Dear Editor-in-Chief, Aquaculture International

Thank you very much for reviewing our manuscript, titled “Recognition of Partially Occluded Soft-shell Mud Crabs Using Deep Learning-based Image Classification and Object Detection” in the previous manuscript and titled “Recognition of Partially Occluded Soft-shell Mud Crabs Using Faster R-CNN and Grad-CAM” in the revised manuscript. We also deeply appreciate the positive, constructive, and valuable suggestions from the reviewers. We have carried out all the issues suggested and revised the manuscript accordingly.

Please find an attachment regarding a point-by-point response to the reviewers’ concerns. We hope that you will find our responses satisfactory and that the revised manuscript is acceptable for publication. Thank you again for your great assistance.

Best Regards,

Thitirat Siriborvornratanakul, Ph.D. (University of Tokyo)
Assistant Professor
Graduate School of Applied Statistics
National Institute of Development Administration (NIDA)
148 SeriThai Rd., Klong-chan, Bangkapi, Bangkok 10240, Thailand
Reviewer 1

This paper presents an interesting study to classify crab mooting status under occluded conditions (covered by lid) as not molting (one crab shell), molting complete (two shell detected) or empty basket. The motivation of the study are clearly outlined and a comprehensive experimental evaluation is conducted. Strengths of the paper are:

• Strong experimental analysis across two datasets consisting of 4022 images collected under both natural and artificial light.

• A robust comparison of different detection backbone algorithms including VGG-16, ResNet-50, InceptionV3, MobileNetV2.

• Analysis of transfer learning to improve performance

• Paper makes an important contribution highlighting feature importance analysis to evaluate model performance. Using Grad-CAM to visualise how each pixel contributes to the final decision of the model, authors illustrate that sometimes the model makes the right decision for the wrong reason. This is an important point to assess the robustness of predictions and diagnose model issues

I found this a good paper and recommend publication

Answer Thank you for assessing our paper and for your recommendation that our manuscript is suitable to be published in Aquaculture International.


**Reviewer 2**

We appreciate the constructive comments from the reviewers. The following are our point-by-point responses.

The reviewed paper presented a deep learning-based proposal to detect and classify soft-shell Mud Crabs. The proposal is very interesting, and generally, the paper was well written. Unfortunately, the paper contained many flaws. The authors exaggerated the results obtained and presented arguments unsupported by scientific evidence. The images do not seem adequate to achieve success in the study. The methodology presents few details, and the selection of techniques was not duly justified. The experiments and the results need to be better explained. In addition, the data must be compelling.

**Answer:** Thank you for your comments. To improve the quality of our manuscript, we revised it according to your comments (removing exaggerated words, discussing experimental results beyond images, adding more implementation detail, reasoning our selection, and explaining results in more detail). Please find them in the revised manuscript.

Please avoid exaggerations. For example, what is the meaning of “yield high accuracies.” What is a good quantity to ensure that the accuracy is high?

**Answer:** Thank you for your suggestion. We removed exaggerated words from the revised manuscript and replaced them with the facts obtained from our experimental results. For example, in the Abstract, “yield high accuracies” was replaced by actual numeric performances from experimental results.

The title is very large and relatively general. Deep learning is a broad field. Moreover, please compare the title and the goal of the paper. With this comparison, the title can be improved.

**Answer:** The title of the revised manuscript is changed to “Recognition of Partially Occluded Soft-shell Mud Crabs Using Faster R-CNN and Grad-CAM” to clearly specify the deep learning techniques used in our work.

The abstract should be rewritten to include information regarding the methodology proposed to solve the problem, a description of the experiment, and numerical information and comparison versus SOTA.

**Answer:** In the revised manuscript, the abstract is revised as follows. However, as our work involves our self-collected dataset and the understudied area of object occlusion, our experimental results cannot be directly compared with results from previous state-of-the-art research.

