Research on Image Recognition of Marine Organisms Based on MIRNet and ViT Models

Xi Chen[1],Yuxi Luo[2], Wenjing Xu[3] and Jiayi Zhao[4,*]

[1]FEIT, University of Melbourne, Melbourne, VIC, 3052, Australia
[2]School of computing and data science,Xiamen University Malaysia ,Sepang, Selangor, 43900, Malaysia
[3]Software college, Jiangxi Agricultural University, Nanchang, Jiangxi Province, 330000, China
[4]Software Engineer, Dalian University of Technology, Dalian, Liaoning Province, 116000, China
*20212251035@mail.dlut.edu.cn

Abstract. With the deterioration of the marine environment, it is crucial to protect marine biodiversity. This paper implements and compares the performance of various models for their accuracy in species identification. ResNet50V2 obtained an accuracy of 79.62% on its validation set, according to the study's findings. 78.79% accuracy was attained by MobileNetV2 on its validation set. On the verification set, EfficientNetB7 achieved an accuracy of 73.57%. These models' performance was not as good as ViT's, which beat the competition with reduced loss rates and an accuracy of 91.51% on the training set and 90.23% on the validation set. Ultimately, the study sought to improve overall identification accuracy by improving low-quality photos. Subsequent studies using the MIRNetV2 model yielded greater results than the MIRNet model; it demonstrated strong image improvement capabilities, achieving an accuracy of 90.34% on the training set and 89.84% on the validation set. The findings suggest that improving image quality significantly enhances species identification accuracy, contributing to preserving marine biodiversity.

Keywords: Image Enhancement, Transfer Learning, ViT.

1 Introduction

In recent decades, the marine environment has faced numerous challenges, including microplastics absorbing organic pollutants, metals, and pathogens, posing threats to marine organisms. Marine Protected Areas (MPAs) are essential for preserving marine biodiversity, and tracking biodiversity over time is crucial.

In the realm of marine biodiversity conservation, traditional methods primarily rely on field surveys and manual records. While these methods are effective to some extent, they have the significant limitation of precision and the high cost of research. Image recognition technology has been widely applied in marine species identification and monitoring to overcome these limitations in recent years. The

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_62

identification of marine species has made greater use of image recognition technology in recent years due to the rapid advancements in computer vision and deep learning. Convolutional Neural Networks (CNNs), one type of deep learning model, are capable of automatically analyzing and processing vast amounts of photos of marine species, leading to accurate and efficient species identification. For instance, the ResNet50V2 and MobileNetV2 models have shown excellent performance in image classification and object detection [1,2]. The EfficientNetB7 model further improves recognition performance by adjusting the network's depth, width, and resolution, while the Vision Transformer (ViT) model demonstrates strong potential in image recognition through self-attention mechanisms [3,4].

The objective of this research is to employ sophisticated picture enhancement and segmentation methods to elevate the caliber of photos depicting marine animals, thus augmenting the precision of species identification. Specifically, the study employs various deep learning models to process and analyze marine animal images, including ResNet50V2, MobileNetV2, EfficientNetB7, and ViT. Comparing these models' performances is the goal, identify the most suitable image recognition algorithm, and evaluate its application effectiveness in improving marine biodiversity monitoring.

Additionally, the study employed a comprehensive stepwise selection method to enhance the quality of images in marine species datasets that exhibit poor performance in terms of contrast, brightness, color bias, and localized imbalances. By applying various image enhancement techniques, the study obtained different experimental results. These results were then combined with high-scoring images from the original quality assessment. Subsequently, the study compared the training and prediction results on the original dataset and the enhanced low-scoring image dataset using the ViT model. The comparison led to the final conclusion that the MIRNetV2 model demonstrated the best enhancement effect on the selected marine dataset, significantly improving the ability to identify the dataset.

