Identifying Pufferfish Specie Using Deep Neural Networks And Face Embedding Method

Yuan Lin✉[1], Shaomin Xie[2], Jari Korhonen[3], Juan Liu✉[2], Xiangrong Liu[2], Junyong You[5], and Debasish Ghose[1]

[1] School of Economics, Innovation, and Technology, Kristiania University College, Kalfaveien, Bergen, Norway
`yuan.lin, debasish.ghose@kristiania.no`

[2] School of Information Science and Technology, Xiamen university, China
`xrliu@xmu.edu.cn, xsmin@stu.xmu.edu.cn`

[3] University of Aberdeen, King's College, Aberdeen AB24 3FX, United Kingdom
`jari.korhonen@abdn.ac.uk`

[4] School of Arsospace Engineering, Xiamen university, China
`cecyliu@xmu.edu.cn`

[5] Norwegian Research Centre (NORCE), Bergen, Norway
`juyo@norceresearch.no`

Abstract. Pufferfish, acclaimed for its distinctive texture and extraordinary delicacy, is however notorious for the highly toxic poison. Identifying individual pufferfish can be very beneficial for the aquaculture and food processing industries, as it tackles the challenges of food security and nutrition strategies, as well as maintenance of a sustainable ecosystem. Current methods of identifying and tracking pufferfish mainly rely on heuristic visual recognition or manual intervention such as RFID. The rapid advances in deep learning together with the presence of large scale database, are now able to solve complex tasks that previously required human expertise. In this work, we have implemented a deep learning framework based on deep Face Recognition (deep FR) techniques, to identify individual pufferfish. First, we created a dataset of labeled and data augmented takifugu bimaculatus fish images, which is publicly accessible as benchmark for interested researchers. Second, we conducted an extensive evaluation of state-of-the-art building blocks of Deep FR, in particular segmentation and loss functions, and conducted an ablation study for their applicability to pufferfish recognition. Third, we proposed a framework named FishIR composed of four deep FR stages. Experiments verified the effectiveness of this framework in terms of learning useful representation of individual pufferfish specie based on the back skin texture pattern. We believe that this approach can generalize to other similar individual recognition tasks, as well as contribute to the massive growth of smart farming and deep ocean fishery.

Keywords: Fish recognition · Deep Face Recognition · Convolutional neural networks.

1 Introduction

Pufferfish, rated for its extraordinary delicacy, distinctive texture, and high nutrition value, has consistently gained popularity as luxury food ingredient especially in eastern hemisphere over centuries. However, as widely known, pufferfish may contain a potent and deadly toxin called tetrodotoxin in the liver and ovarie and not able to be destroyed by cooking procedure. As fish and seafood have grown to be a primary source of protein and essential nutrients, there is a greater consumer awareness about security in the food supply chain. It is difficult to obtain the reliable food information in the complex food supply system which involves multiple economic stakeholders, thus easily lead to food fraud and food safety problems [1].

To tackle the aforementioned problem, methods of identifying and tracking individual pufferfish instance can be very beneficial for the aquaculture and food processing industries. Previous attempts include alphanumerical code or bar code [2], which are carried out without physical engagement directly to the fish individuals, thus prone to questionable reliability in food supply chain. Enplanting Radio-frequency identification (RFID) chips on the other hand faces the risk of potentially injuring the fish [2, 3]. Moreover, those techniques are in general expensive by involving time-consuming and labour-intensive process. On the contrary, a computer vision method to identify individuals would solve this problem by minimizing the impacts from invasive techniques. Traditional methods of identifying fish species are in general using shape and texture feature extraction [4–7]. Mehdi et al. [8] used Haar classifier to classify the shape features modelled by Principal Component Analysis (PCA). However, the main drawbacks of feature based approach come from its sensitivity to background noise, poor generalization ability, and difficulties of finding discriminative features, especially when the goal is to recognizing sub-ordinate object classes or species that are highly characterized by regions or territories.

