance GPUs for model training through the UBELIX HPC cluster maintained by the University of Bern.

References

Allen, P. A., Michael, N. A., D'Arcy, M., Roda-Boluda, D. C., Whittaker, A. C., Duller, R. A., and Armitage, J. J.: Fractionation of grain size in terrestrial sediment routing systems, Basin Research, 29, 180–202, https://doi.org/10.1111/bre.12172, 2017.

Arzt, M., Deschamps, J., Schmied, C., Pietzsch, T., Schmidt, D., Tomancak, P., Haase, R., and Jug, F.: LABKIT: Labeling and Segmentation Toolkit for Big Image Data, Front Comput Sci, 4, https://doi.org/10.3389/fcomp.2022.777728, 2022.

Azzam, F., Blaise, T., and Brigaud, B.: Automated petrographic image analysis by supervised and unsupervised machine learning methods, Sedimentologika, 2, https://doi.org/10.57035/journals/sdk.2024.e22.1594, 2024.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Back, A. L., Kana Tepakbong, C., Bédard, L. P., and Barry, A.: From rocks to pixels: a comprehensive framework for grain shape characterization through the image analysis of size, orientation, and form descriptors, Front Earth Sci (Lausanne), 13, https://doi.org/10.3389/feart.2025.1508690, 2025.

Benet, D., Costa, F., Widiwijayanti, C., Pallister, J., Pedreros, G., Allard, P., Humaida, H., Aoki, Y., and Maeno, F.: VolcAshDB: a Volcanic Ash DataBase of classified particle images and features, Bull Volcanol, 86, https://doi.org/10.1007/s00445-023-01695-4, 2024.

Braylan, A., Alonso, O., and Lease, M.: Measuring Annotator Agreement Generally across Complex Structured, Multi-object, and Free-text Annotation Tasks, in: WWW 2022 - Proceedings of the ACM Web Conference 2022, 1720–1730, https://doi.org/10.1145/3485447.3512242, 2022.

Brayshaw, D. D.: Bankfull and effective discharge in small mountain streams of British Columbia, https://doi.org/10.14288/1.0072555, 2012.

Brodu, N. and Lague, D.: 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology, ISPRS Journal of Photogrammetry and Remote Sensing, 68, 121–134, https://doi.org/10.1016/j.isprsjprs.2012.01.006, 2012.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D.: Language Models are Few-Shot Learners, 2020.

Bunte, K. and Abt, S. R.: Sampling Surface and Subsurface Particle-Size Distributions in Wadable Gravel-and Cobble-Bed Streams for Analyses in Sediment Transport, Hydraulics, and Streambed Monitoring, 2001.

Buscombe, D.: Transferable wavelet method for grain-size distribution from images of sediment surfaces and thin sections, and other natural granular patterns, Sedimentology, 60, 1709–1732, https://doi.org/10.1111/sed.12049, 2013.

Buscombe, D.: SediNet: a configurable deep learning model for mixed qualitative and quantitative optical granulometry, Earth Surf Process Landf, 45, 638–651, https://doi.org/10.1002/esp.4760, 2020.

Butler, J. B., Lane, S. N., and Chandler, J. H.: Automated extraction of grain-size data from gravel surfaces using digital image processing, Journal of Hydraulic Research, 39, 519–529, https://doi.org/10.1080/00221686.2001.9628276, 2001.

Carbonneau, P. E., Lane, S. N., and Bergeron, N. E.: Catchment-scale mapping of surface grain size in gravel bed rivers using airborne digital imagery, Water Resour Res, 40, https://doi.org/10.1029/2003WR002759, 2004.

Chan, V., Rheinwalt, A., and Bookhagen, B.: OrthoSAM: Multi-Scale Extension of the Segment Anything Model for River Pebble Delineation from Large Orthophotos, https://doi.org/10.5194/egusphere-2025-4003, 29 August 2025.

Chardon, V., Piasny, G., and Schmitt, L.: Comparison of software accuracy to estimate the bed grain size distribution from digital images: A test performed along the Rhine River, River Res Appl, 38, 358–367, https://doi.org/10.1002/rra.3910, 2022.