2 Dataset and Methods

2.1 Data Sources

The dataset this article chose is a marine life dataset. It contains images of sea turtles, jellyfish, dolphins, and 19 different categories of marine life. Some of the images are from pixabay.com and some are from flickr.com. The image size is (300px, n), where n is the size of pixels smaller than 300px. This marine life image dataset has a balanced distribution of samples, which greatly simplifies the pre-processing process by eliminating the need for complex and tedious data imbalance processing steps. Meanwhile, the species diversity of the dataset provides the study with more trialability for model training, making the model training process more efficient and direct.

2.2 Feature Selection

Underwater images appear blurred, with low contrast and low resolution, because of the complex imaging environment under the ocean. Various image pre-processing methods are used to address this issue[5]. The study analyzed the images in the dataset by calculating brightness, contrast, color shift, sharpness, and anomalies, treating them as key features.

Contrast, which refers to the variation in brightness and darkness across an image, significantly impacts image clarity. High-contrast images appear sharper and more defined, making them easier to interpret. Conversely, low-contrast images may appear dark and blurry, hindering detail recognition. Optimal contrast enhances both visual quality and the performance of image recognition algorithms by aiding in the differentiation of target and background elements.

Brightness, representing the overall light intensity in an image, is crucial for preserving image detail and ensuring accurate recognition. Excessive or insufficient brightness can compromise image clarity, affecting the effectiveness of recognition algorithms. Proper adjustment of brightness highlights key features and enhances overall image readability and recognition.

Image sharpness, determining the visibility of object edges and details, greatly influences image interpretability. Sharp images are easier to understand and recognize, while blurry or distorted images reduce quality and recognition accuracy.

Color bias, which measures color distortion in an image, directly impacts visual appeal and realism, affecting recognition accuracy. Maintaining a balance between contrast, brightness, sharpness, and color bias is essential for effective image recognition. Adjusting these characteristics optimally can enhance image quality, thereby improving the accuracy and reliability of image recognition.

Using the dolphin dataset as an example, the study performed a univariate visualization of the above four eigenvalue parameters. The study found that the brightness values of the images are concentrated around 115, the contrast and color deviation values are concentrated around 35, and the sharpness values are concentrated around 500. This suggests that widespread images of marine organisms suffer from similar problems caused by the rapid decay of light in water, poor lighting control and ubiquitous organic debris. For this reason, this study also investigated the effect of correlation between parameters on the dataset.

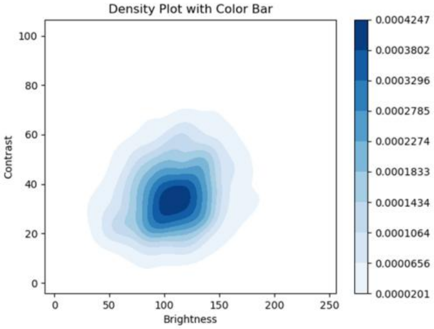


Fig 1. The correlation between brightness and contrast in chosen dataset (Photo/Picture credit : Original)

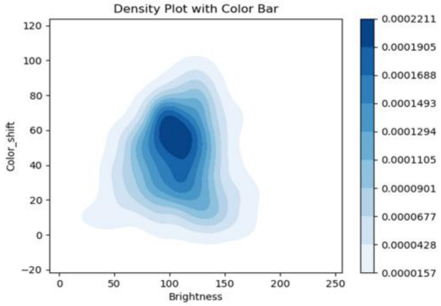


Fig 2. The correlation between brightness and color bias in the chosen dataset(Photo/Picture credit : Original)

According to the correlation analysis in Fig. 1 and Fig. 2, it can be seen that the correlation between these eigenvalue parameters is weak, and relatively speaking, brightness and contrast have a slightly greater impact on the dataset, so to diminish the impact of these two parameters, this study set the weights roughly as 0.9, 0.8, 1 and 1.

