Cover letter with rebuttals

Dear Editor,

Please find attached the revised version of our manuscript. We have addressed the comments of both reviewers, as well as your own comment on the title. Nearly all the suggested corrections have been done. When further clarifications were needed we also answered to the reviewers in the rebuttal below. We hope that you will find this revised manuscript suitable for publication. In closing, I would like to express our gratitude to you and the reviewers for this valuable correspondence.

On behalf of the authors, best regards,

LE Van-Linh

To Reviewer 1:

Dear Vincent,

Thank you very much for your time and comments that helped us greatly improve our manuscript.

Nearly all the suggested corrections have been done (see below).

* Title : Changed according to your suggestion and the suggestion of the editor.

* Abstract: Changed.

* 8-10 : all suggestions done.

* 14: There is now a reference to fig.

* 15-22 : corrected.

* 28 and 351 : We have added the (Porto et al, 2020) reference, thank you. Although we have still kept the 2.2 subsection quite 'general', we completely agree with you on the need for comparison, even benchmarking, in automated landmarking for morphometrics. We think it could be the subject of a separate article with a companion databank and re-using (or re-implementing if necessary) several different image processing algorithms and evaluation metrics. We hope for such collective effort even if it is out of the scope of this particular article.

* 82-113 : all suggestions done.

* 115 and 230 : corrected.

* 133 : sentence corrected.

* 2.2 : Added glossary.

* 140 spatial size : replaced by ' dimension'

* 140-376 : suggestions done if not otherwise stated below.

l 193 : We modified as follow "This problem is related to the construction of a latent variable of shape deformation, I.e. estimating how landmarks move as a group. Of course, this goes further than studying, independently, each landmark correlation between predicted versus manual. At the

very least it means we are interested, for L landmarks, in the whole LxL correlation matrix not just in its diagonal."

* 242: "backpropagation": we have mentioned in settup model part.

* Table 4-6 : yes, thank you very much for pointing this out. We have added other scores.

* Figure 10: Thanks for pointing out the error in the font-size of label when we generate the images. We agree with you that there are many ways to represent this figure information. We settled for this as the least bad and one the simplest.

* line 94 : Concerning the replicability analyses, i'm afraid we are in a 'grey litterature' situation as the document left by the intern are lacking some important details. We know that 12 carabids were photographed twice with manual landmarks set N times (N > 1) by m experimenters (with 0 > m >=2). And we know the variance between replicated landmarks was "small". We are sorry we lost track of those data. The question of intra and interoperator replicability is indeed very interesting. For exemple in (Chang and Alfaro, MEE, 2016), we can see that 90% of non-experts have errors of at most 4.5% of specimen length (SL) while 50% of all landmarks have errors of at most 0.5% of SL. We are below 2% SL error and so are recent algorithms (sometimes better depending on the dataset). At the very least, we can argue that deep learning is already a good alternative to entry-level workers/interns.

* 361: We understand your remark yet labeling/naming was of interest to us. So, we have used the main component of our architecture for it.

* 378: changed as "are available as open-source software"

To Reviewer 2:

Dear Reviewer,

Thank you very much for your time and comments that helped us improve our manuscript. We will answer your remarks in the following paragraphs.

1. We agree that the light conditions are highly variable in field situations. To cope with this issue, the very first step, even more important than the following image processing, would be to optimize the image acquisition setting. For example, as was done for an ongoing community ecology experiment, one could buy or build a portable lighbox with a support for a good camera (e.g. an Olympus tough 6 for example) to dim at most the natural condition. Then by using (i) a set of portable led for homogeneous lighting, (ii) a color calibration target and (iii) homogeneous background to lie down the specimen : image acquisition would be optimal for further image processing. If the image acquisition was *in situ* (e.g. on any background with any lighting), there might be a need for more than data augmentation like custom pre-processing to help standardized images.

2. To remove background for *in situ* acquisition, one can use an array of methods, among which we can cite (a) histogram-based method, when the background is not too complex, (b) optical flow, usually when there is a time lapse and (c) more complex segmentation algorithms when the background is arbitrarily complex. To solve (c), if large training sets are available, efficient CNNs already exist for object detection and segmentation, like Mask-RCNN (He et al, ECCV 2017).

3. We completely understand your remark but we would like to keep the article following the flow of the process that we have worked on. That's why we would like to keep the theory about transfer learning and fine-tuning in the same section as its results, even if it is a bit unusual.