Chen, X., Hassan, M. A., and Fu, X.: Convolutional neural networks for image-based sediment detection applied to a large terrestrial and airborne dataset, Earth Surface Dynamics, 10, 349–366, https://doi.org/10.5194/esurf-10-349-2022, 2022.

Chen, Y., Bao, J., Chen, R., Li, B., Yang, Y., Renteria, L., Delgado, D., Forbes, B., Goldman, A. E., Simhan, M., Barnes, M. E., Laan, M., McKever, S., Hou, Z. J., Chen, X., Scheibe, T., and Stegen, J.: Quantifying Streambed Grain Size, Uncertainty, and Hydrobiogeochemical Parameters Using Machine Learning Model YOLO, Water Resour Res, 60, https://doi.org/10.1029/2023WR036456, 2024.

Chen, Z., Scott, C., Keating, D., Clarke, A., Das, J., and Arrowsmith, R.: Quantifying and analysing rock trait distributions of rocky fault scarps using deep learning, Earth Surf Process Landf, 48, 1234–1250, https://doi.org/10.1002/esp.5545, 2023.

Detert, M. and Weitbrecht, V.: Automatic object detection to analyze the geometry of gravel grains – a free stand-alone tool, in: River Flow 2012, 595–600, 2012.

DiBiase, R. A., Lamb, M. P., Ganti, V., and Booth, A. M.: Slope, grain size, and roughness controls on dry sediment transport and storage on steep hillslopes, J Geophys Res Earth Surf, 122, 941–960, https://doi.org/10.1002/2016JF003970, 2017.

Domokos, G., Jerolmack, D. J., Sipos, A. Á., and Török, Á.: How river rocks round: Resolving the shape-size paradox, PLoS One, 9, https://doi.org/10.1371/journal.pone.0088657, 2014.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.

Eltner, A., Kaiser, A., Castillo, C., Rock, G., Neugirg, F., and Abellán, A.: Image-based surface reconstruction in geomorphometry-merits, limits and developments, Earth Surface Dynamics, 4, 359–389, https://doi.org/10.5194/esurf-4-359-2016, 2016.

von Eynatten, H., Tolosana-Delgado, R., and Karius, V.: Sediment generation in modern glacial settings: Grain-size and source-rock control on sediment composition, Sediment Geol, 280, 80–92, https://doi.org/10.1016/j.sedgeo.2012.03.008, 2012.

Fabbri, S. C., Sabatier, P., Paris, R., Falvard, S., Feuillet, N., Lothoz, A., St-Onge, G., Gailler, A., Cordrie, L., Arnaud, F., Biguenet, M., Coulombier, T., Mitra, S., and Chaumillon, E.: Deciphering the sedimentary imprint of tsunamis and storms in the Lesser Antilles (Saint Martin): A 3500-year record in a coastal lagoon, Mar Geol, 471, https://doi.org/10.1016/j.margeo.2024.107284, 2024.

Fort, S., Ren, J., and Lakshminarayanan, B.: Exploring the Limits of Out-of-Distribution Detection, https://doi.org/10.48550/arXiv.2106.03004, 2021.

Garefalakis, P., do Prado, A. H., Mair, D., Douillet, G. A., Nyffenegger, F., and Schlunegger, F.: Comparison of three grain size measuring methods applied to coarse-grained gravel deposits, Sediment Geol, 446, https://doi.org/10.1016/j.sedgeo.2023.106340, 2023.

Garefalakis, P., do Prado, A. H., Whittaker, A. C., Mair, D., and Schlunegger, F.: Quantification of sediment fluxes and intermittencies from Oligo–Miocene megafan deposits in the Swiss Molasse basin, Basin Research, 36, https://doi.org/10.1111/bre.12865, 2024.

Gatys, L. A., Ecker, A. S., and Bethge, M.: Image Style Transfer Using Convolutional Neural Networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2414–2423, https://doi.org/10.1109/CVPR.2016.265, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, 2017.

Hendrycks, D. and Dietterich, T.: Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, 2019.

Hiller, C., Leistner, S., Helfricht, K., and Achleitner, S.: Robust estimations of areal grain size distribution from geometric surface roughness in a proglacial outwash area, Geomorphology, 439, https://doi.org/10.1016/j.geomorph.2023.108857, 2023.