2.3 Data Pre-processing

Prior to data pre-processing, the study filtered using a 15% threshold and then statistically analyzed the counts and classifications for each species. According to Fig. 3, the dataset was relatively balanced overall, so this study omitted the data imbalance filtering step. In addition, to avoid the influence of single species data on the filtering results, the study randomly merged the datasets of different species before applying the filtering model. The comparison showed that the images to be filtered were not affected by the size of the image set, number of species and other factors, and the filtering results were almost undifferentiated, indicating that the weights were set reasonably and the filtering effect was good.

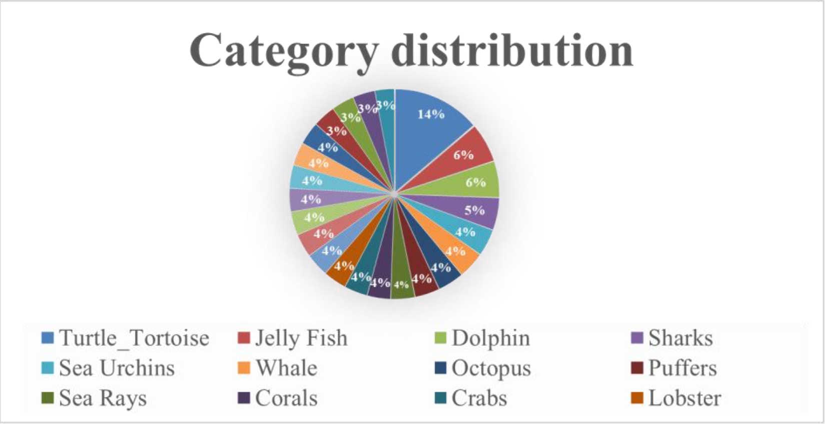


Fig 3. Category distribution (Pie chart)(Photo/Picture credit : Original)

By understanding and accurately analyzing the characteristics of marine data sets, this study can better optimize image quality and annotation. First, image problems such as overexposure or over-darkening are identified and resolved at a local scale. Then images with poor brightness, contrast, color variation and sharpness are filtered out. This study hypothesizes that the higher the normalized values of these features, the more accurate the model predictions will be, especially in the absence of poor-quality images. To fully assess image quality, this study normalized and combined these features into an 'overall' composite score and ranked the scores. Fig. 4 shows the images with the highest (overall score 2.352171) and lowest (overall score 0.737954) scores, where the difference in image quality is visible to the naked eye.

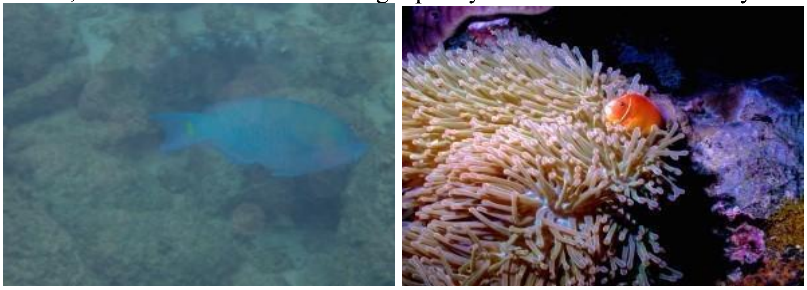


Fig 4. Examples of images with the highest (right) and lowest (left) scores(Photo/Picture credit : Original)

Based on Fig. 5, a critical image was output with a threshold of 15% in the fish dataset. (The left is the first image below the threshold the study set, and the right is the last image above the threshold). This study can see almost no difference between the two images for each feature, indicating that the threshold setting is consistent and reasonable.

