GRAINet review answers

Major changes in the revised version

1. Revised introduction, included additional related work (RC3, RC2)
2. Added CNN architecture illustration (RC2)
3. Added results and visualizations of the learned attention maps (RC2)
4. Extend comparison between digital and field line samples (RC3)
5. Discussion on *Manual component of the presented approach (RC3)*
6. Discussion on *Geographical cross-validation (RC2)*
7. Discussion on *Comparison to past work* (RC2)

All changes in the revised paper are highlighted in blue.
Below is a systematic list of our answers.

RC1

"Why this preprocessing step was necessary since PhotoScan enables user to set the specific spatial resolution when producing orthophotos?"

This is correct, we could have fixed the resolution when generating the orthophotos. In fact the ground sampling distance (GSD) of all orthophotos is not far from 0.25 cm. We preferred to keep the best possible GSD to support the manual data annotation, then resample to exactly 0.25 cm for the CNN. We see this as a minor technical detail.

Minor comments are addressed in the paper.

RC2

Drone Flights and Image resolution

"The method would have much more impact if it could operate on the existing flight patterns (50+ meters flight altitudes) and show that you don't need the near-ground flights to get grain size distribution."

The choice of image resolution was driven by the requirement that a human annotator should be able to reliably label individual grains down to 1 cm in size. We found that a ground sampling distance of 0.25 cm is the limit for this task. To achieve that resolution, the specific UAV used in this study flew 10 m above the ground. However, the flying height depends only on the image magnification, flying at higher altitude is perfectly possible with a suitably chosen camera payload (sensor and lens). Hence, we do not see the input

resolution of 0.25 cm as a serious limitation of our proposed approach, especially given the fast progress of continuously improving and ever cheaper hardware.

"The authors did try to look at the effect of resolution. But in my experience with image texture work, I have repeatedly found that downsampling an image digitally is not the same thing as acquiring the scene from a higher altitude."

We study the effect of lower input resolutions on the performance of the grain size estimation by downsampling the high-res images (with appropriate smoothing to avoid aliasing effects) in section 5.5. To our knowledge this is the best practice and widely used e.g. in the image analysis and computer vision community to study super-resolution tasks. Our study shows that the performance of the CNN is not affected up to 2 cm ground sampling distance (downsampling factor 8), which would be the comparable resolution achieved at 80 m flying altitude with the particular low-cost UAV used in this work. We are confident that, although our synthetic downsampling may indeed miss subtle optical effects, actual images taken with 1 cm GSD would perform at least as well as simulated ones with 2 cm. Since we do not have high-altitude orthophotos for the investigated gravel bars, this is the best study we can provide.

"The authors need to choose their wording more carefully. The innovation here is that for each patch, GRAINet gives a distribution."

Addressed in the paper: lines 82-87, line 492 (blue text)

Network Architecture Description

"Why this network with residual blocks?"

After their introduction by He et al. (2016), CNNs with residual blocks quickly became the preferred design for deep image interpretation models (including many successors that no longer have the "residual" in their names). Residual connections improve gradient flow during training, because they provide shortcuts that mitigate the vanishing gradient problem. We refer to He et al. (2016), Veit et al. (2016) for more details. The model depth and capacity is empirically calibrated on the validation data for best performance on the grain size estimation task. In line with recent literature, our view is that a well-chosen loss function that matches the application objective is key, and more important than the specific choice of CNN architecture. Well-proven, contemporary architectures are rather interchangeable, and the right depth (capacity) depends on the specific task and dataset size. If a larger training set were available, an even deeper network might well perform better.

"At the very least the work needs a figure to detail the network architecture."

Addressed in the paper: line 206, page 35 Figure B1

"It would also be helpful if the authors could show some of the activation maps in order to confirm that the network is 'seeing' reasonable elements that can conceivably lead to a regression to grain size distributions."

Addressed in the paper: page 19 lines 425-433, Figure 10

Training and Overfitting

"As it stands, there is even a footnote saying it is overfit."

