Geochronology Discuss., author comment AC2
https://doi.org/10.5194/gchron-2022-12-AC2, 2022

Reply on RC2

Michael C. Sitar and Ryan J. Leary

Author comment on "Technical note: colab_zirc_dims: a Google Colab-compatible toolset for automated and semi-automated measurement of mineral grains in laser ablation–inductively coupled plasma–mass spectrometry images using deep learning models" by Michael C. Sitar and Ryan J. Leary, Geochronology Discuss., https://doi.org/10.5194/gchron-2022-12-AC2, 2022

We greatly appreciate T. Scharf's constructive comments and corrections regarding our manuscript and are especially glad that she was able to evaluate our work given her unique expertise on this subject. The comments (italicized) highlight many places where we can improve or clarify our manuscript, and we have included a response below each one. We use "revised manuscript" below to refer to a revised copy of our original manuscript that we will submit if invited to do so by the editors.

- *Line 45-46: Colab_zirc_dims works exclusively on reflected light (RL) images of zircons mounted in resin. The authors mention that zircon shape may be partially obscured by resin, resulting in minimum shape measurements instead of true dimensions. As colab_zirc_dims is presented as a tool for zircon shape measurement, could the authors kindly expand the discussion to cover whether the error introduced by reflected light images is significant, and whether or not it predisposes colab_zirc_dims to certain use cases? For example, do we know what proportion of a dataset is typically affected by this phenomenon? Is there any risk in comparative studies in which mounts have been differently handled (e.g. ground to different depths) or where shape measurements have been extracted from a variety of image types?*

This information is certainly important to convey to colab_zirc_dims users. We believe that a study involving preparing and imaging new mounts using different techniques in order to precisely quantify the potential degree to which it impacts grain shape estimation accuracy would very valuable, especially with growing interest in grain size/shape versus age relationships. Though such a study is obviously beyond the scope of our technical note, we think that the data that we have on hand in some cases approximates the errors that the reviewer comments on. We have consequently added the following text to section 5.2 of our draft revised manuscript:

"Because these images are of sufficiently high quality that subsurface grain extents were interpretable by Leary et al. (in press), and because model M-R101-C generally only segments grain areas above resin surfaces, errors in these samples can also be used as a rough proxy for dimensional data loss from using reflected light versus transmitted light images to measure shapes of very poorly exposed grains in cases where reflected light images do not reveal any information about subsurface grain extents (Sect. 1; Leary et

al., 2020a). In the worst-evaluated sample, 1WM-302 (n=180), M-R101-C produces axial measurements that underestimated manually measured grain axes by at least 20% 58.3% of the time, with average grain measurement errors of -16.1% and -21.0% along long and short axes, respectively. Treating these automatically generated axial measurements as ground truth data could result in significantly flawed analysis of relationships between grain size and age. Such shape parameter underestimates present only a minor (though potentially time-consuming) problem for colab_zirc_dims users with poorly exposed grains whose actual areas are still interpretable by humans (e.g., in the case of 1WM-302); erroneous segmentation masks can simply be corrected manually using the GUI. Users who observe that their mounted crystals are both very poorly exposed and invisible below the resin surface in their reflected light images, though, may consider re-imaging their samples using transmitted light and then measuring grains using a different program (e.g., AnalyZr; Scharf et al., 2022) to avoid collecting flawed data."

To address the question of comparison between facilities and/or datasets, we have added the following text:

"Because most facilities aspire to polish their laser ablation zircon mounts to half the thickness of the zircons, we expect that there will be no systematic differences between measurement of grains analysed at different laser ablation labs. However, because there is some variability in quality of polish this was achieved at ALC in the test dataset (Leary et al., 2020a; see above discussion of samples 1WM-302, 5PS-58, 2QZ-9, and 2QZ-272), careful manual checking of polish quality will be required in any dataset as described above. Ultimately, a study in which pre- (e.g., Finzel, 2017) and post-mount (Leary et al., 2020a; Scharf et al., 2022; current study) grain dimension measurements can be collected on the same samples will be the best way to quantify the bias introduced by polishing and/or by different facilities. However, such a test is well beyond the scope of the current study."