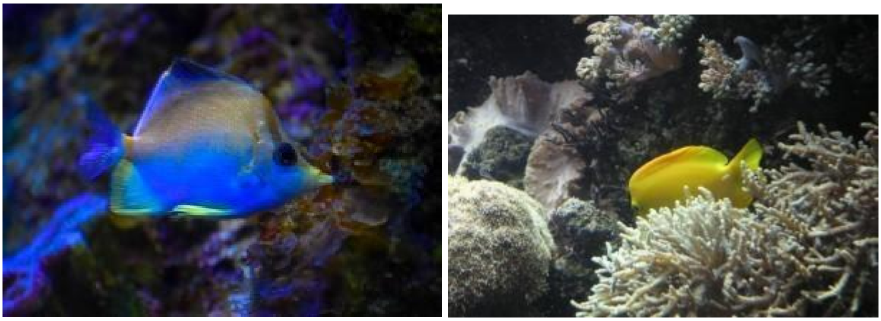


Fig 5. Comparison of images below and above the 15% threshold(Photo/Picture credit : Original)

2.4 Methods and Models

This article used ResNet50V2, MobileNetV2, EfficientNetB7, and ViT deep learning models to recognize marine organism images. The study used CNNs, MIRNet and MIRNetV2 models for enhancement of marine organism images.

Overview of the Models. ResNet50V2 introduces adjustments in the residual blocks by incorporating the 'bottleneck' structure and adding batch normalization and ReLU activation functions between layers [6].

MobileNetV2 introduces the Inverted Residual Block, which reduces computational complexity through channel expansion and compression.

EfficientNetB7 is a deep CNN model based on the EfficientNet architecture. EfficientNetB7 is one of the largest EfficientNet models in terms of depth, width and resolution.

The ViT model is a deep learning framework for image recognition based on the Transformer architecture.

The MIRNet model is a deep learning solution designed to significantly improve the quality of images captured under varying lighting conditions [7,8]. MIRNet uses a multi-scale approach with recursive residual groups and selective kernel feature fusion to dynamically improve low-light images.

MIRNetV2 is a state-of-the-art neural network designed to perform various image restoration and enhancement tasks. It retains the multi-scale feature aggregation strategy, but refines the architecture to make it more efficient in terms of computation and performance.

Model Training. For the ResNet50V2, MobileNetV2 and EfficientNetB7 models, this study used EarlyStopping and ModelCheckpoint callback functions to improve training efficiency and robustness.

ViT uses self-attention mechanisms to process image data [9]. This study mainly used this model for marine organism image recognition and made it more adaptable to the task requirements through transfer learning and fine tuning [10].

Image Quality Evaluation and Enhancement. This study identified 15 percentiles ranging from 0.85 to 0.15, ranked them by 'overall' score, and trained and predicted them using the ViT model to find the percentile eigenvalues that improve accuracy as an image enhancement criterion. 'Bad' images within the bottom 15% based on the 'overall' score were selected for enhancement and their performance re-evaluated using the same ViT model.

Finally, this paper explores the use of CNN and MIRNet models for image restoration and enhancement, especially for low-light and complex scenes. MIRNetV2 has an improved architecture to improve computational efficiency and performance further.

2.5 Evaluation

This study evaluates the model’s performance primarily using three metrics: Validation Loss, Accuracy, and F1-score. These metrics provide an overall comprehension of the model's effectiveness and efficiency in classifying the different categories of the sea animal images.

Accuracy, defined as the proportion of correct predictions out of the total predictions made in the dataset. It is defined as:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

This metric is straightforward and gives a general sense of how often the model is correct. This study chose accuracy because it is a fundamental measure of a classifier's performance, offering an overall picture of the model's effectiveness across all classes.

The F1-Score, which combines precision and recall into a single metric, is especially useful for imbalanced class distributions. It is calculated as the harmonic mean of precision and recall:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

Precision is the fraction of true positive predictions out of all predicted positives, while Recall is the fraction of true positive predictions out of all actual positives.

The study selected the F1-score because it provides a balance between precision and recall, which is critical when dealing with uneven class distributions. It ensures that the model performs well not only by making correct predictions but also by minimizing false positives and false negatives.