> Soft shell crab production is a labor-intensive process that necessitates constant inspections by aquacultural farmers to determine the completion of each crab’s molting. The need for continuous visual inspections throughout the harvest days is imperative to prevent the new shell from hardening, a process that commences within three hours after molting. This study proposes a vision-based deep learning approach to automate this repetitive visual inspection task, specifically targeting the challenge of locating and detecting crabs concealed within box baskets, which obstruct the camera's view. Initially, we explored the well-established technique of image classification through transfer learning
and fine-tuning a pre-trained VGG-16 model, achieving an accuracy of 99%. However, our investigation using Gradient-weighted Class Activation Mapping (Grad-CAM) uncovered the sensitivity of this mature classification method to image bias, leading to the inadvertent learning of false features. In pursuit of a more robust alternative, we turned to object detection techniques, particularly Faster R-CNN, which proved to be better suited to our problem and dataset. Through experimentation with various backbones and bounding box confidence thresholds, our resulting model demonstrated the capability to detect crabs with an average precision (AP 0.5) ranging from 83% to 91%. By translating crab counting results into image classification, the accuracy remained consistently high at 98-99%. This innovative approach offers a promising solution to streamline soft-shell crab production without compromising accuracy.

**What is Grad-CAM? All the acronyms should be explained on the first usage. How false features are detected?**

**Answer:** Information about Grad-CAM is added to Section 4.1.2 (Visual explanation algorithm) of the revised manuscript as follow:

*Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) serves as a visual explanation algorithm designed to shed light on the decision-making process within CNNs, often seen as black-box models. It achieves this by computing the gradients of a class-specific prediction score with respect to the 2D feature maps derived from the final convolutional layer, which is expected to capture the highest-level features. Grad-CAM subsequently generates a heatmap image that visually illustrates how each pixel in an input image influences the prediction score for a specific class. Consequently, it provides valuable insights into whether our CNN model is making decisions based on the relevant regions of the input image. For instance, in our image classification experiment (Section 4.2.1), we anticipated the heatmap to emphasize the presence of crabs in the image. However, it unexpectedly highlighted the basket instead. This observation indicates that the model has erroneously learned features unrelated to crabs, which are essentially considered false features in our context of classification.*

**Please avoid repetitions. Many sentences are very repetitive.**

**Answer:** Thank you for your comment. We thoroughly read the whole manuscript, revised it, and removed sentence repetitions.

**The introduction is very well-written and easy to read. However, authors should include more references. For example, the authors claim that deep learning has an impressive performance in image processing. However, value judgments are not valid in science. Therefore, to solve the problem, the authors must include a reference. This procedure is needed in many parts of the paper.**

**Answer:** As suggested, we added three new references (Yang et al., 2020; Chen et al., 2022; Liu et al., 2023) to Section 1 (Introduction) to support the claim regarding the impressive performance of deep learning in image processing.
Please explain how many classes the crabs are classified (this is explained until chapter three).

**Answer:** The following explanation about the four classes in our work is included in Section 3.1 of the revised manuscript, along with the revised Table 2.

We organized our image samples into three distinct groups for classification purposes. The first group signifies empty baskets (class 0), the second group comprises hard-shell crabs (class 1), and the third group includes soft-shell crabs along with their discarded shells (class 2). In the context of image classification (Section 3.2.1), the model’s task is to categorize images into these three designated groups. However, the objective of object detection (Section 3.2.2) differs somewhat. In this scenario, the model is tasked with detecting crabs, if any, within the image and quantifying the number of crabs present. Subsequently, this count is compared to the sample groups. In cases where the object detection model mistakenly identifies more than two crabs in a single image or encounters difficulties fitting the result into any of the three groups due to errors, the outcome is classified as an “overcount” and placed into its own group (class 3). The implications of each technique are detailed in Table 2.

**Table 2. Four classes regarding each implementation technique of this paper**

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of crab</th>
<th>Farming scenarios</th>
<th>Image classification</th>
<th>Object detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No crab</td>
<td>Empty basket</td>
<td>An empty basket with no crab</td>
<td>An empty basket with no crab</td>
</tr>
<tr>
<td>1</td>
<td>One crab</td>
<td>Hard-shell crab</td>
<td>Molting incompletion</td>
<td>Molting incompletion</td>
</tr>
<tr>
<td>2</td>
<td>Two crabs</td>
<td>Soft-shell crab and its shell</td>
<td>Molting completion</td>
<td>Molting completion</td>
</tr>
<tr>
<td>3*</td>
<td>More than two crabs</td>
<td>N/A</td>
<td>N/A</td>
<td>Over counting</td>
</tr>
</tbody>
</table>

* Class 3 does not correspond to any of the genuine soft-shell farming scenarios; instead, it is utilized to denote counting errors that may occur in the object detection technique.