Recent years have seen rapid advances of applying deep convolutional neural networks (CNNs) in the fields of computer vision and machine learning [9]. The integration of deep learning into marine science has achieved remarkable success, i.e., applying underwater video and acoustic surveillance to monitor or identify fish species [10]. Deep learning methods has been deployed for fish recognition competition in the Kaggle challenge [11]. However, to the best of our knowledge, very limited work has been conducted in fish individual recognition (IR), partly because the statistical approach has been dominant in addressing marine biodiversity and ecosystem, and also because the scale and diversity of ocean fish species makes individual recognition impractical. Nevertheless, it is foreseeable that learning individual features has already drawn drastic attention and requires robust while flexible methodology in aquaculture and food traceability.

We have observed that each individual pufferfish has a unique skin texture pattern, suggesting that such texture patterns may be utilized for identifying each individual pufferfish, analogically to the process of learning feature representations of individual human faces. In this work, we introduce a well engineered deep face recognition (deep FR) technology, to identify takifugu bimaculatus, one type of pufferfish that inhabits in East China sea. By training a similar architecture on images of fish, we enable accurate identification of each individual pufferfish without physical intervention. Previous attempt can be found in [17], where the authors have proposed an architecture based on FaceNet [15] to identify salmon. It is worth noting that our work does not simply adopt any state of the art face recognition framework. Instead it investigates elementary building blocks of the architecture and therefore gains a better insight of well engineered features.

In summary, we make the following main contributions: (i) we have proposed a system that focuses on individual recognition of fish specie, by introducing a well engineered face recognition technology; and remarkable result has been achieved through an ablation study; (ii) we have created a labeled dataset of takifugu bimaculatus fish that is publicly available; (iii) by introducing individual pufferfish identification and tracking we could also enable new research areas that require monitoring of individuals over time such as feeding behavior, detection of diseases and social behavior. FishIR can facilitate such research by offering a non-invasive and efficient approach for identifying pufferfish, thus contribute to the massive growth of smart farming and deep ocean fishery.

The rest of this paper is structured as follows. In Section 2, we outline the related work. The construction of dataset is presented in Section 3. In Section 4, we present a comparison and analysis on elementary building blocks that are of vital importance for overall performance. Following this, we present our approach FishIR, a complete framework to identify individual pufferfish. Section 5 presents conclusions and thoughts on future research directions for this work.

2 Background

Face recognition is a typical example of individual recognition. Inspired by the breakthrough work launched by DeepFace [13] and DeepID [14], research in face recognition [12] focus has shifted to deep learning based approaches, and the accuracy was dramatically boosted to above 99.80% in just three years [16]. The facial recognition process normally has four interrelated phases or steps that are essential components of a facial recognition system and depend on each other. First, an image is fed to an FR module and segmented, followed by image processing and alignment [34], handling intra-personal variations before training and testing, such as poses, illuminations, expressions, and occlusions. Then, the feature extractor is learned by a loss function when training, and it is utilized to extract features of faces when testing. Last but not least, a face matching algorithm is used to compute similarity scores of the features to determine the specific identity of faces [18].

2.1 Segmentation

The widely used segmentation models are in general grouped into several categories based on the underlying model architecture. We introduce several representative relevant models and evaluate some novel approaches in our study.

Fully Convolutional Networks Fully convolutional networks (FCNs), which include only convolutional layers, have been applied to a variety of segmentation problems, such as brain tumor segmentation [20], skin lesion segmentation [21], etc. In this work, we applied FCN-8, first proposed by Long et al. [19] for semantic image segmentation.

Multi-Scale and Pyramid Network Based Models Feature Pyramid Network (FPN) was first proposed by Lin et al. [22]. In their work, the inherent multi-scale, pyramidal hierarchy of deep CNNs was used to construct feature pyramids with marginal extra cost. The Pyramid Scene Parsing Network (PSPNet), a multi-scale network to better learn the global context representation of a scene, was proposed by Zhao et al. [23]. In this work, we have evaluated PSPNet as one of the benchmark methods for performance comparison.