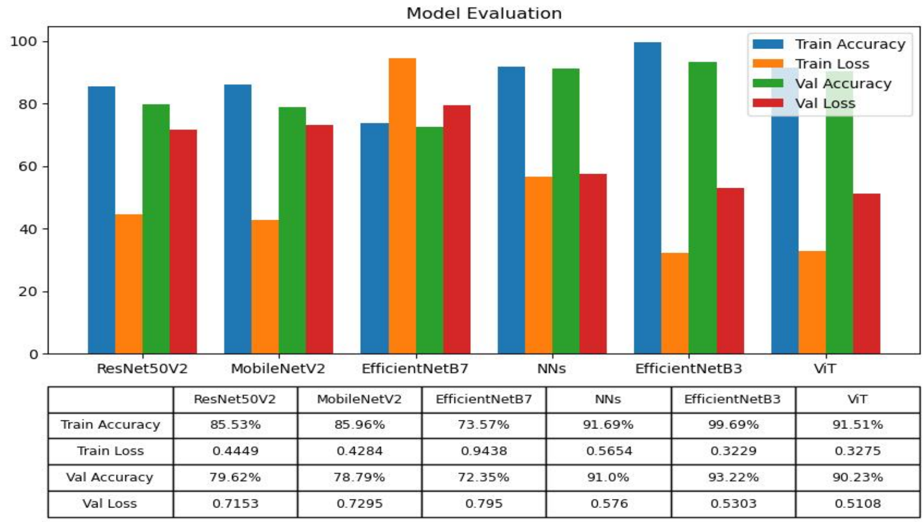
Validation Loss assesses the model's ability to generalize to new data. It is determined by evaluating the model on the validation dataset and calculating the loss. Lower validation loss signifies better generalization. This study monitors validation loss to ensure that the model does not overfit the training data and performs well on new, unseen images.

Using these three metrics together gives a well-rounded evaluation of the model, covering its accuracy and robustness against overfitting and class imbalance.

3 Results and Discussions

3.1 Results Before Enhancement

The study presents the evaluation results of various models used for classifying the sea animal images. The models are assessed based on Accuracy, F1-Score, and Validation Loss. Below is the summary table of the performance metrics for each model:



	ResNet50V2	MobileNetV2	EfficientNetB7	NNs	EfficientNetB3	ViT
Train Accuracy	85.53%	85.96%	73.57%	91.69%	99.69%	91.51%
Train Loss	0.4449	0.4284	0.9438	0.5654	0.3229	0.3275
Val Accuracy	79.62%	78.79%	72.35%	91.0%	93.22%	90.23%
Val Loss	0.7153	0.7295	0.795	0.576	0.5303	0.5108

Fig 6. Evaluation results of various models(Photo/Picture credit : Original)

From Fig. 6, the comparative analysis showed that while ResNet50V2 and MobileNetV2 had similar performance metrics, ViT offered the best balance between accuracy and loss rate. EfficientNetB7, although competitive, did not match the performance of ViT in this context. The EfficientNetB3 model achieved the highest accuracy at 93.22%. However, it also showed a very high training accuracy which is 99.69% and relatively high validation loss of 0.5303. These two metrics are crucial indicators of how well the model generalizes to unseen data. High validation loss and extremely high training accuracy suggest that the model might be overfitting the training data and not performing as well on new images.

Considering the balance between accuracy and generalization, this study selected the ViT model after image enhancement as the final classification model. The ViT model achieved a high accuracy of 90.23% and maintained the lowest validation loss at 0.5108. This indicates that the ViT model performs well on unseen data, making it a robust choice for the image classification task after applying image enhancement techniques.

3.2 Enhancement Methods

This study used MIRNet and MIRNetV2 to enhance images. MIRNet enhances image quality using fixed convolutional kernels, which apply the same parameters to all input images. Its straightforward multi-scale feature fusion approach merges features from different scales but may not effectively retain detailed and global information. This can lead to suboptimal enhancement effects.