The authors should review and rewrite the contributions. The current was not written correctly. Also, after reviewing the article, it is clear that the list should include more contributions. Please add a paragraph at the end of the introduction to describe the organization of the project.

**Answer:** Our contributions are revised and summarized at the end of Section 1 (Introduction) as follows:

1. **This research presents a series of experiments along with their outcomes, demonstrating the limitations of well-known ImageNet classification backbones in detecting false features when applied to partially occluded crabs within our self-collected image dataset.**

2. **We introduce an alternative solution involving the Faster R-CNN object detector, leveraging various backbone architectures. Our findings validate that this approach exhibits superior robustness when confronted with partially occluded crabs and our self-collected image dataset, addressing the limitations encountered with traditional ImageNet classification backbones.**
The analysis of each work presented in Section II should be extended. Just a few details were offered for each work. Consequently, the information is not valuable to the reader. Including a table to compare the features of each work reviewed is recommended.

**Answer:** Thank you for your suggestion. In the revised manuscript, a new Table 1 is included and discussed in Section 2 (Related Works) as follows:

*Table 1 offers a comprehensive summary of the aquaculture studies discussed earlier. Notably, while numerous studies have explored image classification and object detection, there has been limited attention directed toward scenarios involving significant occlusion of target objects, such as crabs. This uniqueness within the domain distinguishes our work as it addresses this specific challenge.*

**Table 1. Summary of related works in aquaculture as discussed in Section 2**

<table>
<thead>
<tr>
<th>Work</th>
<th>Occlusion</th>
<th>Method</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Xuan et al. 2018) Pearls’ defects classification</td>
<td>No</td>
<td>MS-CNN</td>
<td>Two-class 92.14% accuracy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>seven-class 91.24% accuracy</td>
</tr>
<tr>
<td>(Liu, 2020) Shrimp quality classification</td>
<td>No</td>
<td>Deep-ShrimpNet (an improved AlexNet) with IMJA algorithm</td>
<td>98% accuracy on validation</td>
</tr>
<tr>
<td>(Wang et al., 2016) Counting fully-visible swimming crabs</td>
<td>No</td>
<td>Haar operator and Adaboost cascade</td>
<td>79.5% precision</td>
</tr>
<tr>
<td>(Pitakphongmetha et al., 2021) Crab detection in infrared images</td>
<td>No</td>
<td>YOLOv2</td>
<td>Daytime 55.3% AP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Night-time 78.6% AP</td>
</tr>
<tr>
<td>(Tang et al., 2020) Crab detection under turbulent water</td>
<td>No</td>
<td>YOLOv3 with adaptive dark-channel defogging algorithm</td>
<td>91% precision</td>
</tr>
<tr>
<td>(Nimitkul et al., 2022) Molting crab classification</td>
<td>No</td>
<td>Pixel count by a threshold of molting and non-molting pixel</td>
<td>22.45% accuracy</td>
</tr>
<tr>
<td><strong>Ours:</strong> Counting crabs in a basket</td>
<td>Yes</td>
<td>VGG-16 image classification</td>
<td>Detect false features regarding non-crab areas</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>Faster R-CNN</td>
<td>83-91% AP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>98.99% accuracy</td>
</tr>
</tbody>
</table>

Please include details about the existence of class zero. Therefore, in Table 1 and class zero, it is unclear which object is occluded.

**Answer:** Class 0 refers to a situation of an empty basket where no crab is present in the basket. In the revised manuscript, we changed the wording in the table to “An empty basket with no crab” for better clarification. Please note that, due to a new table being added, this table is numbered as Table 2 in the revised manuscript.
Historically, deep learning techniques are characterized by using many samples as input. Considering this, please explain why only 4022 images were acquired. Moreover, how authors deal with the problem of class imbalance.

**Answer:** Please find our response to this comment in the following paragraph which is included in Section 3.1 (Dataset) of the revised manuscript:

Despite the typical requirement for large training datasets in deep learning, our approach leverages models pre-trained on extensive image datasets. Utilizing these pre-trained models offers the advantage of reusing previously learned features and knowledge, significantly reducing the need for extensive training resources. However, in order to enhance the diversity of our image dataset and address class imbalance issues, we employ image augmentation techniques to oversample our dataset. In the task of image classification (Section 3.2.1), the original images undergo separate transformations, including cropping (“cropped images”) and augmentation (“augmented images”). These transformed images are then proportionally combined with the original images to create two distinct datasets: Dataset A and Dataset B.