- *Line 72-76: Please include an image that compares the segmentation achieved by traditional methods such as Otsu thresholding against those of colab_zirc_dims, when artefacts are present (e.g. anomalous bright spots, bubbles), to support this assertion. Alternatively, please include supporting literature references.*

We have amended revised versions of Figs. 1 and 2 (see attached) to include selections of images displaying artefacts and fractured grains (e.g., mosaic stitching boundaries, anomalous bright spots, and a bubble), along with basic Otsu thresholding masking results (Fig. 1d) and colab_zirc_dims segmentation and measurement results (Fig. 2d). We hope that these images adequately show differences in performance, which are significant in the cases of some atypical images (e.g., the top row in Figs. 1d and 2d), but minor in others (e.g., the bottom row in Figs. 1d and 2d).

- *Line 88: I am perhaps confused by the term "zonal area" – does this refer to the banding seen in cathodoluminescence (CL) images of zircon? Have the authors used AnalyZr to extract bands from within grains? Unfortunately, as AnalyZr was not developed for CL images, full grain segmentation from CL images is expected to fail. Perhaps reword the text to clarify, as it might mistakenly be interpreted as a recommendation to use AnalyZr for CL image grain segmentation.*

Reply from the first author: Due to a combination of a) my misunderstanding of the intended use of the AnalyZr spot picking/CL texture recording UI in the Scharf et al. (2022) paper and b) a not unsuccessful test of AnalyZr with a CL zircon grain image (loaded as _RL), I was in fact under the mistaken impression that AnalyZr could be used for full grain segmentation in CL images. "Zonal area" is vague but refers here to the CL image zoning textures that users can localize to analytical spots using the UI; I erroneously believed that users could identify and mark these textures in-situ while

segmenting grains in CL images. I apologize for mischaracterizing your work and should note that the Scharf et al. (2022) paper is quite unambiguous with regards to the intended use cases and files for AnalyZr.

To eliminate any possibility of confusion, we have removed "…zonal area from cathodoluminescence images…" from our revised manuscript and added the following text:

"Analytical spot identification and localization in AnalyZr is done manually through an interface which also allows input of spot-specific comments and qualitative internal grain zoning descriptors that persist into the program's exports (Scharf et al., 2022)."

- *Potentially inconsistent terminology. Do the authors intend these terms to have different meanings, or do they all refer to a MaskRCNN implementation with FPN, using a ResNet backbone? If the latter, I'd recommend that terminology be standardised throughout the paper.*

We agree that our terminology here is inconsistent and potentially confusing. These terms are used somewhat interchangeably in deep learning literature and documentation, but the technically correct terminology as per (He et al., 2018) should be "Mask RCNN ResNet-FPN" when referring to specific model designs. We have updated our manuscript for consistency as follows:

- *Mask RCNN FPN (line 139 & 147)*

Changed both to "Mask RCNN" (reference to general model architecture)

- *Mask RCNN (line 151, Fig 2 caption)*

Kept as is due to reference to specific model in Table 1.

- *Mask RCNN ResNet-FPN (line 178)*

Kept as is (already correct).

We have additionally revised all backbone names in Table 1 (see attached) to reflect the fact that all the backbone networks that we used incorporate FPNs, and revised the text of our manuscript to state this specifically where appropriate.

- *Line 155: Table 1: The authors have selected training iterations of 4000-7000. Fig 4 shows that models stabilise at approximately 2000 iterations. Could the authors please include their reasoning for selecting model checkpoints at such high iterations? Is there any risk that these models are comparatively overtrained (meaningfully less generalised) than those around ~2000 iterations (e.g. how do they compare on the test dataset used for Table 3)?*

We agree that these questions should be answered for readers in the manuscript and have consequently revised figure 4 (attached) to include average absolute long axis error from evaluations of model checkpoints saved during training, and revised text at the end of section 5 to read:

"We picked "best" model checkpoints (Table 1) at checkpoints beyond 3000 iterations where models achieved apparent local maxima in validation accuracies (i.e., Fig. 4) and local minima or plateaus in various measurement error metrics (e.g., failure rate and absolute long axis error; Table 3; Fig. 4) when evaluated on the full Leary et al. (in press) test dataset. We set our threshold (greater than 3000 iterations) for checkpoint picking based on qualitative observations that grain masks for all models appeared to be more

"blobby" (i.e., more refined to actual grain areas) at lower training iterations, though it is worth noting that we fail to see conclusive evidence for this relationship in training, validation mask loss, or test accuracy curves (Fig. 4). Changes in most evaluation accuracy metrics (roughly represented by average absolute long axis error in Fig. 4) for the models trained with image augmentations were largely stochastic after ~2000 (for the pretrained models) to ~3000 iterations (for the randomly initialized models; Fig. 4). This suggests a lack of meaningful overfitting (possibly attributable to a combination of learning rate drawdown and training image augmentation) in relation to our test dataset and probable negligible negative effects on model generalization abilities from our selecting models at relatively high training iterations."

For the reviewer's benefit, we would like to note that we do see significant evidence for overfitting against our test dataset for our single model trained without image augmentation (M-R50-S-NA) after ~4000 iterations. Due to much higher overall test dataset error values than for other models, however, we are unable to fit it on the new panel of our revised Fig. 4 without obscuring test results for our better-performing models. We have attached a version of Figure 4 that includes model M-R50-S-NA test results for the reviewer to evaluate – see "revised fig 4 with unaugmented model test results.png".

- *Line 184-185: The meaning of "sample-dependent…resolutions…(194 by 194 pixels…)" was not clear to me. Does this refer to the "max_zircon_size" criterion in the "mosaic_info" data table of the "Data Matching and Preparation" Colab notebook (lines 262-266)? Consider rewording the text to clarify.*

This does indeed refer to the "max_zircon_size" ("Max_grain_size" as of colab_zirc_dims v1.0.9). To clarify our meaning, we have revised the sentence to read:

"ALC images (Table 2) were algorithmically extracted from mosaic images at scales and resolutions (194 by 194 pixels to 398 by 398 pixels) that respectively varied sample-to-sample based on imaging parameters during analysis and grain size (i.e., "Max_grain_size"; Sect. 4.3.1)."

- *Lines 184-188: Could the authors please provide an indication of the nature of these zircon grains (e.g. sources, ages, sedimentary environment, histograms of shape parameter variation etc.) so that the reader has an understanding of how diverse the image test and validation datasets are? As the authors are using a small dataset to train deep convolutional neural networks, a reader may wonder how generalised the trained models are, and whether they will perform as well on zircon grains from different regions. Alternatively, if the authors feel that the small training dataset inhibits generalisation, please expand on this in the discussion.*

We have added the following texts to section 3.2.3 to provide additional information on the provenance of training dataset images:

"These samples contain a wide range of zircon ages from Proterozoic to late Palaeozoic and represent a variety of terrestrial and marine depositional environments including fluvial, delta plain, nearshore, and continental shelf environments. See Leary et al. (2020a; 2020b) for detailed discussion of these samples."

and

"These images are of grains derived from samples of Late Mesozoic-Early Cenozoic rocks interpreted to have been deposited by braided stream systems. Dated zircon grains from these samples indicate the presence of mixed populations of Proterozoic grains that likely record long-range tectonic-fluvial transport (e.g., from the Grenville Orogen to modern

day Nevada, USA; Rainbird et al., 1997; Gehrels et al., 2000) and iterative recycling prior to their most recent deposition. These grains are combined in approximately equal proportions with minimally transported Early Cretaceous grains presumably sourced from the ancient Sierran Arc. Images from the UCSB training set consequently include variable mixtures of very well-rounded and relatively fresh, euhedral grains."