Ibbeken, H. and Schleyer, R.: Photo-sieving: A method for grain-size analysis of coarse-grained, unconsolidated bedding surfaces, Earth Surf Process Landf, 11, 59–77, https://doi.org/10.1002/esp.3290110108, 1986.

Israel, U., Marks, M., Dilip, R., Li, Q., Yu, C., Laubscher, E., Iqbal, A., Pradhan, E., Ates, A., Abt, M., Brown, C., Pao, E., Li, S., Pearson-Goulart, A., Perona, P., Gkioxari, G., Barnowski, R., Yue, Y., and Van Valen, D.: CellSAM: A Foundation Model for Cell Segmentation, https://doi.org/10.1101/2023.11.17.567630, 20 November 2023.

Kettler, C., Phillips, E., Pichler, K., Smrzka, D., Vandyk, T. M., and Le Heron, D. P.: 3D macro- and microfabric analyses of Neoproterozoic diamictites from the Valjean Hills, California (United States), Front Earth Sci (Lausanne), 11, https://doi.org/10.3389/feart.2023.929011, 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, 2023.

Lang, N., Irniger, A., Rozniak, A., Hunziker, R., DIrk Wegner, J., and Schindler, K.: GRAINet: Mapping grain size distributions in river beds from UAV images with convolutional neural networks, Hydrol Earth Syst Sci, 25, 2567–2597, https://doi.org/10.5194/hess-25-2567-2021, 2021.

Lepp, A. P., Miller, L. E., Anderson, J. B., O'regan, M., Winsborrow, M. C. M., Smith, J. A., Hillenbrand, C. D., Wellner, J. S., Prothro, L. O., and Podolskiy, E. A.: Insights into glacial processes from micromorphology of silt-sized sediment, Cryosphere, 18, 2297–2319, https://doi.org/10.5194/tc-18-2297-2024, 2024.

Litty, C. and Schlunegger, F.: Controls on pebbles' size and shape in streams of the Swiss alps, Journal of Geology, 125, 101–112, https://doi.org/10.1086/689183, 2017.

Li, Y., Mao, H., Girshick, R., and He, K.: Exploring Plain Vision Transformer Backbones for Object Detection, 2022.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, 2019.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Mair, D.: Dataset and model weights for ImageGrains (Version v1), Zenodo, https://doi.org/10.5281/zenodo.8005771, 2023.
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.
Mair, D., do Prado, A., Garefalakis, P., Wild, A. L., Ville, F., Schuster, B., Österle, J., Fabbri, S. C., Litty, C., Achleitner, S., Leistner, S., Hiller, C., & Schlunegger, F.: ImageGrains 2.0 dataset (Version v2), Zenodo. https://doi.org/10.5281/zenodo.17866827, 2025a.
Mair, D., do Prado, A., Garefalakis, P., Wild, A. L., Ville, F., Schuster, B., Österle, J., Fabbri, S. C., Litty, C., Achleitner, S., Leistner, S., Hiller, C., & Schlunegger, F.: ImageGrains 2.0 dataset (Version v2), Zenodo. https://doi.org/10.5281/zenodo.17866827, 2025b.
Marchetti, G., Bizzi, S., Belletti, B., Lastoria, B., Comiti, F., and Carbonneau, P. E.: Mapping riverbed sediment size from Sentinel-2 satellite data, Earth Surf Process Landf, 47, 2544–2559, https://doi.org/10.1002/esp.5394, 2022.
Miazza, R., Pascal, I., and Ancey, C.: Automated grain sizing from uncrewed aerial vehicles imagery of a gravel-bed river: Benchmarking of three object-based methods, Earth Surf Process Landf, 49, 1503–1514, https://doi.org/10.1002/esp.5782, 2024.
Miller, K. L., Szabó, T., Jerolmack, D. J., and Domokos, G.: Quantifying the significance of abrasion and selective transport for downstream fluvial grain size evolution, J Geophys Res Earth Surf, 119, 2412–2429, https://doi.org/10.1002/2014JF003156, 2014.
Mörtl, C., Baratier, A., Berthet, J., Duvillard, P. A., and De Cesare, G.: GALET: A deep learning image segmentation model for drone based grain size analysis of gravel bars, in: Proceedings of the IAHR World Congress, 5326–5335, https://doi.org/10.3850/IAHR-39WC2521716X2022895, 2022.
Na, S., Guo, Y., Jiang, F., Ma, H., Gao, J., and Huang, J.: Segment Any Cell: A SAM-Based Auto-Prompting Fine-Tuning Framework for Nuclei Segmentation, IEEE Trans Neural Netw Learn Syst, 1–10, https://doi.org/10.1109/TNNLS.2025.3611322, 2025.
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.
Padilla, R., Netto, S. L., and da Silva, E. A. B.: A Survey on Performance Metrics for Object-Detection Algorithms, in: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 237–242, https://doi.org/10.1109/IWSSIP48289.2020.9145130, 2020.
Patel, N. K., Schlunegger, F., Mair, D., Pati, P., do Prado, A. H., Garefalakis, P., and Choudhury, R. K.: Source-to-sink patterns of grain size along the Yamuna River in the Indian Himalaya, Geomorphology, 486, 109909, https://doi.org/10.1016/j.geomorph.2025.109909, 2025.
Pokhrel, P., Attal, M., Sinclair, H. D., Mudd, S. M., and Naylor, M.: Downstream rounding rate of pebbles in the Himalaya, Earth Surface Dynamics, 12, 515–536, https://doi.org/10.5194/esurf-12-515-2024, 2024.
Prieur, N. C., Amaro, B., Gonzalez, E., Kerner, H., Medvedev, S., Rubanenko, L., Werner, S. C., Xiao, Z., Zastrozhnov, D., and Lapôtre, M. G. A.: Automatic Characterization of Boulders on Planetary Surfaces From High-Resolution Satellite Images, J Geophys Res Planets, 128, https://doi.org/10.1029/2023JE008013, 2023.
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.
Quick, L., Sinclair, H. D., Attal, M., and Singh, V.: Conglomerate recycling in the Himalayan foreland basin: Implications for grain size and provenance, GSA Bulletin, 132, 1639–1656, https://doi.org/10.1130/B35334.1, 2020.
Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos, 2024.
Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection, 2016.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Rezwan, N., Mair, D., Whittaker, A. C., Schlunegger, F., do Prado, A. H., and Setu, S. H.: Assessing Flood-Influenced Sediment Dynamics Using UAV Photogrammetry and Machine Learning: Insights from River Sense, Switzerland, https://doi.org/10.5194/egusphere-2025-5145, 28 October 2025.

