Dear Dr. Vermeesch,

We want to thank you for your constructive comments on our manuscript and for the several long extensions that you and the other Geochronology editors have granted us over the past several months. Your suggestions have been incorporated in a majorly revised manuscript and via several additional pages on the colab_zirc_dims GitHub page. Please see below for our responses (plain text; with occasional references to archived emails that we have exchanged) to each of your comments (italicized):

1. *I found the paper easy to read up to Section 3.2, when its complexity suddenly increases. Here the text contains a frightening number of acronyms and technical terms, including Mask RCNN, MS COCO, NMS, Detectron2, Swin-T, FPN, ResNet50, ResNet-101, "backbone network", Centermask etc. There are three problems with this complexity. First, the AI jargon won't make sense to the vast majority of GChron readers, who are not experts in this field. Second, given the rapidly changing landscape in AI technology, it won't be long before the specific tools used in colab_zircon_dims are superseded by more performant alternatives. Therefore, even AI experts may have trouble understanding the paper in the future. Third, whilst it is easy to update software to keep up with technical developments, the same is not true for academic papers. If you swap out some components in colab_zirc_dims, then the notebook will be 'out of sync' with the GChron paper. In order to make the paper more future proof, I suggest rewriting the text in a more generic form. Please explain the AI segmentation algorithm in general terms and dedicate fewer words to the specific implementation. Technical notes are meant to be short anyway, so much of the specific details could be moved to the online documentation of the Jupyter notebook. I appreciate that it is not possible or desirable to remove all jargon from the paper. It would be useful to add a table to the revised manuscript, listing the remaining definitions and acronyms.*

We have moved all discussion of differences between deep learning model architectures, training regimes, and their respective results when applied to our test datasets to a 'Model Library' subpage of the project GitHub page, which we refer to in our revised manuscript, with most remaining deep learning technical terms and associated references removed to an appendix table (B1). Granular information on our new training-validation dataset and on exported shape parameters are also now hosted on our project GitHub page. We sincerely appreciate your advice here; these edits should give us significant leeway to update models in the future without depreciating the manuscript. We hope that they also improve the interpretability of the text.

Regarding manuscript length: the net result of these edits is a three-page reduction in total manuscript length versus the manuscript version incorporating requested edits from the referees. The main body of our updated manuscript (i.e., ignoring the two appendices, which could be moved to our supplement if absolutely necessary) is approximately 5 pages shorter. We hope that this is closer, at least, to an appropriate length for a technical note.

2. *It is not clear how the apparent grain sizes measured by the AI algorithm on 2D images relate to the actual size distribution in 3D. The introduction mentions some published studies investigating age-size relationships in zircon U-Pb geochronology. Some of these studies (Lawrence et al., 2011) used sieves, whereas others (Cantine et al., 2021) used images. Are there any studies that have compared both approaches? I would imagine that their results can differ significantly. On a related note, the reviewers have already highlighted some issues caused by colab_zirc_dims's reliance on reflected light images. As pointed out in your manuscript, reflected light images of polishing surfaces tend to underestimate the actual grain sizes. Cross sectional areas also depend*

on the polishing depth and on whether the grains are mounted as SIMSstyle epoxy pucks, or on glass slides (Figure 1). As a consequence, cross sectional grain sizes can only be used to compare samples prepared in the same lab and by the same analyst. I do not think that they can be used to compare samples from different labs or prepared by different analysts. This greatly reduces their usefulness. The revised manuscript needs to assess these limitations upfront.

We basically agree with your concerns here, and especially appreciate the figure that you created to show how grain mounts themselves likely influence surface-level grain exposures. Also, to answer your question: we are not aware of any such study. In the latest version of our revised manuscript we include distinct "Impact of grain exposure" and "Limitations" sections that note potential uncertainties stemming from differential preparation methods and/or grain exposure and highlight the need for additional study. We also mention these uncertainties in the revised abstract and conclusion. In total, these changes will hopefully impart to readers the issues (both well-understood and unresolved) inherent to our approach to grain dimension measurement.

