**Response to Reviewers**

***Review Comments to the Author***  
Please use the space provided to explain your answers to the questions above. You may also include additional comments for the author, including concerns about dual publication, research ethics, or publication ethics. (Please upload your review as an attachment if it exceeds 20,000 characters)  
  
Reviewer #1: After the corrections made, the article can be published.  
  
Reviewer #2: (No Response)  
  
Reviewer #4: The authors have addressed some of the reviewers' comments, and improved the experiments to some extent. For example, authors have added comparisons of Renset18/34/152 and DenseNet for whale/water classification.  
However, there are still major concerns in the experimental part:

1. They way authors chose Learning rate = 0.0009 in Table 1 seems ad-hoc.

**Reply:** The choice of this learning rate was the result of experimentation, though we had not included a great deal of detail on this in the manuscript. We have now included the results of further learning rates in Table 1 which we believe demonstrate that we have tuned this parameter with intention [Line 242]. We have previously described the effect of learning rate on model performance and included a figure to this effect [Lines 237-240; Fig. 3A]. We have updated this figure to make the effect of learning rate clearer. While it is possible that a better model could be trained with a slower learning rate and more epochs, our aim here is to demonstrate that, even with 2-4 hours of training, one can arrive at model weights that can classify whales quite accurately.

2. In Table 3. how do authors explain that Renset18 is better than Resnet34/152 given the latters are proven to be better in image classification?

**Reply:** Having investigated this further, we believe that the inverse performance is likely attributable to overfitting in ResNet-34 and -152. We tested the model weights from earlier epochs and found the pattern to indeed be as one might expect – ResNet-152 outperforms the others, though the results of ResNet-18 and -34 are very similar. We have updated Table 3 to reflect this and include some discussion of this possibility to aid readers in making decisions with their own models if they attempt the same [Lines 307-314; 361-370]. We have also included a new figure [Fig. 3B] that shows the training/testing curve for our best model.

3. In Table 2, what happened to the 10th fold, the recall is abnormally small. Seems like the same problem happened to the first fold as well.

**Reply:** We separated our folds broadly based upon different scenes, as we considered that a complete randomization would defeat the purpose of the multi-fold process. The characteristics of individual parts of a water scene are quite similar to one another, but very different from other scenes in which the sea conditions and water coloration vary. In this case, Fold 10 is testing on several scenes, including one which has rougher sea conditions than are found in the other scenes. Since the model has only trained on calmer conditions, we believe it considers the edges and contrast in the rough-water scene to be whales. This certainly suggests that future improvements should focus on incorporating additional rough water, as we consider in the Discussion. We have added additional discussion of this particular instance [Lines 256-262] and included a supplement [S2 Fig] to demonstrate this issue.

4. Again, the authors claim that this is the first work to apply CNN to classify cetaceans from satellite imagery. Then how do we compare the results to previous researches? A traditional method must be present. E.g., given the fact that Resnet18 is better than deeper and wider networks, will simpler classifiers like SVM classifiers work even better in this case, I wish to see the results.  
  
In summary, I appreciate the efforts of this work, but I hope the experimental setup can be organized better.

**Reply:** We have now included two non-CNN approaches: a ridge regression and an SVM approach (C-SVC) and describe the results [Lines 222-228; 284-296]. We are grateful for this suggestion as including these methods demonstrates the fact that CNNs are better at subtle classification than methods like SVM, though in the absence of a CNN, those methods would still provide some classification benefit. With regard to previous research, our aim was not to demonstrate that a CNN provides an improvement over previous classification methods but rather that a CNN can be used to augment and improve existing visual survey methods currently standard in the field of marine mammal biology. Satellite imagery has not been used widely and we are unaware of any previous attempts to automate the detection of whales in satellite imagery at this scale. It is impossible to directly compare our method to field surveys as the limitations of satellite imagery (namely the revisit time and potential for cloud cover to obscure the target area) require a much different analytical approach. As such, we don’t view this method as replacing an inferior method but rather as providing an additional data stream that can help fill data gaps in both spatial and temporal coverage of study areas. We discuss this at length in the Introduction and Discussion [Lines 54-91; Lines 363-374]. We also discuss the one previous attempt to tackle this problem in the Introduction [Lines 76-79], where Fretwell et al. used maximum likelihood supervised classification and isoData and k-means unsupervised methods to threshold the spectral bands in imagery to isolate whales. We have added a description of why we believe these methods would not be robust to differences in satellite scenes from different water bodies in the Discussion, and thus how this method marks a clear improvement [Lines 315-324].