Response to comments received by Reviewer #1

The article presents the use of deep learning methods in the problem of the classification of whales (binary classification whales or water). The convolutional networks have been used for this.

Below are some of my comments regarding the methods presented and the results achieved.

1. In line 98 you wrote about using machine learning method in many problems. Expand the use of machine learning in medicine problem for example in the article: A Novel Machine Learning Approach for Early Detection of Hepatocellular Carcinoma Patients.

REPLY: We have added two new example citations describing the use of deep learning in a medical and urban planning setting (Lines 105-106).

2. Prepare a chapter: Deep learning with convolutional neural networks in a more scientific (technical) form. Describe the different types of layers available (not only convolutional, but also dropout layer, pooling layer). Introduce some schemes. Write about the most popular convolutional neural networks and the importance of image processing.

REPLY: We have added a description of the rapid pace of progress in this field (Lines 189-  
190). We have also further expanded our discussion of the models we selected (Lines 204-207). We have elected to avoid extra detail on the individual networks we’ve chosen, as their characteristics are well documented in the literature and our paper is primarily intended for marine mammal biologists interested in more efficient survey methods. As these models are well known, those with prior knowledge will not require this discussion and those without will find it difficult to understand in a condensed form. Better resources are available for discussion of these aspects and we have provided citations which should assist the reader (Lines 176-177; Lines 205-206).

3. Extend chapter: Model Training with detailed information about your model. Prepare a table describing the 18 layers of your model. What are the layers (convolutional, dropout, pooling)? How many neurons are in these layers? What are the activation functions? What is the optimizer and batch size?

REPLY: As discussed in the above reply, these are pre-trained and well-documented models with specifics that are well documented elsewhere and we have provided citations to direct the reader to a more complete discussion of the technical details (Lines 176-177; Lines 205-206).

4. In Table 2, we can see a larger number of examples in one of the classes. This may be the reason for the high accuracy of the model. Prepare the results also for 10-fold validation (4-fold validation is not typical). Please, show the confusion matrix and F1- score value.

REPLY: We increased our validation to 10-fold and tested on each of the model architectures (see comment 7). As stated, we implemented a weighted sampler to avoid this issue, as we do have very unbalanced classes (Lines 218-221). As noted in our manuscript, we are first and foremost interested in maximizing precision; maximizing recall is of secondary importance because false positives can be thrown out by the observer but false negatives cannot. Our automated pipeline is, at this point, best considered a means of filtering images that greatly reduces the time required by (but does not yet eliminate) human annotators. As suggested by the Reviewer, we do now include the F1-scores for our models and describe how the model with the highest F1-score may not optimal if another model that produces a higher precision with a reasonable recall rate (Lines 271-274). A confusion matrix can be found on Line 263 (Fig 4B), and others are included in supporting materials (S2 Figure).

5. Try to compare the results of accuracy with previous research on this topic.

REPLY: This is the first attempt of which we are aware to deploy machine learning for the survey of whales in sub-meter resolution imagery, and thus there are no alternative benchmarks against which this model can be compared.

6. Present the network learning time on ImageNet.

REPLY: ResNet18 and the other models used in this paper are available pre-trained on   
ImageNet (and we implemented them as such). For this reason, and in the interest of being concise, we omit these details from our manuscript as they are readily available in the cited literature. It is also worth noting that users will not need to train the network themselves.

7. Prepare the results for other typical network. Perhaps this problem is such simple that very good results can be obtained on a large number of models. The most popular models are available here: keras.io/applications/. Use one of them in PyTorch.

REPLY: We selected ResNet-32, DenseNet and ResNeXT-101 (Lines 204-207) and trained and validated the model using each. We compared model results (Line 266-273; Fig 4; Table 2) and described differences in training success and validation (Lines 233-240; Fig 3).

8. How learning rate it affects the results? Present the results for larger values learning rate, e.g. 0.1, 0.2, 0.3 or similar, together with different numbers of epochs. The results can be presented in a graph or in a table. The article will be able to be published after significant corrections. A more detailed description of the methods used (convolutional networks), detailed description of the model used, comparing it with some other model is required. A very extensive presentation of results is also required (confusion matrix, F1-score).