The quality of Figure 1 must be improved. The current version is completely useless.

**Answer:** In the revised manuscript, we updated all figures to their maximum resolution possible. Also, zooming effects were added to Figure 1 for better visualization as follows.

![Figure 1](image-url)
Please add details about the transformations conducted during the stage of data augmentation.

**Answer:** Detailed settings of our image augmentation are provided as follows in Section 3.1 (Dataset) of the revised manuscript.

To generate the augmented images, sections of the original images and cropped images are labeled and resized to 224x224 pixels. We employ Albumentations version 1.1.0 (https://github.com/albumentations-team/albumentations) for image augmentation, using the following parameter settings: [A.HorizontalFlip \( p=0.5 \), A.VerticalFlip \( p=0.5 \), A.GaussNoise \( \text{var}_\text{limit}=(10.0, 50.0) \), \( p=0.5 \), A.Rotate \( \text{limit}=(-5,5) \), \( p=1 \), \( \text{border}_\text{mode}=5 \), A.HueSaturationValue \( \text{hue}_\text{shift}_\text{limit}=(-5,5) \), \( \text{sat}_\text{shift}_\text{limit}=(0,20) \), \( \text{val}_\text{shift}_\text{limit}=0 \), \( p=0.5 \), A.RandomBrightnessContrast \( \text{brightness}_\text{limit}=0.1 \), \( \text{contrast}_\text{limit}=0.1 \), \( \text{brightness}_\text{by}_\text{max}=\text{True} \), \text{always}_\text{apply}=False, \( p=0.5 \), A.RandomCrop \( \text{int}(\text{img}.\text{shape}[0] \ast \text{random}.\text{randint}(85,90)/100), \text{int}(\text{img}.\text{shape}[1] \ast \text{random}.\text{randint}(85,90)/100), p=0.7 \)]}. In total, Dataset A comprises 9,741 images, and Dataset B contains 9,654 images. The distribution of images used in this technique, categorized by image alterations, is provided in Table 5, while examples of original and cropped images are illustrated in Figure 2.

Figure 2 has the same problem as Figure 1. That is, nothing can be seen. Maybe the authors should include a zoomable image. I do not understand how these images were used for the detection and classification.

**Answer:** The purpose of Figure 2 is to make a side-by-side comparison between an original image before cropping (left) and the cropped image (right). We revised the figure caption for better clarification as follows: Example of how an original image (left) is cropped along the red rectangular box, resulting in the cropped image (right) used in the image classification techniques (Section 3.2.1).

In the first paragraph of subsection 3.2.1, please insert scientific arguments to justify the selection of VGG-16. What were the other architectures considered? Why were the other architectures discarded? Moreover, please explain why the fully connected layers were removed, the feature extractor was kept, and the ReLu was reattached.

**Answer:** Please find our response regarding this comment in the following paragraphs included in Section 3.2.1 of the revised manuscript:

In the realm of deep learning-based image classification, a typical architecture comprises two main components: a feature extractor and a classifier. The feature extractor is responsible for learning meaningful representations of images, while the classifier assigns these representations to specific target classes. As outlined in Table 4, our image classification task involves a 3-class classification objective. To achieve this, we opted to employ the pre-trained VGG-16 model (Simonyan and Zisserman, 2015). VGG-16 is a state-of-the-art convolutional neural network (CNN) renowned for achieving a test accuracy of 92.7% on the ImageNet dataset. Moreover, it is frequently used as a feature extractor in object detection models like Fast R-CNN or Faster R-CNN, making it an ideal choice for a fair comparison with our subsequent technique (Section 3.2.2) involving Faster R-CNN.

The original feature extraction component of VGG-16 consists of 13 convolutional layers with ReLU activation functions and 5 max-pooling layers. Meanwhile, the classifier component comprises two fully-connected layers with 4096 channels, each employing ReLU activation functions, followed by a flattening layer that reduces the output to 1000 channels before applying the softmax function to determine image classes. Given that our image dataset significantly differs from the ImageNet dataset
used for pre-training VGG-16, we made the decision to remove the pre-trained classifier section (which classifies 1000 ImageNet classes) from VGG-16. In its place, we introduced our custom classifier tailored for our 3 classes. Our custom classifier consists of four sequential fully-connected layers with 1024 (ReLU activation), 512 (ReLU activation), 512 (ReLU activation), and 3 (Softmax activation) nodes, respectively. This adaptation ensures that VGG-16 is well-suited to our specific classification task.

