Research on Fur-seal Face Recognition

Zhang Xingzhe/张星喆[*]
09020115
1357738012@qq.com

Chen Minheng/陈闵恒[†]
09020119
minhengchen@qq.com

Abstract

Human face recognition is considered to be one of the hottest fields at the moment，the recognition methods based on machine learning have achieved good performances. Some scholars are now trying to apply these method to animals. This report is about our group project in pattern recognition class which is related to Fur-seal face recognition. In this report, we first talk about the existing method for object detection and face recognition, and then we will introduce our proposed method. For evaluation, experiment were performed on a task dataset provided by unit chair. Experiment results show that the performance of the proposed method on the dataset is acceptable.

1 Introduction

1.1 Object Detection

Object detection is a research hotspot in computer vision and is in demand for applications in many fields, such as surveillance security, autonomous driving, traffic monitoring and robot vision scenarios(1). Object detection generally involves the detection of some predefined classes of target instances (e.g. people and vehicles, etc.).

Traditional target detection relies on elaborate manual feature design and extraction(1; 2), such as the Histogram of Oriented Gradient (HOG)(3). 2012, AlexNet, based on deep convolutional neural network (CNN), won the ImageNet image recognition competition with a significant advantage, and deep learning has been gaining widespread attention since then. Target detection has also gradually entered the era of deep learning.

[*] School of computer science and engineering , Southeast University, Nanjing, China.

[†] School of computer science and engineering , Southeast University, Nanjing, China.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

In 2014, Girshick et al.(4) proposed the R-CNN (region with CNN feature) network at the CVPR (IEEE conference on computer vision and pattern recognition) conference. He K et al. added a spatial pyramid polling layer in front of the fully connected layer of the network, which makes the network not require normalized image size for the fully connected layer input. The extracted features have better scale invariance, reducing the possibility of overfitting, and improving the mean average precision (mAP) to 53.3% on the VOC (visual object classes) 2012 dataset. In 2015, Girshick et al.(5) and Ren et al.(6) proposed Fast R-CNN and Faster R-CNN models to improve the detection performance of the model. In 2016, Redmon et al.(7) proposed YOLO (You only Look Once) V1 with 45 FPS (frames per second) to truly achieve the speed of real-time video detection, which pointed the way for real-time detection of video motion targets; Liu et al.(8) proposed SSD (Single Liu et al. proposed the SSD (Single Shot MultiBox Detector) object detection model, which was higher than YOLO V1 in terms of real-time and accuracy, and achieved 74.3% mAP and 46FPS in the VoC dataset; subsequently, Redmon et al. proposed YOLO V2, which improved the mAP to 78.6% in the VOC dataset and significantly increased the detection speed to 67 FPS in real-time detection. At the same time, CNN target detection based on region suggestion and weakly supervised learning methods based on coupled CNNs have also been proposed, resulting in a significant reduction in tagging cost.