We have added the following text to the end of section 5.2 of our revised manuscript in order to better inform readers on potential uncertainties related to our small training dataset:

"Though most of our models evidentially generalize well to our test set, and we believe that they will most likely generalize well to other datasets, they are still untested on data from facilities not represented in their training dataset (i.e., besides ALC and UCSB). And, though they have been exposed to some relatively euhedral detrital zircon grains in the UCSB training images, our models are notably also untested on crystals derived from primary igneous and volcanic rocks. Some uncertainty remains in how well our models will work when applied to more diverse data by colab_zirc_dims users. Since our training dataset is quite small and lacking in diversity of image sources, increasing the size and diversity of our training dataset before training updated models will likely yield some improvements in model generalization ability. We plan to expand our training dataset and release new models as we maintain colab_zirc_dims and will make this our priority should users inform us that our current models fail to generalize well."

- *Line 190: Kindly indicate which of the images were hand-selected (perhaps rename the image files and refer the reader to the supplementary data).*

We appreciate this suggestion and certainly plan to incorporate it. We are working on a Python script to persist filename changes into training annotation .json files and will get this done in advance of submitting supplementary data for a revised manuscript or possibly sooner (e.g., for colab_zirc_dims 1.0.10).

- *Lines 202-204: Please clarify what an iteration refers to, in this training regime. Additionally, consider specifying epochs and batch size in Table 1, for those readers who may wish to test the reported training strategy within their own Python implementation of MaskRCNN.*

We are glad that the reviewer pointed this out – a Google search reveals several completely different definitions of "iteration" in the context of deep learning, which will surely be confusing to readers. We have added the following text to clarify what the term means for Detectron2 training:

"Detectron2 exclusively uses the term "iteration" to define the extent of a model's exposure to training data. To avoid semantic confusion, it is worth noting that a Detectron2 "iteration" is synonymous with a "batch" in other deep learning libraries (e.g., PyTorch and TensorFlow), and the number of iterations in an "epoch" is thus equivalent to training set size divided by iteration batch size (Paszke et al., 2019; TensorFlow Developers, 2022; Detectron2)."

- *Line 216: Kindly provide definitions, using simple terminology or mathematical formula, for the training mask loss and average precisions shown in Fig 4. Is there a reference for the source of these definitions that could be provided?*

We have added the following text to our revised manuscript draft to clarify training mask loss and COCO AP metrics:

"…Mask RCNN mask loss, which is defined by He et al. (2018) as average binary cross

entropy loss calculated over each sigmoid-activated mask prediction for each class in each ROI of each image in a batch, is a component of the loss functions for all models and thus can be compared one-to-one between them (e.g., Fig. 4)."

"Model performance on the validation set was evaluated using bounding box and mask MS COCO AP (mean average precision) values, which are themselves arithmetic means of mean average precisions calculated at 10 segmentation intersection over union thresholds between 0.5 and 0.95 (Lin et al., 2015; COCO - Common Objects in Context, 2022). These evaluations (Fig. 4) were run every 200 iterations…"

- *Line 259, Section 4.3.1 Dataset Preparation tools: Colab_zirc_dims has a unique work flow with specific data and metadata requirements. I suggest including a flow diagram illustrating the segmentation and shape measurement procedure, inputs and outputs, including detail on image size and channels. This would help the reader understand the data requirements and process flow of colab_zirc_dims.*

We have significantly expanded Figure 5 (revised version attached) to provide a full overview of potential colab_zirc_dims workflows and potential dataset inputs. We hope that this diagram addresses the reviewer's concerns raised here and in comment 11 below. We have also added the following text to section 4.3.2:

"Researchers with datasets comprised of reflected light images that are not shot-centred and lack Chromium metadata can adapt (i.e., Fig 5a) their image datasets for use with colab_zirc_dims by either using Chromium Offline (Teledyne Photon Machines, 2020) to generate scaling and/or shot placement metadata or by manually cropping shot-centred images from mosaics (e.g., using ImageJ's "multicrop" function; Schindelin et al., 2012). Such a workflow will, however, bypass most of the automation in the colab_zirc_dims data loading process, and potential users are advised that collecting grain measurements using existing software (i.e., AnalyZr; Scharf et al., 2022) will likely be less arduous."