REPLY: For each of our model architectures, we have tested learning rates at 0.0009 (our best LR from the previous iteration), 0.001, 0.01, 0.1, and 0.2 and present the results in a table (Line 244; Table 1). We also describe this in more detail, especially for a non-expert (Lines 229-235). We also added a figure demonstrating the effect of LR on training and overall performance (Lines 237-243). We prepared a confusion matrix and table of F-1 scores (Lines 270-273; Table 3, Figure 4).

Response to Comments by Reviewer #2

This paper presents a pipeline to conducted cetacean surveys from satellite imagery, aiming to minimize the time of analyses of huge amount of images. This pipeline altogether with the satellites imagery analyses, certainly will help to improve the knowledge about vast areas of the ocean difficult to survey.  
  
General comments  
  
Title  
The title seems to be too much specific but certainly looks fine for the aim of the paper.

REPLY: We have added some clarity to the title, specifically to increase the visibility of the use   
 of aerial imagery, which is the element which solves the major problem outlined in this paper  
 (Line 4).  
  
Introduction:  
The introduction is well write. The paper lacks of hypotheses, however the authors stated very well the aims of the paper, making the reader knows exactly what the paper will be about. Thus, no hypothesis is need in this case.

REPLY: We agree with the Reviewer’s assessment.  
  
Methods  
Methods section is well write and easy to understand and follow.  
Results  
The results section is well write  
  
Discussion  
The discussion is well write however I think they should discuss a little bit more about:  
- How they can identify different whale species?

REPLY: We included reference to the only paper which has attempted species-level identification from satellite imagery and suggest that this may become possible but would require a much larger training set (Lines 313-316).

- How the price of this method is comparable with the classic surveys?

REPLY: We have added language suggesting that the cost may be comparable to aerial surveys in   
 coastal regions but likely a considerable savings in remote areas (Lines 291-296).  
  
- How other artifacts can be included? They mentioned it however it is not clear how the machine learning can works with artifacts such as boats and rocks that changes of size and shapes.  
- Although it seems to be a very nice tool, what is the feasibility of conduct this kind of survey in open waters (e.g. have good quality images) and how is the feasibility to train the model with huge areas with whales and with different other objects?

REPLY: We consider this to be a pilot study to establish method feasibility and acknowledge that environmental conditions will limit success at times, though this is equally the case for aerial and ship surveys (Lines 320-326). Further, we describe how this should not impact cost, as imagery providers do not charge for cloudy images (Lines 330-332). The question of scaling up is important and we have previously touched on it in Lines 74-77, but we include further discussion on Lines 338-343. We include an additional figure illustrating boats and discuss how this might be incorporated in future versions to improve performance (Lines 308-311; Fig 5).

Bibliography  
Reference 4 (Bedriñana et al. ), is not Biodiv, but Diversity.

REPLY: We appreciate this catch and have changed the journal name to Divers Distrib (Line 396).

Specific comments  
  
Line 53: Add a recent paper by Hucke-Gaete et al. 2018 which shows a very interesting pattern of migration in blue whales (Hucke-Gaete R., Bedriñana-Romano, L., Viddi F.A., Ruiz J.E., Torres-Florez, J.P. & Zerbini A.N. Blue whale individual movement patterns from Chilean Patagonia to Galapagos, Ecuador: novel insights on migratory pathways along the Eastern South Pacific. PeerJ 6:e4695)

REPLY: We appreciate the Reviewer alerting us to this reference, which is certainly related to   
ongoing work on whale distributions, but we believe this reference is not required to make the point being addressed here.

Line 142: E. australis

REPLY: We have corrected this (Line 146).

Response to comments by Reviewer #3:

Reviewer #3: This paper provides an algorithm for the automated detection of whales in satellite imagery. This paper is novel and timely given that satelitle imagery is becoming more readily available and that monitoring for whales in increasingly important because of environmental changes.