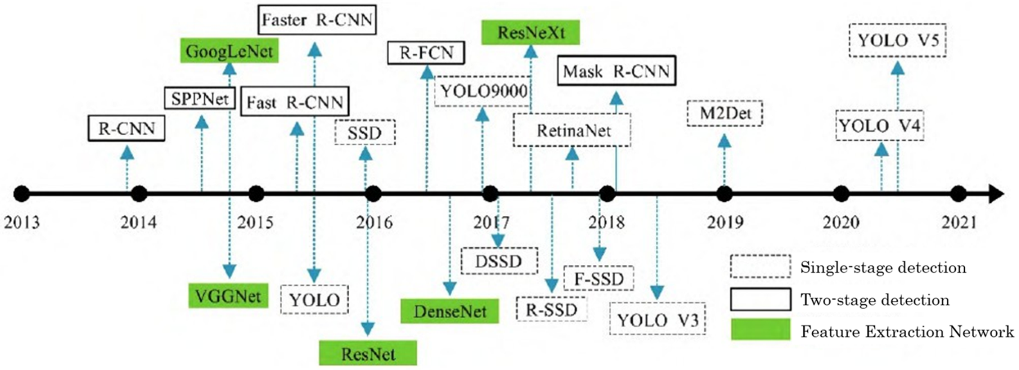


图 1: Mainstream algorithm development history

The two-step detection algorithm based on region suggestions has a good performance in terms of accuracy, but cannot reach the requirements of real-time detection in terms of detection speed. The regression-based algorithm makes concessions in accuracy compared to the area-based suggestion algorithm, but improves greatly in detection speed and can fully meet the requirements of real-time detection(9). The YOLO algorithm is comparable to the SSD algorithm in terms of performance during the YOLO v3 period, but the subsequent YOLO v4(10) and YOLOv5(11), which have been introduced, have far surpassed the SSD algorithm in terms of performance.

1.2 Fur-seal Face Recognition

Recognition of animals in the wild is critical for understanding the evolutionary processes that guide biodiversity. Researchers must reliably recognize each individual animal in order to observe that animal's variation within a population. Unique appearance-based cues, such as body size, presence of scars and marks, and coloring, are often used for interim studies(12; 13), but these attributes are subjective and vary over time. Therefore, they are unreliable in longitudinal studies, which are necessary for the study of long-term population health and behavior, group dynamics, and the heritability and effects of traits(14).

Biologists and anthropologists have started to adopt more objective and rigorous tracking methods, such as collars or tags. While these approaches have been successfully used in several long-term in the Wild marine animals studies, Such as studies of seal behavior, they are problematic in a number of ways. First, the devices can be expensive and time-consuming to apply. Second, tagging requires capture of the animal, which has demonstrably negative effects - it can disrupt social behavior, and cause intense stress, injury, and even death. In contrast, automatic facial recognition is a promising method to accurately identify individuals with minimal risk to these already threatened species.

This project is mainly inspired by the recent development of deep learning in face verification: does these two pictures represent the same individual?, in face recognition: who is this individual? and in face clustering: group these pictures by individuals (15). Therefore, our work aims thus at developing the existing methods by using recent deep learning research in face verification and recognition on a novel dataset of seal faces.

2 Methods

The target of our project can be divided into two parts, face detection and face recognition. In face detection part we need to detects all the faces from a image that containing fur-seals. And in face recognition we need to determines which faces are from the same fur-seal on an unlabeled data set (task dataset). If done appropriately, your model can help answer the following questions: How many individual fur-seals in a given set of photos?Which fur-seals are often seen on the same beach at the same time (using image meta data)? or for simplification, appear on the same images. What is the potential relationship between those co-occurred fur-seals?

The dataset of our task are provided by our unite chair which contains 74 fur-seal images(416X416). A labelled data set available at Github and its format is completely consistent with the public MS COCO dataset. There are 96 images in the dataset,

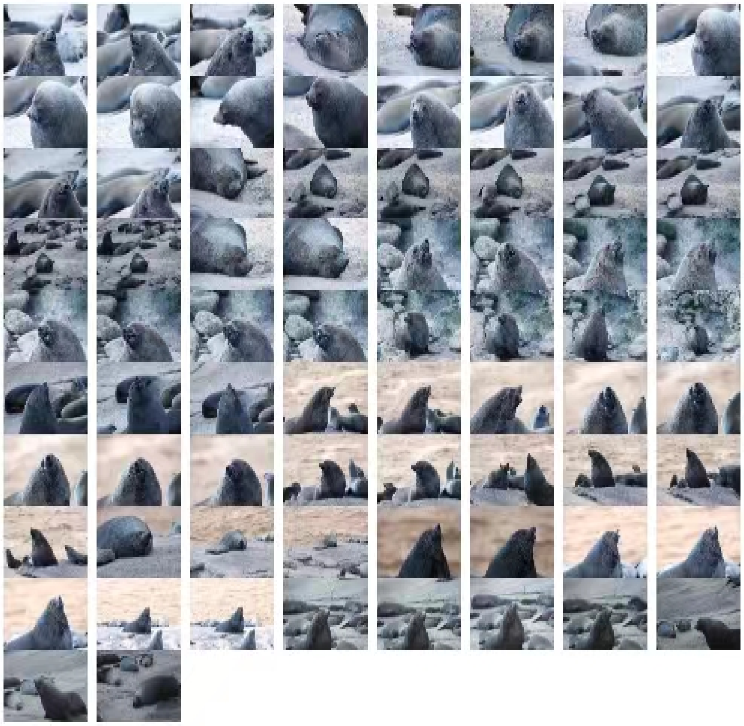


图 2: Task Dataset

which have be divided into three folders. All the images have been highly augmented in advance.