- *Lines 277-280: Colab_zirc_dims is designed for data output by facilities using the Chromium software. It therefore has specific data and metadata requirements. There appears some flexibility around metadata, which the authors touch on in lines 277-280. However, after reading this text I felt I lacked a clear understanding of (1) whether or not any reflected light image dataset could be adapted for use in colab_zirc_dims and (2) how this may be done (e.g. automatically generating the necessary metadata files from user inputs via script). Please could the authors expand on the flexibility and limitations surrounding the application of this tool to reflected light datasets in general. This would help readers quickly identify whether this tool can be used on their datasets. The flow diagram suggested in the previous comment may help in this regard.*

We agree that more specific information on the input requirements for colab_zirc_dims processing notebooks is in order, and hope that our revisions to components A and B of Figure 5 (see response to comment 10) adequately cover this. We also hope that the reviewer's concerns regarding a lack of discussion of limitations related to different reflected light datasets (e.g., ones varying in image quality and/or grain exposure) are addressed by our additions to section 5.2 of our revised manuscript, as described in response to comment 1.

- *Line 305, Table 3. Are the authors using the term "spot" as a synonym for "grain" in "Average segmentation time per spot"? Additionally, the footnote describes the metric as the time required for image segmentation. Does the reader interpret this metric as segmentation time per image, per grain in the image, or per analytical spot (potentially more than one spot per grain) in the image?*

We were in this case using spot as a synonym for grain (i.e., an analytical spot placed on

a grain, with one image corresponding to that grain). This is clearly confusing and inconsistent, and we have consequently revised the table column name to "Average segmentation time per image" and the footnote to "Average time for model/method to successfully segment an image and return a measurable mask. Actual per-image processing times will be higher due to additional automated mask measurement and verification image saving time. Measured in Colab notebook with NVIDIA T4 GPU."

- *Line 305, Table 3: Please provide definitions, using simple terminology or mathematical formula, for each of the tabulated metrics. Is there an appropriate literature reference for the definitions, which could be provided?*

We have revised this table to include in its footnotes written definitions for metrics "n" and "Average segmentation time per image" and formulas for each of the other calculated metrics. See the attached file 'Table3_revised_as_word' for optimal viewing of the formulas.

- *Line 305, Table 3: Please could the authors add a description of the test dataset to help the reader understand how similar the test dataset is to the training and validation datasets. This adds additional context to the performance evaluation results.*

We have added the following text to the beginning of section 5 to address this comment:

"We assessed the accuracy of our segmentation models by comparing a manually generated grain-dimension dataset (Leary et al., in press) to automatically generated grain dimensions from the same samples measured using colab_zirc_dims. The test dataset from Leary et al. (in press) consists of samples collected from late Palaeozoic strata exposed across Arizona, USA. These samples were deposited in the same orogenic system—the Ancestral Rocky Mountains—as the Leary et al. (2020a) training dataset images, and the grain ages and depositional environments are largely similar. The test dataset is unrelated to the training dataset images from UCSB (see above)."

- *Line 309: The authors refer the reader to Fig 4 and Table 3, in which models are differently named. Kindly standardise model names throughout the paper, thus facilitating quick comparison of models across tables and figures.*

We have standardized the names in our revised figures (i.e., Fig. 4; attached) and in the text of our revised manuscript where appropriate.

- *Line 310: Consider amending "training loss" to "training mask loss", to be consistent with Fig 4.*

We have made this correction to our revised manuscript.

- *Line 335: Please clarify the meaning of "skew slightly negative".*

We are referring here to skewness in our error results that is quantifiable using Pearson's skewness coefficient; this can be illustrated by the two histogram plots (for long and short axis error) that have added to a revised version of Figure 6 as Fig. 6b (see attached). To define Pearson's skewness coefficient, we have added the following equation (which will be rendered by the MS Word equation formatter as Equation 2) to section 5.2 of our revised manuscript:

Pearson's skewness coefficient = 3(mean-median)/(standard deviation)

We have also revised the beginning of section 5.2 to read:

"Per-grain automated (M-R101-C) measurements for the full Leary et al. (in press) dataset generally hew close to ground-truth measurements but with a significant number of datapoints plotting well below the 1:1 measured versus ground truth (i.e., Leary et al., in press) line (Fig. 6). The apparent dominant cause of this negative skew (i.e., Equation 2, Fig. 6B) is…"

Technical Corrections

- *Line 68: "with via" amend to "via".*
- *Line 269: "…allows to users to generate…" amend to "…allows users to generate".*
- *Line 292: "…are can…" amend to "can".*
- *Line 323: "Centermaks2" amend to "Centermask2"*

The technical corrections above have been incorporated in our draft revised manuscript. Thank you to the reviewer for identifying them!