This seems to be a misunderstanding. To clarify, the footnote number 2 on page 5, referring to line 108, is talking about the related work by Buscombe (2019), where it appears that model selection and performance evaluation were done with the same data, which typically leads to an overfit with overly optimistic results. "Model selection" in this context means choosing at which iteration of the learning sequence the network parameters are read out. That hyper-parameter must be tuned on a validation set, without looking at the performance on the actual test data (see also Goodfellow et al. (2016), Chapter 5, section 5.3). Besides sanity-checking the training, a main purpose of monitoring the validation loss is to detect when the relatively best parameters have been found, as overfitting to the training data begins.

Contrary to our understanding of Buscombe's work, we do use the correct procedure (line 326 in the paper), i.e., our model is *not* overfitted and the reported performance is an unbiased estimator of the expected performance on unseen images.

To avoid any confusions, we cite Buscombe (2019) explicitly in this footnote 2 on page 5.

"…the authors must establish that their network is well trained."

We carried out a range of careful experiments to evaluate the performance of the network, and present its performance change w.r.t. random training-test splits and w.r.t. particular training-test gravel bars. In our opinion this is a rather comprehensive and transparent form of evaluation. We also tried to explicitly discuss both advantages and weaknesses.

Validation

"But what they call 'geographical cross-validation' is nothing else but bar-scale boot-strapping (sometimes called jack-knifing). It is not appropriate or reliable in this context. The main issue is that even if the authors hold out a whole bar, there remain samples from the same river. Given that river properties vary relatively slowly with geology and sometimes tributary inputs, the boot-strapped training samples will still have data that is very similar to the label data. So this is not a good test of generalisation. A much better approach to this would be to hold out an entire river."

Addressed in the paper on page 31 lines 560-586

"As it stands, I think the results show that the network is not well trained, has too many parameters or too little training data."

We do not understand this statement. Since no further explanation is given, we cannot properly address it. We can only hypothesize that this claim is based on the misunderstanding of footnote 2, which we have discussed above.

In fact, our experimental setup quite clearly reveals in which specific cases the limited and unbalanced training data causes performance issues. Namely, our results degrade for coarse gravel bars, which is linked to the fact that only 14% of the training data have Dm > 10 cm (see page 33 lines 616-621 in the paper).

Comparison to past work

"They must show that this is not just a new way of doing existing jobs, it represents real progress. And to do this they must cite errors from other work and clearly demonstrate that their results are better. This should include SedNet which is the rival CNN approach. And given that the title has the term UAV, they must be much more explicit in discussing their method in relation to other UAV (and airborne) particle sizing methods. Ideally, they should try to adapt the method so that it represents real progress with respect to other UAV methods in terms of time spent in acquisition, pre-processing and final data quality."

Addressed in the paper on page 32 lines 588-604

Code

The project is implemented using Keras, with Tensorflow as a backend. The code and a demo on a subset of the data is published here: https://github.com/langnico/GRAINet
Unfortunately, we are unable to distribute the full dataset for commercial reasons, as it is owned by a private company (who also created it at their own cost).
Nevertheless, the dataset may be requested for research purposes by directly contacting andrea.irniger@hzp.ch. (see revised paper "Code and data availability")

References

Buscombe, D. (2020). SediNet: A configurable deep learning model for mixed qualitative and quantitative optical granulometry. Earth Surface Processes and Landforms, 45(3), 638-651.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning, MIT Press, http://www.deeplearningbook.org.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems* (pp. 550-558).

Purinton, B., & Bookhagen, B. (2019). Introducing PebbleCounts: a grain-sizing tool for photo surveys of dynamic gravel-bed rivers. *Earth Surface Dynamics*, *7*(3).

Minor comments in the annotated pdf (supplement) are addressed in the paper.