2.1 Face Detection

The basic model we decide to use in this part is YOLO.YOLO is believed to be the SOTA model in object detection fields, it has many versions, In our experiment we evaluate the efficiency of two models belongs to YOLO series, YOLOv5(11) and YOLOv7(16). And we find that YOLOv5 performs better on our dataset. The labelled fur-seal dataset only contains very few samples, so it's impossible to use the conventional method to detect object. We believe that this is a few-shot learning problem. Few-shot learning is the application of Meta Learning in the field of supervised learning. Meta Learning, also known as learning to learn, decomposes the data set into different meta tasks in the meta training stage to learn the generalization ability of the model when the category changes. So in the meta testing stage, when facing a new category, the classification can be completed without modifying the existing model.

In YOLOv5 model, they provide a meta learning method by using hyperparameter evolution.Hyperparameter evolution is a method of Hyperparameter Optimization using a Genetic Algorithm (GA) for optimization. Hyperparameters in ML control various aspects of training, and finding optimal values for them can be a challenge. Traditional

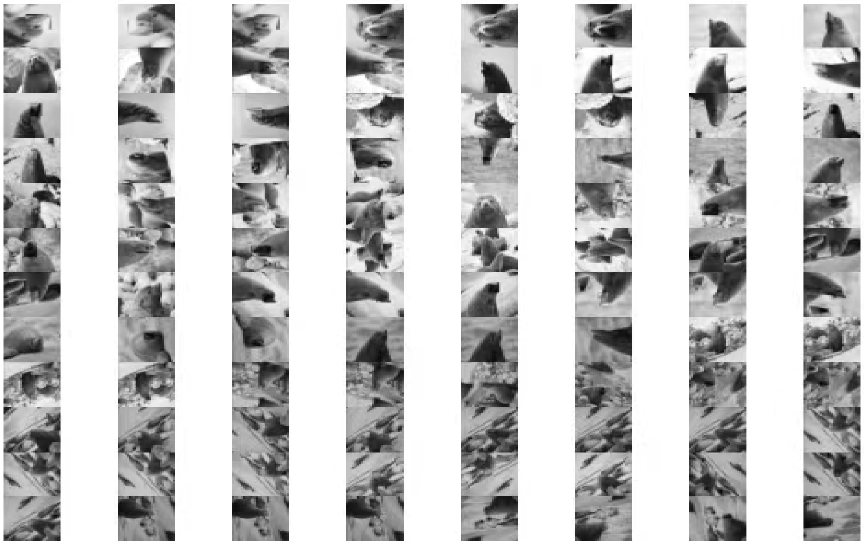


图 3: Labelled Dataset

methods like grid searches can quickly become intractable due to 1) the high dimensional search space 2) unknown correlations among the dimensions, and 3) expensive nature of evaluating the fitness at each point, making GA a suitable candidate for hyperparameter searches.

However, hyperparameter evolution could obtain a better set of model initialization parameters, which means let the model learn to initialize itself. Face detection is a complicated task, and so does the YOLO network. We have to find a way to achieve the task without overfitting on small amount of data. Transfer learning is the ability of a system to recognize and apply knowledge and skills learned in previous domains to novel domains. We propose to apply transfer learning method training on the previous pretrained YOLO models.

2.2 Face Recognition

For a face recognition assessment, it can be divided into two steps, feature extraction and feature clustering. In the first step, we extract the feature from the fur-seal faces which we have detected by using the detector we designed and mentioned in para 2.1.