In general, some of the technical terms related to neural network could be explained a bit better to help the readers that are not familiar with deep learning methods.

Lines 184-185: Please, clarify. Do you mean that you divided the images in 4 folds, and within each of the 4 folds, you pick 75% of the photos for training and 25% for test. Or, do you mean that you divided the 25% of the testing photos into 4 folds?

REPLY: At the suggestion of another reviewer we changed the validation stage to incorporate 10-fold rather than 4-fold validation and have added language to improve the clarity (Lines 194-195).

Line 198. What do you mean by 100 classes?

REPLY: We have provided new examples on Lines 212-213).

Line 224-225: For the recall, does it include only “true whales”?

REPLY: We have rewritten this to add that recall refers to ‘labeled whales.’ (Line 252-253)

Line 229: Could you comment on what causes the false positive

REPLY: [[[[I’ll get back to this when get a GPU node again]]] – Add about center crops, losing   
edge, probably improved with overlapping tiles.

Discussion: it would be interesting to write a paragraph in which the precision and recall of the algorithm is compared to other similar detector.

REPLY: Following the suggestions of another reviewer, we have implemented several different model architectures with which to compare our original model (Lines 204-207).

Line 263: Could you provide more detail about the sea condition of the images used in the study? Could you provide a value on the Beaufort Sea scale?

REPLY: We have provided some further discussion on sea-state limitations (Lines 324-330). It is   
not possible to make a clear assessment of sea state given the imagery resolution, but we discuss how any image below Beaufort 4 should appear similar in quality.

Lines 269-272: Could you comment on the minimum resolution needed for detection in satellite imagery and the resolution of the other imagery currently available. Similarly, could you comment on the minimal size of a whale that could be detected by the algorithm.

REPLY: We have added the resolution of the previous use of satellite imagery for manual whale detection (Line 84). We would prefer not to comment on a minimum size as we have no firm evidence for a minimum. We do note however that the model was trained on images of minke whales, the smallest baleen whale and roughly the size of an orca or beaked whale (Lines 302-304).

Line 276: Could you comment on the likelihood of getting daily images for a given location. Our experience in the Arctic is that images are hard to get

REPLY: A great deal of detail may be outside of the scope of this paper, but we agree that this is   
an important consideration, especially for those new to satellite imagery and have added some clarifying language about the process of image collection and the current orbits of the commercial constellation. We specifically address the question of high-latitude image collection which we agree can be challenging (Lines 293-296).

Supplement files: the names of the files are confusing and contradictory. For example, the file names Supplement S1 has for title Supplement S3.

REPLY: We have edited the supplements and corrected any errors.

Response to comments by Reviewer #4:

This paper propose to use CNNs to classify satellite cetacean image tiles in order to eliminate laborious human screening efforts. The idea is technically intuitive and sound. The experimental setup is clearly described. However, I have strong concerns about the novelty of this paper. The proposed CNN is a common classification network. The authors used by pytorch as the deep learning framework. By checking the implementation details I feel like the implementation is merely re-organizing customized cetacean training data with out-of-the-box pytorch training pipelines. What make things worse, the results are plain and have no comparison with state-of-the-art works. Based on these reasons, my recommandation is Major Revision to Reject.

REPLY: While the reviewer is correct that we have employed a common CNN, novelty is not a criterion for publication in PLOS ONE and an off-the-shelf CNN improves the portability of this method to new users who are more familiar with field-based studies than machine-learning solutions. With this method, a potential user has a wealth of support available from other users and will be able to troubleshoot common problems. Complicating the process by creating a more bespoke approach will limit the utility and intelligibility of this method. We hope that our approach is accessible enough that those with minimal experience with machine learning could feel empowered to take our code and begin to deploy it for their own regions and species.

With regard to comments regarding the plainness of results and comparisons with other methods, we have elected to compare the ResNet-18 implementation with two other models: DenseNet and ResNeXt, and another implementation of ResNet, the ResNet-34 architecture (Lines 204-207; 266-267; Table 3, Fig. 4).