In contrast, MIRNetV2 introduces several advancements. It employs dynamic convolution operations, adjusting convolutional kernel parameters based on the specific input image. This allows for more flexible and context-aware processing. Additionally, MIRNetV2 utilizes a more sophisticated feature fusion mechanism, enabling more effective integration of multi-scale information. It also incorporates advanced attention mechanisms, such as the Non-Local Attention Mechanism, which better capture and merge important features [11]. These enhancements allow MIRNetV2 to better retain and integrate details and global information, leading to improved image enhancement performance over MIRNet.

3.3 Results After Enhancement

Finally, the study compared the results of training and predicting with the ViT model on the original dataset and the dataset where the lower-tier images had been enhanced. This comparison aimed to assess the overall impact of data enhancement strategies on model performance.

As shown in Table 1, the results after enhancing images with MIRNet did not achieve a higher accuracy compared to the unenhanced images. This study believes the reason for this outcome is that the MIRNet model was specifically trained to enhance images from the LOL dataset, which may not align perfectly with the characteristics of the dataset. However, in the MIRNetV2 model, the study got better results than before the enhancement.

Table 1. The comparison of results after being enhanced by MIRNet and MIRNetV2

Metrics	Unenhanced	MIRNet	MIRNetV2
Accuracy	0.8982	0.8663	0.8984
F1-Score	0.8854	0.8535	0.8813
Validation Loss	0.452	0.677	0.370

For Performance Evaluation, each model was trained with appropriate callback functions, including EarlyStopping and ModelCheckpoint, to monitor training progress and save the best-performing models. The models were evaluated based on their accuracy on training and validation sets.

3.4 Discussion and Suggestions

From Table 1, it can be observed that the metrics of images enhanced by MIRNet are inferior to those of unenhanced images, and the parameters of images enhanced by MIRNetV2 are only slightly better than those of unenhanced images. To analyze this issue, some images for the three scenarios were exported. As shown in Fig. 7, it is evident that MIRNet performs poorly compared to Unenhanced and MIRNetV2. This report believes this is due to inherent deficiencies in the MIRNet model. Specifically, this may include but is not limited to the following: MIRNet uses fixed convolutional kernels to process all input images, whereas MIRNetV2 introduces dynamic convolution operations that adjust kernel parameters dynamically based on the input images. The feature fusion mechanism in MIRNet is relatively simple and may not sufficiently retain detailed and global information when merging features at different scales, resulting in suboptimal enhancement effects. MIRNetV2 optimizes the feature fusion mechanism, enabling more effective integration of multi-scale information and improving feature representation capabilities. Additionally, MIRNetV2 introduces more powerful attention mechanisms, such as the Non-Local Attention Mechanism, which better capture and integrate important features, enhancing image enhancement effects [11].