Encoder-Decoder Based Models The encoder/decoder based model consists of an encoder using convolutional layers adopted from the VGG16 network and a corresponding deconvolutional network, that takes the feature vector as input and generates a map of pixel-wise class probabilities and predict segmentation masks. We evaluated a novel encoder-decoder based design, named SegNet [24], in this work.

R-CNN Based Models (for Object Instance Segmentation) He et al. [25] proposed Mask R-CNN, which efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Mask Scoring R-CNN [26], proposed a mask scoring strategy which calibrates the misalignment between mask quality and mask score, brings consistent and noticeable gain with different models, and outperforms the state-of-the-art Mask R-CNN. Both Mask R-CNN and Mask Scoring R-CNN are evaluated in this work.

Dilated Convolutional Models and DeepLab Family Dilated convolution introduces the dilation rate to convolutional layers, based on sparse convolution kernels to enlarge the perception field. Some of the most important works include the DeepLab family [27] [28] [29]. In 2018, Chen et al. [30] released DeepLabv3+, which is composed of a depthwise convolution and pointwise convolution operations, has obtained a 89.0% mIoU score on the 2012 PASCAL VOC challenge best pretrained on the COCO and the JFT datasets.

2.2 Loss Functions

Intuitively, face features are discriminative if their intra-class compactness and inter-class separability are well maximized. As pointed out by [32], the classical softmax loss lacks the power of feature discrimination. In face recognition, designing margin-based (e.g., angular, additive, additive angular margins) softmax loss functions plays an important role in learning discriminative features. Given an input feature vector x with its ground truth label y, below equation shows a re-formulated softmax loss for face recognition, where $w_k \in R^d$ is the k-th classier ($k \in 1, 2, ..., K$) and K is the number of classes. $x \in R^d$ denotes the feature belonging to the y-th class and d is the feature dimension. $\cos(\theta_{w_k,x})$

is the cosine similarity and $\theta_{w_k,x}$ is the angle between w_k and x. s is the scale parameter.

$$L = -\log \frac{e^{s\cos\left(\theta_{w_y,x}\right)}}{e^{s\cos\left(\theta_{w_y,x}\right)} + \sum_{k\neq y}^{K} e^{s\cos\left(\theta_{w_k,x}\right)}} \tag{1}$$

3 Dataset

Recently, the creation of large annotated databases has encouraged the development of highly discriminative, state-of-the-art deep learning face recognition, from which rich and compact representations of faces are learned and leveraged. Several benchmark datasets are released for researchers to verify their algorithms, such as PASCAL [35], MS COCO [36], and ILSVRC [37]. In this work, we have constructed a properly annotated pufferfish database to facilitate identification of individual pufferfish.

3.1 Data Collection and Splitting

We constructed our own dataset with labeled pictures of takifugu bimaculatus, one type of pufferfish that inhabits in East China sea, as the starting point of our dataset construction. We acquired the dataset by obtaining video clips and extracting frames from the video stream.

The implementation aspects are described as below:

1. Specify the photographing conditions and device for image acquisition. A light background with a clear contrast to the fish color is required to ensure that fish's outline is clearly visible. We used D65 light source with a color temperature of 6500K.
2. Collect fish characteristics by shooting from various angles. The video clip is taken from carefully chosen angles, i.e. 30 or 45 degrees, with the camera rotating for about 60 seconds. Around 200 images are extracted from each recorded video clip.
3. Categorize and annotate images according to the chosen naming convention.

After acquiring the data, we applied MS COCO-style [36] to create the dataset for segmentation task. Since segmentation task is relatively simple, we used 30 different takifugu bimaculatus fish with in total 1267 images as training dataset. LabelImg software was used to manually label and annotate images. For feature extration, we applied pair matching to organize and split the dataset. The dataset was split into verification and identification subsets with 126 and 20 fish included, respectively. The verification subset was further divided into training, verification and recognition datasets with a proportion of 8:1:1. In total, $20,793$ pictures of 142 takifugu bimaculatus individuals were collected for feature extration task, and each image was scaled to 1500×800.