HOG(Histogram of Oriented Gradient)(3) feature has been proved efficient in human detection. The basic idea is that local object appearance and shape can often be characterizedrather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In practice this is implemented by dividing the image window into small spatial regions which are called “cells”, for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation. For better invariance to illumination, shadowing, etc., it is also useful to contrast-normalize the local responses before using them. This can be done by accumulating a measure of local histogram "energy" over somewhat larger spatial regions which are called "blocks" and using the results to normalize all of the cells in the block. We will refer to the normalized descriptor blocks as Histogram of Oriented Gradient (HOG) descriptors.The process of extracting HOG features in our method can be divided into three steps. First, we need to resize it into a picture with the same size as the input image of the detection network, then convert it from RGB image to grayscale image, and finally Then extract its HOG features from this grayscale image.

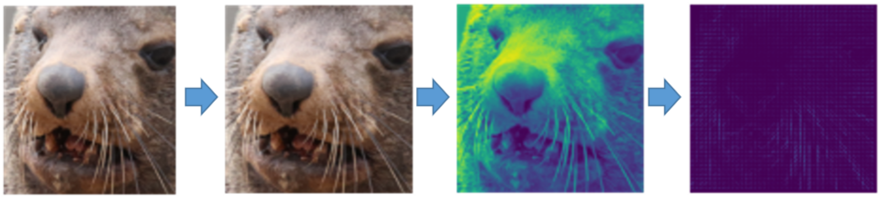
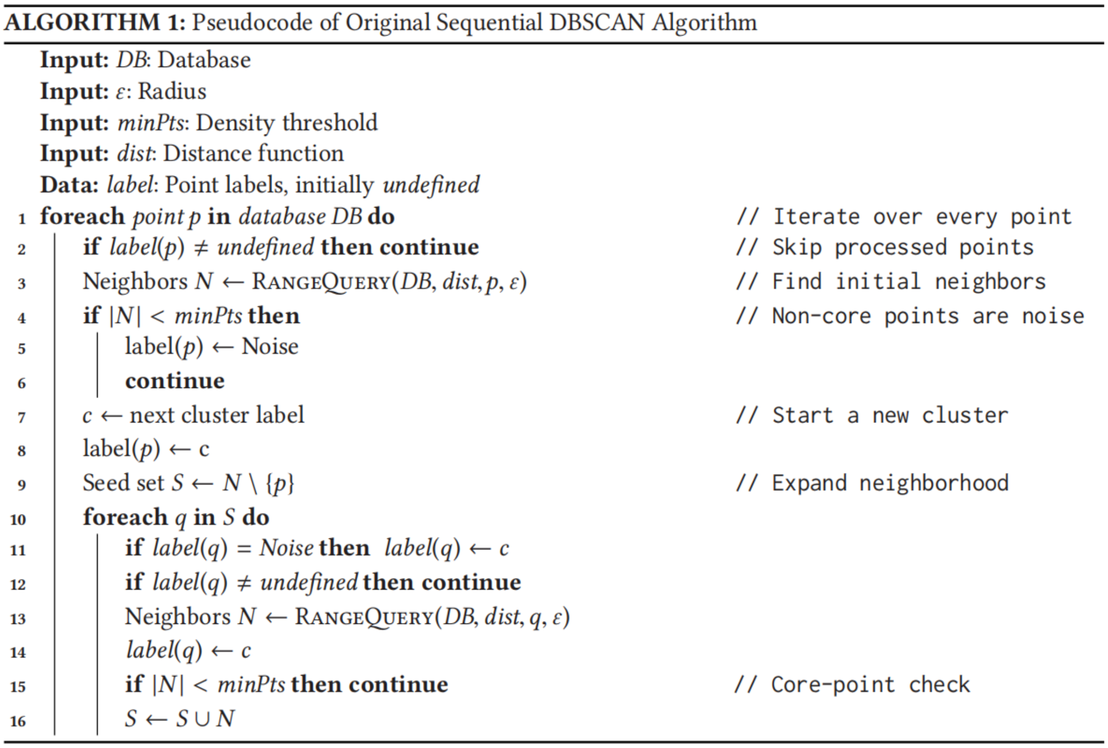