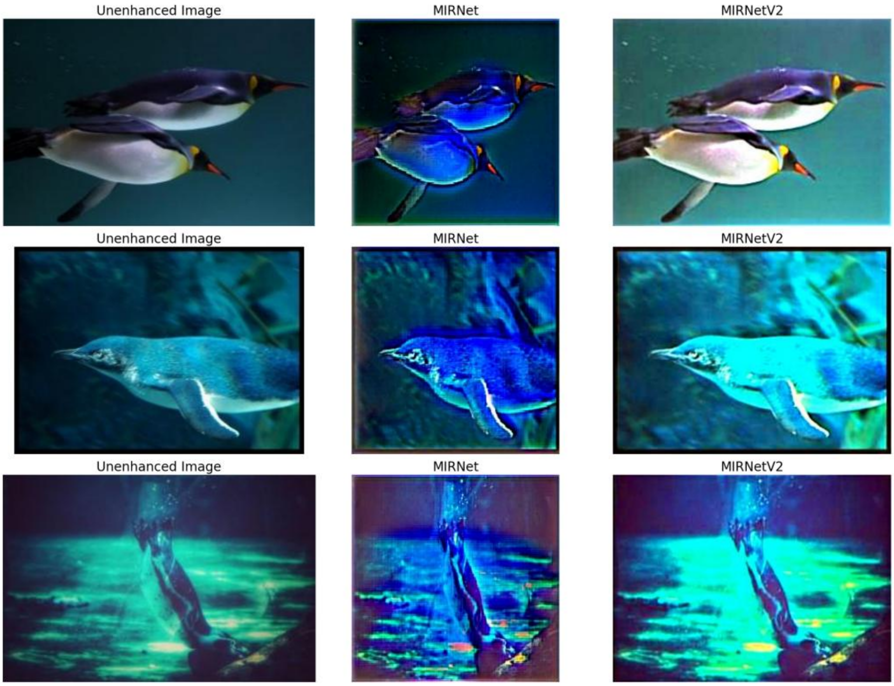


Fig 7. Image results of enhanced and unenhanced cases

By comparing Unenhanced and MIRNetV2, it is found that MIRNetV2 performs well in enhancing images under black dark light conditions, as shown in the first row of penguin images. However, when the light is deep blue or green, the images enhanced by MIRNetV2 may be overexposed or fail to enhance properly. Analyzing the training set and training process of the MIRNetV2 model reveals that the input images enhanced by MIRNetV2 exhibit characteristics of overall low brightness and a tendency towards black colors, as shown in Fig. 8. Therefore, this is the main reason why the enhancement results of ocean images are unsatisfactory. Although dark conditions exist underwater, blue dominates the color of underwater images, not the black found in the MIRNetV2 training set.

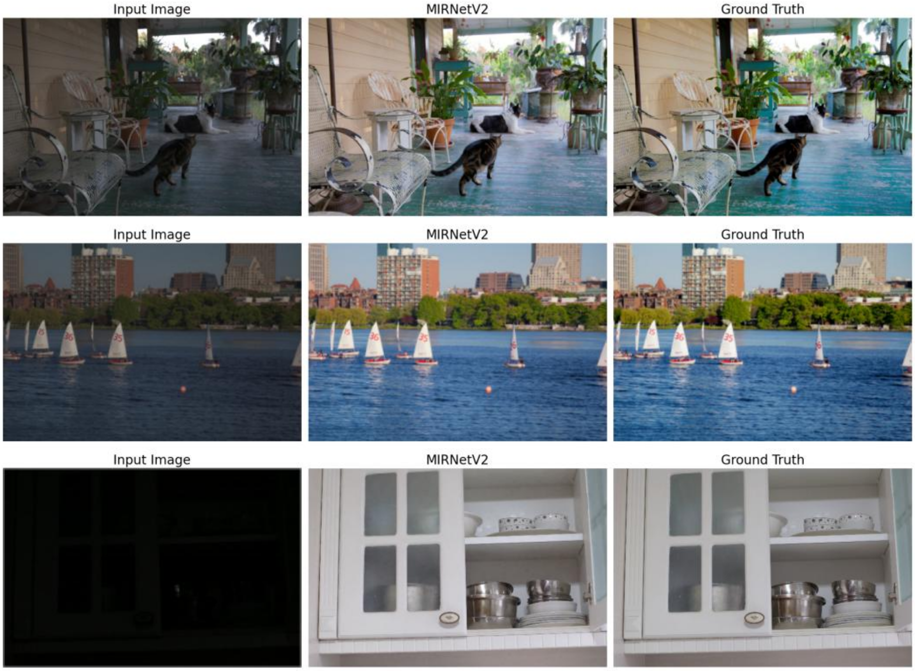


Fig 8. Image results of enhanced and unenhanced cases fromMIT-Adobe FiveK and LoL dataset [11]

The study finds that the ViT model has dependencies and high computational cost when dealing with small datasets and there are also many tuning strategies in existing Vit models [12]. Although the experiments in this report did not significantly improve prediction accuracy using MIRNetV2-enhanced images, this report demonstrates that with unenhanced input images and manually adjusted ground truth images specific to the ocean environment, a model trained using MIRNetV2 methods can achieve a much better prediction results.

Furthermore, by comparing MIRNetV2 with MIRNet, we can see significant improvements in various metrics of MIRNetV2. Therefore, this report proves that image enhancement technology has great potential in improving image prediction accuracy.