Additional corrections:

Line 110: We erroneously state that PyTorch is developed by Google:

Corrected to "…also developed by Facebook…"

References:

COCO - Common Objects in Context: https://cocodataset.org/#detection-eval, last access: 14 July 2022.

Finzel, E. S.: Detrital zircon microtextures and U-PB geochronology of Upper Jurassic to Paleocene strata in the distal North American Cordillera foreland basin, Tectonics, 36, 1295–1316, https://doi.org/10.1002/2017TC004549, 2017.

Gehrels, G. E., Dickinson, W. R., Riley, B. C. D., Finney, S. C., and Smith, M. T.: Detrital zircon geochronology of the Roberts Mountains allochthon, Nevada, in: Special Paper 347: Paleozoic and Triassic paleogeography and tectonics of western Nevada and Northern California, vol. 347, Geological Society of America, 19–42, https://doi.org/10.1130/0-8137-2347-7.19, 2000.

He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, ArXiv170306870 Cs, 2018.

Leary, R., Smith, M., and Umhoefer, P.: Grain□Size Control on Detrital Zircon Cycloprovenance in the Late Paleozoic Paradox and Eagle Basins, USA, J. Geophys. Res. Solid Earth, 125, https://doi.org/10.1029/2019JB019226, 2020a.

Leary, R., Smith, M. E., and Umhoefer, P.: Mixed eolian-longshore sediment transport in the Late Paleozoic Arizona Pedregosa basin, USA: a case study in grain-size analysis of detrital zircon datasets, J. Sediment. Res., in press.

Leary, R. J., Umhoefer, P., Smith, M. E., Smith, T. M., Saylor, J. E., Riggs, N., Burr, G., Lodes, E., Foley, D., Licht, A., Mueller, M. A., and Baird, C.: Provenance of Pennsylvanian–Permian sedimentary rocks associated with the Ancestral Rocky Mountains orogeny in southwestern Laurentia: Implications for continental-scale Laurentian sediment transport systems, Lithosphere, 12, 88–121, https://doi.org/10.1130/L1115.1, 2020b.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P.,

Ramanan, D., Zitnick, C. L., and Dollár, P.: Microsoft COCO: Common Objects in Context, ArXiv14050312 Cs, 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, ArXiv191201703 Cs Stat, 2019.

Rainbird, R. H., McNicoll, V. J., Thériault, R. J., Heaman, L. M., Abbott, J. G., Long, D. G. F., and Thorkelson, D. J.: Pan-Continental River System Draining Grenville Orogen Recorded by U-Pb and Sm-Nd Geochronology of Neoproterozoic Quartzarenites and Mudrocks, Northwestern Canada, J. Geol., 105, 1–17, https://doi.org/10.1086/606144, 1997.

Scharf, T., Kirkland, C. L., Daggitt, M. L., Barham, M., and Puzyrev, V.: AnalyZr: A Python application for zircon grain image segmentation and shape analysis, Comput. Geosci., 162, 105057, https://doi.org/10.1016/j.cageo.2022.105057, 2022.

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, Nat. Methods, 9, 676–682, https://doi.org/10.1038/nmeth.2019, 2012.

TensorFlow Developers: TensorFlow, Zenodo, https://doi.org/10.5281/zenodo.5949169, 2022.

Detectron2: https://github.com/facebookresearch/detectron2.

Please also note the supplement to this comment:
https://gchron.copernicus.org/preprints/gchron-2022-12/gchron-2022-12-AC2-supplement.zip