图 4: Flow Chart of Extracting HOG from Fur-seal Face

Feature extracted by CNN is another feature we applied in our method. A pre-trained convolutional neural network can map the image to the corresponding feature space very well. So if we can find an appropriate pretrained CNN model such as google facenet or model pretrained by Imagenet, We can get the ideal embedded vectors.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)(17)(18) is a density-based spatial clustering algorithm. The algorithm divides regions with sufficient density into clusters and discovers clusters of arbitrary shape in a noisy spatial database, which defines a cluster as the largest collection of density-connected points. We believe that this unsupervised density-based algorithm is suitable in the feature clustering stage. The benefit of being based on density is that other method such as k-means clustering algorithm can only handle spherical clusters, that is, a solid cluster. this is because the algorithm itself is limited in calculating the average distance. But often there are various shapes in reality. At this time, those traditional clustering algorithms are obviously tragic.

The method for exploring potential relationship between these fur-seals is based on an assumption. If two seals appear in the same photo multiple times, we consider them to belong to the same family/group.



ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

```
Input: $DB$: Database
Input: $\varepsilon$: Radius
Input: $minPts$: Density threshold
Input: $dist$: Distance function
Data: $label$: Point labels, initially $undefined$
foreach point $p$ in database $DB$ do                          // Iterate over every point
    if $label(p) \neq undefined$ then continue                 // Skip processed points
    Neighbors $N \leftarrow$ RangeQuery($DB, dist, p, \varepsilon$)   // Find initial neighbors
    if $|N| < minPts$ then                                     // Non-core points are noise
        label($p$) $\leftarrow$ Noise
        continue
    $c \leftarrow$ next cluster label                          // Start a new cluster
    label($p$) $\leftarrow$ c
    Seed set $S \leftarrow N \setminus \{p\}$                  // Expand neighborhood
    foreach $q$ in $S$ do
        if $label(q) = Noise$ then $label(q) \leftarrow c$
        if $label(q) \neq undefined$ then continue
        Neighbors $N \leftarrow$ RangeQuery($DB, dist, q, \varepsilon$)
        $label(q) \leftarrow c$
        if $|N| < minPts$ then continue                        // Core-point check
        $S \leftarrow S \cup N$
```

图 5: DBSCAN Algorithm

3 Experiments

The hyperparameter has been evolved for about 100 generations, and the best generation is the 28 epoch. After hyperparameter evolution, we start to train the YOLO model by using transfer learning. YOLOv5 provides many different version, they are

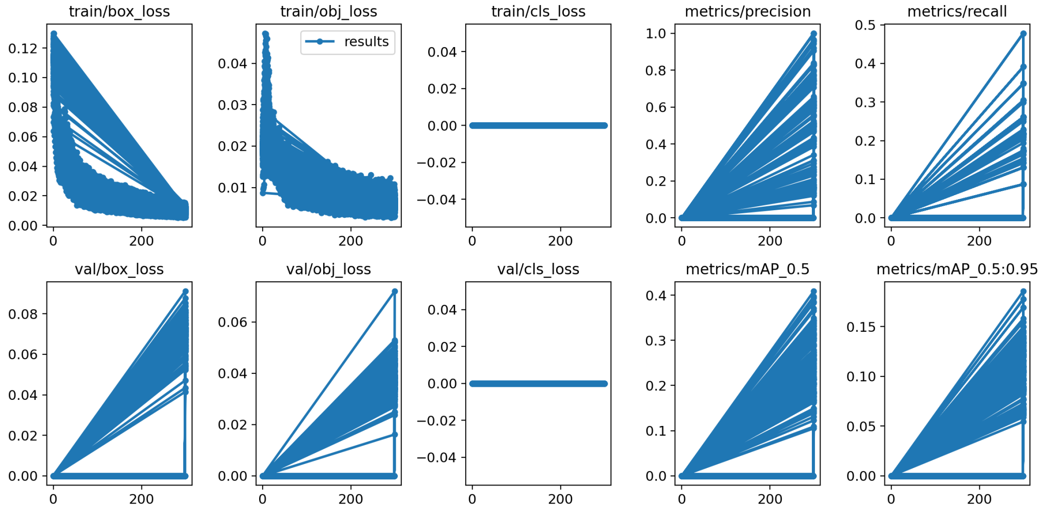


图 6: Evolution loss

different in depth of the network and the accuracy of prediction. The training dataset has been mentioned in para 2. The training version is YOLOv5x, which is the deepest network and has the largest amount of parameters. Our model and other comparative deep learning models are all implemented with Pytorch and trained on RTX 3090 GPU. We set learning rate and other hyperparameter by the result of hyperparameter evolution. The batchsize is 42, epoch is 10000 while with a patience parameter 300 which means if the result haven't got any progress in 300 epoch the will be terminated.