4 Conclusion

The study aimed to enhance the identification accuracy of marine species by addressing common image quality issues such as poor contrast, brightness, color bias, and sharpness. A dataset of photos of marine animals was subjected to sophisticated image improvement and segmentation techniques, and the study assessed the effectiveness of several deep learning models, such as ResNet50V2, MobileNetV2, EfficientNetB7, and ViT.

Several deep learning models were trained and put into practice by the study. The study compared the performance of ResNet50V2 and MobileNetV2 in image categorization and task-based object detection. Using the initial training set, ResNet50V2's accuracy was 85.53%, while on the validation set, it was 79.62%. This demonstrates good overall efficiency, albeit with a relatively high loss rate.With an accuracy of 85.96% on the initial training set and 78.79% on the set for validation, along with a significant loss rate, MobileNetV2 likewise showed comparable performance. Utilizing EfficientNetB7, which achieved a precision of 73.57% on the set used for validation and 72.35% on the initial training set, the advantages of scaling network depth, width, and resolution were investigated. Ultimately, the ViT demonstrated strong feature extraction and effective learning capabilities by outperforming other models with higher precision of 91.51% on the data used for training and 90.23% on the set used for validation, as well as lower validation loss. This was achieved by using self-attention mechanisms to improve image recognition.

The study employed techniques for image improvement and segmentation. These approaches are crucial for accurately identifying species because they increase the quality of photos with poor contrast, brightness, color bias, and sharpness. The findings indicate that advanced image enhancement techniques, coupled with sophisticated deep learning models like ViT, have a significant potential to improve the accuracy and efficiency of marine species identification. This improvement is crucial for effectively monitoring and preserving marine biodiversity within MPAs.

In conclusion, this study comprehensively evaluates various deep learning models for marine species identification and highlights the importance of image quality enhancement. Future studies and applications in the conservation of marine biodiversity can build on the techniques and findings reported in this work.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. He, K., Zhang, X., Ren, S., & Sun, J.: Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778 (2016).
2. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. :MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510-4520 (2018).
3. Tan, M., & Le, Q. V. : EfficientNetV2: Smaller Models and Faster Training. In Proceedings of the 38th International Conference on Machine Learning (ICML), 10096-10106 (2021).
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929 (2020).

5. Qi Q., Li K., Zheng H., Gao X., Hou G., Sun K.: SGUIE-net: semantic attention guided underwater image enhancement with multi-scale perception. IEEE Trans. Image Process. 4, 6816–6830 (2022).
6. Shao D, Shao X, Liu P, Zhao C, Tao Q. Infrared and visible light image fusion algorithm based on ResNet50 and convolutional sparse representation. Computer Applications and Software, 41(05): 189-196 (2024).
7. Purbandini, Fatichah C. and Amaliah B.,:Digital Image Enhancement Using MirNet and Zero-Deep Curve Estimation (Zero-DCE), 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Purwokerto, Indonesia, pp. 290-295,(2023).
8. Luo L., Yu L. & Zheng M.: MIRNet-Plus: An improved method for low-light image enhancement based on rich feature learning. Microcomputer Systems (03),664-669 (2024).
9. Abdallah M, Younis S, Wu S & Ding X.: Automated deformation detection and interpretation using InSAR data and a multi-task ViT model. International Journal of Applied Earth Observation and Geoinformation, 128 (2024).
10. İncir R ,Bozkurt F . Improving brain tumor classification with combined convolutional neural networks and transfer learning. Knowledge-Based Systems, 299 (2024).
11. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2022). Learning enriched features for fast image restoration and enhancement. IEEE transactions on pattern analysis and machine intelligence, 45(2), 1934-1948.
12. Khan M., Naseer M., Hayat S., Ali M., Shahzad F., Khan M., and Porikli F., :Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work,"arXiv preprint arXiv:2203.01536, 2022. Retrieved from arXiv (2023).