3. *According to Section 5.2 of the paper, it would take an estimated six hours to determine the grain size distribution of the Leary dataset. This is impractical. Unless the grain size measurements are truly effortless, I'm afraid that the AI approach won't get much use. I understand that the Jupyter notebook is a proof-of-concept product. What would need to be done to make it faster or easier and use?*

As noted in our prior email exchange, most of the time estimated here would be spent manually reviewing (~1.4 hours) and correcting (~4.2 hours) grain segmentations. The best way to make this process 'effortless', then, would be to train a model that makes very few mistakes when segmenting grains. I (the first author) sought to do this by creating a new, much larger grain image-annotation dataset (1558 images with 16464 grains annotated) and retraining models using it. This succeeded to some degree: probably as a direct result of the conservative annotation strategy used on the training-validation dataset, the most performant re-trained model still fails to interpret larger subsurface grain extents where suggested by subtle "shadows" surrounding some poorly exposed grains in the test dataset (i.e., from Leary et al., 2022). It is, however, practically free (≤ 0.33%) from other segmentation error types. The estimated total time required to semi-automatically measure the ~5000-grain Leary et al. (2022) dataset using colab_zirc_dims is still, unfortunately, around 6 hours, largely dependent on surface-level grain exposure in images. For samples within the dataset with uniformly well-exposed grains, semi-automated processing is much quicker (~15 minutes/600 grains; as shown in the v1.0.10 video tutorial).

Correcting the under-segmentation errors within the confines of the existing colab_zirc_dims framework would require manually reannotating large portions (i.e., all grains with "shadows") of the new training dataset. Such is the double-edged sword of deep learning algorithms. We might do this in the future but elected not to at present; you might agree that post-preprint development of our new dataset and models has taken too long already. Our deep-learning-based models' performance when compared to algorithms that have been previously applied to this problem will hopefully still serve as an adequate proof of utility, persistent under-segmentation errors in some ALC images non-withstanding. As for the (admittedly, long) ~6 hour time estimate for semi-automated large-n dataset measurement: we agree that this probably discourages widespread adoption of colab_zirc_dims for grain measurement. We are, however, still cautiously hopeful that researchers who have already committed to the many hours of often-tedious work required to produce a large-n detrital zircon age dataset (e.g., ~60 hours of LA-ICP-MS time in the case of Leary et al. (2022)) will not be overly averse to spending several more measuring their grains.

4. *Section 3 lists the advantages and disadvantages of Google Colab. It does not mention two problems. First, Colab requires an internet connection, yet many lab computers are not connected to the web due for security reasons (automatic OS updates are disabled on most lab computers). Second, Google products are not accessible from China, which is a huge 'geochronological market'.*

We have addressed these issues by updating the code such that, given some hardware-specific additional installation steps (detailed on the project GitHub page), the code and notebooks can be run on local machines. Said instructions also detail an approach that *should* allow code and/or notebook execution on machines that are not connected to the internet, though I (first author) have not yet been able to properly test this on an out-of-date lab computer.

5. *In your response to Dr. Nachtergaele, you chose not to follow his suggestion to add a plot of grain size vs. U-Pb age to your paper, because such a plot is already scheduled to appear in an upcoming JSR paper by Leary et al. I would urge you to reconsider this decision, for two reasons. First, the Leary et al. study only presents the manual measurements, and not your automated results. Second, the paper will be stuck behind a paywall, so not all GChron readers will be able to check this useful figure. I suggest that you replace Figure 8 with a scatter plot of grain size vs. U-Pb age for a representative sample, with a box plot and KDE shown along the y- and x-axis, respectively.*

As noted in my prior email, this is now addressed with a compilation figure (Figure 5; expanded from Figure 8 of the pre-print manuscript) which illustrates the apparent reproducibility of grain age-shape relationships identified by Leary et al. (2022) using fully-automated measurement data. These relationships are explained in some detail within a new "Viability of fully automated measurement" section.

Thank you again for your time and constructive evaluation of our manuscript. We look forward to receiving your feedback on the revised copy!

Sincerely,
Michael C. Sitar and Ryan J. Leary

Reference:

Leary, R. J., Smith, M. E., and Umhoefer, P.: Mixed eolian–longshore sediment transport in the late Paleozoic Arizona shelf and Pedregosa basin, U.S.A.: A case study in grain-size analysis of detrital-zircon datasets, Journal of Sedimentary Research, 92, 676–694, https://doi.org/10.2110/jsr.2021.101, 2022.