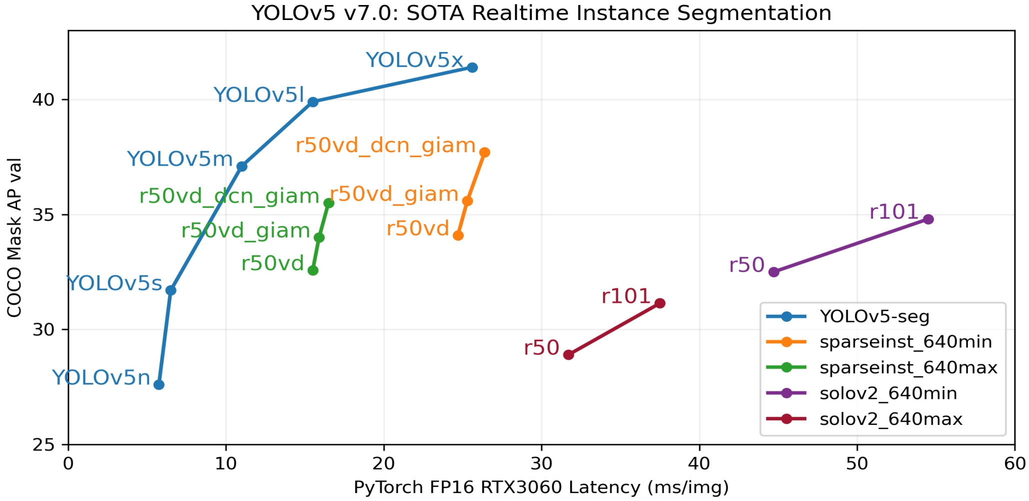


图 7: YOLOv5 model

We have tried to translate the image for validation from RGB to gray scale. But actually this approach doesn't make much sense. We stop training at about 1000 epochs. And it performs exellent on test data. We use the last model make a detection on the taske dataset and find 86 possible faces. Then we start to findthe most suitable parameter for DBSCAN clustering approach. We have two types of feature, HOG and embedded vector extracted by CNN, because their dimension and characteristic have enormous discrepancy. So we need to find their parameter separately. We defined the distance as Minkowski distance to the fourth power. Becasue for HOG feature its hard

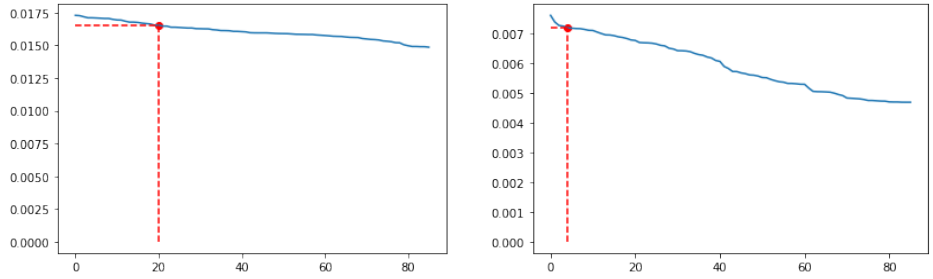


图 8: best parameter for HOG(left) and embedded vector(right) to find distinct clusters，so we only used the embedded vector extracted by CNN. The CNN model used in our project is VGG-face, which is a pretrained model provided by DeepFace package. We set eps=0.022 and min samples=2. We find 10 fur-seals after clustering. We named them fur-seal 1, fur-seal 2 e.g. in turn. And find their potential relationship. We find that fur-seal 0 is often seen on the same picture and fur-seal 1 and fur-seal 6 are family, fur-seal 1 and fur-seal 9 are family,fur-seal 2 and fur-seal 3 are family.