Rheinwalt, A., Purinton, B., and Bookhagen, B.: Curvature-based pebble segmentation for reconstructed surface meshes, Earth Surface Dynamics, 13, 923–940, https://doi.org/10.5194/esurf-13-923-2025, 2025.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/arXiv.1505.04597, 2015.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J. Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., and Cardona, A.: Fiji: An open-source platform for biological-image analysis, https://doi.org/10.1038/nmeth.2019, July 2012.

Schuster, B., Gegg, L., Schaller, S., Buechi, M. W., Tanner, D. C., Wielandt-Schuster, U., Anselmetti, F. S., and Preusser, F.: Shaped and filled by the Rhine Glacier: the overdeepened Tannwald Basin in southwestern Germany, Scientific Drilling, 33, 191–206, https://doi.org/10.5194/sd-33-191-2024, 2024.

Schuster, B., Mair, D., Schmid, T. C., Gegg, L., Buechi, M. W., Schaller, S., Anselmetti, F. S., and Preusser, F.: Automated X-ray computed tomography-based analysis of clast fabric in drill-cores of glacial diamicts using deep learning models, Boreas, https://doi.org/10.1111/bor.70023, 2025.

Sklar, L. S.: Grain Size in Landscapes, Annu Rev Earth Planet Sci, 52, 663–692, https://doi.org/10.1146/annurev-earth-052623-075856, 2024.

Soloy, A., Turki, I., Fournier, M., Costa, S., Peuziat, B., and Lecoq, N.: A deep learning-based method for quantifying and mapping the grain size on pebble beaches, Remote Sens (Basel), 12, 1–23, https://doi.org/10.3390/rs12213659, 2020.

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.

Sylvester, Z., Stockli, D. F., Howes, N., Roberts, K., Malkowski, M. A., Poros, Z., Martindale, R. C., and Bai, W.: Segmenteverygrain: A Python module for segmentation of grains in images, J Open Source Softw, 10, 7953, https://doi.org/10.21105/joss.07953, 2025.