In Table 6, please explain what a parameter is and how the number of parameters was computed.

**Answer:** In Section 3.2.1, this table represents the numbers of trainable and non-trainable parameters regarding each training strategy. Trainable parameters are those subject to updates and adjustments during training, while non-trainable parameters remain either unmodifiable or intentionally frozen throughout the training process. To provide further clarity on this distinction, we have included this explanatory information in Section 3.2.1 of the revised manuscript.

Why was the model trained in 100 epochs? Why was ADAM selected? How the net hyperparameters were calculated?

**Answer:** Determining the ideal number of training epochs for a model can be somewhat arbitrary, as there is no universal rule of thumb. In this study, we opted for 100 epochs, primarily because it proved sufficient for our model's training loss to converge effectively. We also chose to employ the Adam optimizer (Kingma and Ba, 2015), owing to its ability to facilitate rapid convergence during training while demanding relatively low memory resources. The selection of other hyperparameters was based on a series of trial-and-error experiments, where we fine-tuned these parameters to optimize model performance. We have incorporated this additional information into Sections 3 and 4 of the revised manuscript to provide a comprehensive understanding of our approach.

Please explain in depth the process of image classification.

**Answer:** An explanation regarding our image classification process is included in Section 3.2.1 as follows:

*In the realm of deep learning-based image classification, a typical architecture comprises two main components: a feature extractor and a classifier. The feature extractor is responsible for learning meaningful representations of images, while the classifier assigns these representations to specific target classes. As outlined in Table 4, our image classification task involves a 3-class classification objective. To achieve this, we opted to employ the pre-trained VGG-16 model (Simonyan and Zisserman, 2015). VGG-16 is a state-of-the-art convolutional neural network (CNN) renowned for achieving a test accuracy of 92.7% on the ImageNet dataset. Moreover, it is frequently used as a feature extractor in object detection models like Fast R-CNN or Faster R-CNN, making it an ideal choice for a fair comparison with our subsequent technique (Section 3.2.2) involving Faster R-CNN.*
Please review the first paragraph of subsection 3.2.2.

**Answer:** We revised the first paragraph of Section 3.2.2 to include a more detailed explanation as follows:

While image classification has been a prevalent and well-established solution, as discussed in Section 2, our subsequent investigation (Section 4.2.1) uncovered that image classification models are susceptible to dataset bias and can achieve high performance even without focusing on the crab(s) inside the basket. Therefore, the primary goal of our object detection technique is to address this identified issue and to leverage the classification result by inference based on the number of detected crabs. In essence, the number of detected crabs in the box basket in this technique will be interpreted and classified into the four previously summarized classes, as presented in Table 2.

Table 3 shows three classes; however, the paragraph refers to four.

**Answer:** This table refers to the number of original images categorized by classes and lighting conditions. However, the actual numbers of classes to be used during image classification and object detection are 3 and 4 respectively, as stated in Table 2. Please note that Table 3 in the submitted manuscript is shifted to Table 4 in the revised manuscript.

All the information presented in the second paragraph in subsection 3.2.2 should be explained using scientific arguments. Why was the faster-R-CNN selected? What other detectors were considered, and why were they discarded? Which technique was employed for calculating the hyperparameters? How were the thresholds of the bounding boxes computed? How were the sizes of anchor boxes determined? Many details are needed to ensure that other researchers can repeat the methodology.

**Answer:** Please find our response to this comment in Section 3.2.2 of the revised manuscript as follows:

In the realm of CNN-based object detectors, there exist two primary paradigms: the one-stage detector and the two-stage detector. One-stage detectors include well-known models such as YOLO, SSD, and RetinaNet, while the majority of two-stage detectors are built upon the foundational principles of Faster R-CNN, often incorporating various feature extraction backbones. Our research opts for Faster R-CNN (Ren et al. 2015), primarily because it excels at providing bounding boxes of diverse sizes. This choice was made to ensure that our model can effectively address any potential challenges related to object size variations.