RC3

"I would avoid the reference to Fehr (1987)"

Addressed in the paper: line 40

The main purpose of the paper is to propose and evaluate a new method to estimate grain size distribution from raw UAV images, using convolutional neural networks. We see that the standard methodology to measure grain sizes in the field may vary between countries. Fortunately, there is no tight coupling, the CNN is agnostic and will learn to replicate the outcome of any consistent procedure for creating ground truth grain size distribution samples. The annotation strategy can easily be exchanged and the network retrained to estimate grain size distribution according to a different National or regional standard (e.g., grid sampling). However, since the line sampling by Fehr continues to be the standard field method in the German speaking world, we found our digital line sampling to be most efficient, while providing representative reference data with respect to the "gold standard" of line sampling in the field.

Introduction

"Reference to traditional approaches is very poor"

Addressed in the paper: lines, 36, 40, 52

"L 15-28. This part is not very useful."

In our opinion it is important to give a broad motivation for our research and explain its relevance for society. If the reviewer (respectively, editor) insists we will of course remove the paragraph, but we would much prefer to keep it.

"I would say that automatic grain size is much less time consuming but it is also, commonly, less accurate."

We agree and mention that the automatic grain size estimation from images is still limited in terms of accuracy when it comes to large scale applications (line 61 in the paper).

Ground truth

“Is this really a ground truth? These measurements of grains are obtained from images not from direct measurements. I understand that this can be the way for training the model, but I would not say that these are ground truth.”

Addressed in the paper: line 166

Comparison with field measurements

“How field measurements were carried out should be explained in detail (in the Method section).”

We explain that field samples were measured according to the line sampling proposed by Fehr (1987) (page 15 line 354). But we feel it would be too much to add another section in the methodology, given that this is not the focus of the paper, and a widely used standard practice.

“At least for those field measurements of known location, it would be crucial to show the real difference with digital line samples.”

As mentioned in line 355, the location of the field measurements is not known exactly, because the field measurements were originally recorded for other purposes, within other projects. The best we can do with the available data is to evaluate the bar-level agreement between (independent) field work and our digital line sampling approach. For the training and evaluation of the CNN we consider the digital line samples to be representative (line 344). We do agree that a comparison between digital line samples and geolocalized field samples would be interesting, but since this data is not available, we cannot provide it in this paper.

“A better comparison with digital line samples should be carried out: I do not agree that “. . .overall, no bias exists between the field measurements and the digital line samples” (L 343-344; figure 6).”

Addressed in the paper: line 359-361, 363-365

Robust estimation

Figure 11 shows the estimated grain size distribution for entire gravel bars in comparison to the ground truth data. Given appropriate training data, the model estimates the grain size distributions very accurately. Due to the inherent regularisation, deep learning models tend to be fairly robust.

Manual component of the presented approach

Addressed in the paper: page 31 lines 543-558

“Comparison with human performance (section 5.4.4). Errors are not so small, see figure 15”

The vertical axis label in Figure 15 seems to have been mangled during pdf generation. To clarify, the Y-axis in Figure 15 corresponds to the Dm, not the error. We compare the mean diameter (Dm) estimated by the CNN to the Dm from multiple annotations by different operators. This comparison requires a lot of manual labour (repeated, independent annotation by different people), thus only a small number of samples could be processed, which is not ideal for statistical analysis. However, it still gives a feeling for the human variation in the annotation process, with an average standard deviation across different operators of 0.5 cm for Dm. The max standard deviation from repeated annotation is 2.0 cm. We correct a small mistake in line 488: The standard deviation of the labels (0.5 cm) should be compared with the root mean squared error (RMSE) of the model predictions, not with the mean absolute error MAE.

Presence of fine material

We have already included two examples in Figure 8 b) and c) that show that the network can make robust predictions even in the case of slight disturbances caused by colmation through fine material, given such samples are provided during training.

References

Wohl, E. E., Anthony, D. J., Madsen, S. W., & Thompson, D. M. (1996). A comparison of surface sampling methods for coarse fluvial sediments. *Water Resources Research*, *32*(10), 3219-3226.

Minor comments are addressed in the paper.