Szabó, T., Domokos, G., Grotzinger, J. P., and Jerolmack, D. J.: Reconstructing the transport history of pebbles on Mars, Nat Commun, 6, https://doi.org/10.1038/ncomms9366, 2015.

Walicka, A. and Pfeifer, N.: Automatic Segmentation of Individual Grains From a Terrestrial Laser Scanning Point Cloud of a Mountain River Bed, IEEE J Sel Top Appl Earth Obs Remote Sens, 15, 1389–1410, https://doi.org/10.1109/JSTARS.2022.3141892, 2022.

Van Der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T.: Scikit-image: Image processing in python, PeerJ, 2014, https://doi.org/10.7717/peerj.453, 2014.

Williams, R. M. E., Grotzinger, J. P., Dietrich, W. E., Gupta, S., Sumner, D. Y., Wiens, R. C., Mangold, N., Malin, M. C., Edgett, K. S., Maurice, S., Forni, O., Gasnault, O., Ollila, A., Newsom, H. E., Dromart, G., Palucis, M. C., Yingst, R. A., Anderson, R. B., Herkenhoff, K. E., Le Mouélic, S., Goetz, W., Madsen, M. B., Koefoed, A., Jensen, J. K., Bridges, J. C., Schwenzer, S. P., Lewis, K. W., Stack, K. M., Rubin, D., Kah, L. C., Bell, J. F., Farmer, J. D., Sullivan, R., Van Beek, T., Blaney, D. L., Pariser, O., Deen, R. G., Kemppinen, O., Bridges, N., Johnson, J. R., Minitti, M., Cremers, D., Edgar, L., Godber, A., Wadhwa, M., Wellington, D., McEwan, I., Newman, C., Richardson, M., Charpentier, A., Peret, L., King, P., Blank, J., Weigle, G., Schmidt, M., Li, S., Milliken, R., Robertson, K., Sun, V., Baker, M., Edwards, C., Ehlmann, B., Farley, K., Griffes, J., Miller, H., Newcombe, M., Pilorget, C., Rice, M., Siebach, K., Stolper, E., Brunet, C., Hipkin, V., Léveillé, R., Marchand, G., Sánchez, P. S., Favot, L., Cody, G., Steele, A., Flückiger, L., Lees, D., Nefian, A., Martin, M., Gailhanou, M., Westall, F., Israël, G., Agard, C., Baroukh, J., Donny, C., Gaboriaud, A., Guillemot, P., Lafaille, V., Lorigny, E., Paillet, A., Pérez, R., Saccoccio, M., Yana, C., Aparicio, C. A., Rodríguez, J. C., Blázquez, I. C., et al.: Martian fluvial conglomerates at gale crater, Science (1979), 340, 1068–1072, https://doi.org/10.1126/science.1237317, 2013.

https://doi.org/10.5194/egusphere-2025-6346
Preprint. Discussion started: 20 January 2026

Woodget, A. S., Fyffe, C., and Carbonneau, P. E.: From manned to unmanned aircraft: Adapting airborne particle size mapping methodologies to the characteristics of sUAS and SfM, Earth Surf Process Landf, 43, 857–870, https://doi.org/10.1002/esp.4285, 2018.

Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., and Antani, S. K.: Assessing Inter-Annotator Agreement for Medical Image Segmentation, IEEE Access, 11, 21300–21312, https://doi.org/10.1109/ACCESS.2023.3249759, 2023.

Zegers, G., Hayashi, M., and Garcés, A.: Distributed estimation of surface sediment size in paraglacial and periglacial environments using drone photogrammetry, Earth Surf Process Landf, 50, https://doi.org/10.1002/esp.70093, 2025.

Zhou, F. Y., Marin, Z., Yapp, C., Zou, Q., Nanes, B. A., Daetwyler, S., Jamieson, A. R., Islam, M. T., Jenkins, E., Gihana, G. M., Lin, J., Borges, H. M., Chang, B.-J., Weems, A., Morrison, S. J., Sorger, P. K., Fiolka, R., Dean, K. M., and Danuser, G.: Universal consensus 3D segmentation of cells from 2D segmented stacks, Nat Methods, 22, 2386–2399, https://doi.org/10.1038/s41592-025-02887-w, 2025.