Faster R-CNN comprises two key sub-networks: the Region Proposal Network (RPN) and the classifier network. In our study, we employ Faster R-CNN object detectors with various image classification backbones, including VGG-16, ResNet50 (He et al., 2016), InceptionV3 (Szegedy et al., 2016), and MobileNetV2 (Sandler et al., 2018) with an alpha value of 1.0. Both sub-networks are trained using the Adam optimizer, initialized with a learning rate of 1e-5. We fine-tune all hyperparameters of the entire model using Stochastic Gradient Descent (SGD) with an initial learning rate set to 0.01. Our optimization objective is to minimize the mean absolute error (MAE) loss, which encompasses both the class prediction error and the bounding box coordinate error from both sub-networks. Our models undergo training for a total of 50 epochs. For implementation, we utilize the Keras deep learning API in Python, with code modifications adapted from the following source: https://github.com/kbardool/keras-frcnn.

For the purpose of performance comparison, we configure the bounding box class confidence thresholds for each backbone at three different levels: 0.80, 0.85, and 0.90. These thresholds are applied to the bounding boxes’ class probabilities determined by the classifier sub-networks using the
Softmax function. This comparison plays a crucial role in deciding whether to retain or discard the candidate bounding boxes generated by the model. In our approach, a total of 9 anchor boxes are employed, each with sizes of 128, 256, and 512 pixels, while maintaining aspect ratios of 1:1, $\frac{1}{\sqrt{2}}$, and $\frac{2}{\sqrt{2}}$. This selection of anchor boxes is designed to cover a range of object sizes and aspect ratios effectively. To capture more intricate feature details, as advocated in the original Faster R-CNN study, we utilize a region of interest (ROI) pooling size of 7x7 instead of 14x14. This adjustment serves to expedite the training process without sacrificing feature quality. All object detection models in our study are trained on Google Colab Pro, making use of the Tesla P100 GPU for computational acceleration.

Please be careful when reporting any result in which 100% success was obtained. The first thing that can be considered is that the experiment was improper. Also, considering the images presented, it seems almost impossible to obtain 100% results.

**Answer:** We agreed that 100% results are uncommon in scientific reports. As mentioned in Section 4.2.1 of the revised manuscript, our image classification models incorrectly learned non-crab areas instead of crab areas, resulting in the incorrect 100% results that are revealed later by Grad-CAM.

Please explain how the Grad-Cam map was computed.

**Answer:** Computation of Grad-CAM is added to Section 4.1.2 of the revised manuscript as follows:

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) serves as a visual explanation algorithm designed to shed light on the decision-making process within CNNs, often seen as black-box models. It achieves this by computing the gradients of a class-specific prediction score with respect to the 2D feature maps derived from the final convolutional layer, which is expected to capture the highest-level features. Grad-CAM subsequently generates a heatmap image that visually illustrates how each pixel in an input image influences the prediction score for a specific class. Consequently, it provides valuable insights into whether our CNN model is making decisions based on the relevant regions of the input image. For instance, in our image classification experiment (Section 4.2.1), we anticipated the heatmap to emphasize the presence of crabs in the image. However, it unexpectedly highlighted the basket instead. This observation indicates that the model has erroneously learned features unrelated to crabs, which are essentially considered false features in our context of classification.

Clear explanations about the information shown in Figures 5 and 8 should be included.

**Answer:** We revised the captions of all figures in the paper to include more explanation detail. Please find them in the revised manuscript.

Tables 7 and 8 were never explained.

**Answer:** Thank you for your correction. Both tables (which have become Tables 8 and 9 in the revised manuscript) are mentioned and explained in Section 4.2.2 of the revised manuscript.
Section 4 leaves a lot to be desired and needs to be rewritten. Experiments and results must be clearly explained. Please do not include misleading information. Explain to the readers why good results are obtained, which is a good result. How many tests were done? Show that the differences between the methods are significant. What were the authors based on to say their results are good or better? Please insert a scientific discussion of the results.

**Answer:** We thoroughly revised Section 4 according to your suggestion, including more scientific discussion about experimental results. Please find the revised section in the revised manuscript.

Please rewrite the conclusions and extend the information regarding further works. The methodology presents few details, and the selection of techniques was not duly justified.