4 Conclusion and Discussion

In this report, we presented our method for fur-seal recognition, which has two stage face detection and face recognition. And with our method, we can answer the three question:How many individual fur-seals in a given set of photos?Which fur-seals are often seen on the same beach at the same time (using image meta data)? or for simplification, appear on the same images. What is the potential relationship between those co-occurred fur-seals?

But our method our method has certain limitations and flaws. Our detector cannot find all the faces in the dataset. The possible reason is the leakage of training samples. Besides, the clustering result havn't split all the fur-seals. Many sample points are still clustered together in the category of fur-seal 0, and have not been separated. We think the main reason for this is that the features extracted by the feature extraction module are not clear enough, which makes us spend a lot of effort to achieve clustering. If some larger and deeper pre-trained network models are used, or a classifier specially trained with fur-seal faces, the effect should be improved to a certain extent. In short, from the experimental test results, we can see that the method we proposed has been able to realize the basic function of fur-seal face recognition, and has a certain degree of robustness.

References

[1] Liu.L, O.Wanli, W. Xiaogang, et al. Deeplearning for generic object detection: A survey[J].International Journal of Computer Vision, 2020, 128(2):261–318. doi:10.1007/s11263-019-01247-4.

[2] Zou Zhengxia, Shi Zhenwei, Guo Yuhong, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv :1905.05055,2019.

[3] Dalal, N., Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). Ieee.

[4] Girshick R, Donahue J, DARREL T, et al.Rich feature hierarchies for accurate objeot detection and semantic segmentation[C]12014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus,OH, USA.IEEE,2014:580-587.

[5] Girshick R.Fast R-CNN[CJ12015 IEEE International Conference on Computer Vision. Santiago, Chile. IEEE,2015:1440-1448.

[6] Ren S Q, H.Girshick R, et al.Faster R-CNV.towards real-ime object detection with region proposal networks[J. IEEE Trans Pattern Anal Mach lntell, 2017, 39(6):1137-1149.

[7] HE K M.Gckiloxarl G, Dolar P, et a Mask R-CNNC]I120171EEE Intermnational Conference on Computer Vision. Venic, italy. IEEE, 2017:2980-2988

[8] W.XL, Shrivastavaa, Gupta A.A-fastRCNN.hard positive generation via adversary for object detection[C]12017IEEEConference on ComputerVision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017: 3039-3048.

[9] Shao Yanhua, Zhang Duo, Chu Hongyu, Zhang Xiaoqiang, Rao Yunbo. Overview of YOLO target detection based on deep learning [J]. Journal of Electronics and Information, 2022,44(10):3697-3708.

[10] Bochkovskiy A, Wang C Y, and L. H Y M.YOLOv4:Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[11] Jocher G, Stoken A,Borovec J, et al.Ultralytics/YOLOv5:V3.1-bug fives and performance improvements [EB/OL.hips:/do.org10.5281/zenodo.4154370, 2020. doi:10.5281/zenodo.4154370, 2020.

[12] Carson M Murray, Margaret A Stanton, Kaitlin R Wellens, Rachel M Santymire, Matthew R Heintz and Elizabeth V Lonsdorf, "Maternal effects on offspring stress physiology in wild chimpanzees", American Journal of Primatology.

[13] Serge A Wich, S Suci Utami-Atmoko, T Mitra Setia, Herman D Rijksen, C Schurmann, J.A. Van Hooff, et al., "Life history of wild sumatran orangutans (pongo abelii)", Journal of Human Evolution, vol. 47, no. 6, pp. 385-398, 2004.

[14] Tim Clutton-Brock and Ben C Sheldon, "Individuals and populations: the role of long-term individual-based studies of animals in ecology and evolutionary biology", Trends in Ecology Evolution, vol. 25, no. 10, pp. 562-573, 2010.

[15] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. CoRR abs/1503.03832 (2015). http://arxiv.org/abs/1503.03832

[16] Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.

[17] Ester, M., Kriegel, H. P., Sander, J., Xu, X. (1996, August). Density-based spatial clustering of applications with noise. In Int. Conf. Knowledge Discovery and Data Mining (Vol. 240, No. 6).

[18] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21.