**Answer:** Section 6 (Conclusion and Future Works) is revised as follows:

This study addresses a less-explored challenge involving the detection of molted mud crabs placed inside baskets, where the lid obstructs a complete view of the crabs. Despite achieving nearly 100% accuracy, as indicated by the confusion matrices, the image classification models ultimately fail to meet the study’s objective due to the issue of false feature recognition, as confirmed by Grad-CAM analysis. Conversely, among the object detection models, Faster R-CNN with VGG-16 backbone only misclassifies one image, resulting in an accuracy of 99.87% at a 0.8 bounding confidence threshold. While all object detection models successfully detect crabs without false positives, the difference in size between predicted bounding boxes and ground-truth bounding boxes prevents the crab class from reaching 100% accuracy.

In considering future developments, it’s imperative to explore the design of a mechanical handling system that operates in synchronization with the camera when implementing these solutions in actual farming scenarios. This integration would ensure a seamless and efficient workflow for crab detection and classification. Moreover, if video analytics is considered a potential enhancement in the future, there is a need to address the possibility of false crab detections during the transition period between baskets. Specifically, the model may detect crabs both within the processed basket and the upcoming basket within the same frame, potentially leading to erroneous classifications. To mitigate this issue, it might be necessary to implement object tracking mechanisms that can monitor and follow the detected crabs in motion, thereby preventing such instances of false detection and classification.

Thank you very much again for your consideration. We hope that you find our responses satisfactory and that the revised manuscript is now acceptable for publication.
Reviewer 3

We appreciate the constructive comments from the reviewers. The following are our point-by-point responses.

The manuscript describes a system for automatic recognition and/or detection of crab molting completion using Deep Learning techniques with Computer Vision. The images were split into three categories or classes: empty basket, one crab, and two crabs. The two crabs mean the molting was completed. The results show high accuracy in classification regardless of a visual inspection of the system with GradCam that shows the model is paying attention to features other than the crabs inside the basket. Detection results look more promising with only one error in the test data.

The manuscript is well-written, and the results look ok. Since this is a practical application of existing techniques, it is unclear what problem it solves.

Answer: Thank you for your comments. In our study focused on classifying molting crabs within baskets under occlusion conditions, we initially adopted established image classification techniques that exhibited robustness against occlusion in prior literature. However, upon applying the Grad-CAM technique (Selvaraju et al., 2017), we observed that these conventional methods did not perform as anticipated when applied to our self-collected image dataset. In response to this emerging challenge, we conducted extensive experimentation with alternative approaches and ultimately found that object detection techniques proved to be highly effective in addressing this issue. These techniques not only successfully handled occlusion but also exhibited resilience in the face of certain biases present in our image dataset. This explanation is added to Section 1 (Introduction) of the revised manuscript.

1) Is this planned to replace the manual labor of people performing this classification?

Answer: The labor-intensive nature of handling soft-shell crabs arises from the rapid hardening of their new shells within three hours post-molting. As a result, each basket of crabs must undergo a thorough inspection within a 180-minute timeframe. Given that each inspection typically takes around 30 seconds, a single worker during a shift can manage approximately 360 baskets in a day. This repetitive task can become monotonous for farmers, leading to a desire for more efficient processes. It’s essential to emphasize that the implementation of this automated inspection system is not intended to replace human labor. Instead, it aims to address the challenges in current production methods, improving efficiency and potentially reducing the labor-intensive aspects of soft-shell crab handling. This information is added to Section 1 (Introduction) of the revised manuscript.

2) Can you show examples of images with detection annotations?

Answer: As suggested, a new Figure 3 is included in the revised manuscript. This figure illustrates detection annotations, with red boxes representing ground-truth boxes. The blue and green boxes depict the model’s predictions for the basket and crab, respectively.
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{\text{Intersection}}{\text{Detect box}}.$$

**Figure 3.** IoU calculation regarding Sample of True Positive (TP) and False Positive (FP). The image on the right illustrates detection annotations, with red boxes representing ground-truth boxes. The blue and green boxes depict the model’s predictions for the basket and crab, respectively.

3) Improve the captions of the Tables and Figures. Do not start with "The ...". For example, in Figure 1 instead of "The sample of images collected under three situations and two lighting conditions" it can be "Sample of images collected under three situations and two lighting conditions"

**Answer:** Thank you for your suggestion. In the revised manuscript, we revised the captions of all tables and figures while avoiding starting the captions with “The” (unless necessary).

Thank you very much again for your consideration. We hope that you find our responses satisfactory and that the revised manuscript is now acceptable for publication.