3.2 Data Augmentation

Data augmentation increases the number of training samples by applying a set of transformations to the images, which typically include translation, reflection, rotation, warping, scaling, color space shifting, cropping, and projections onto principal components. Data augmentation has proven to improve the performance of the neural network models in terms of yielding faster convergence, decreasing the chance of over-fitting, and enhancing generalization, especially when learning from limited datasets. We applied several different augmentation methods to datasets both for segmentation and feature extraction tasks, increasing the number of images use for segmentation from 1267 to 7620, and for feature extraction from $20,793$ to $55,837$.

4 Experiments

We adopted synchronized stochastic gradient descend (SGD) training on NVIDIA GeForceGTX 2080 Ti GPUs based on Linux Ubuntu 16.04 LTS. Python and Pytorch [40] were applied for all the experiments.

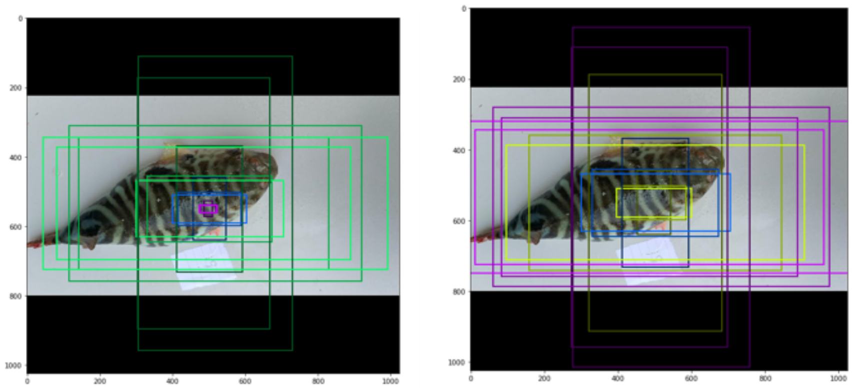
4.1 Segmentation Experiments

For segmentation experiments, 30 different takifugu bimaculatus fish with in total 1267 images were used as training dataset.

We performed an elaborated comparison of state-of-the-art image segmentation methods along with comprehensive ablation experiments. Mask R-CNN [25], Mask Scoring R-CNN [26] in instance segmentation category, and DeepLabV3+ [30], PSPNet [23], FCN8 [19], SegNet [24] were evaluated in this work. For Mask R-CNN and Mask Scoring R-CNN, learning rate was initialized to 0.0025. For the other models, learning rate was initialized to 0.01. The standard semantic segmentation metrics including MPA (Mean Pixel Accuracy) and MIoU (Mean Intersection over Union) [41] were used in our experiments as performance criteria.

Ablation Study First, we evaluated the proper anchor size for Mask R-CNN and Mask Scoring R-CNN. Anchor boxes provide a predefined set of bounding boxes of different sizes and ratios as a first reference for predicting object locations for the Regional Proposal Network (RPN). Inspired by YOLO [42], we apply K-means clustering to cluster the bounding boxes of objects in the training data to determine suitable anchor box sizes. These anchor boxes essentially serve as a guideline for our algorithm to look for objects of similar size and shape.

Based on the clustering results, we propose two sets of anchor boxes with different sizes, which are (32,128,256,512,600) and (128,256,512,600,680) respectively, and the aspect ratio is set to (0.5,1.8,2.5). Fig. 1 shows two different sets of anchor boxes. According to the training result after $10,000$ iterations, the first set of anchor boxes achieved slightly better results in terms of MPA and MIoU. Therefore, the result suggests us to apply the first set of anchor boxes of size



(a) (b)

Fig. 1. Two different sets of anchor boxes.

(32,128,256,512,600) and aspect ratio of (0.5,1.8,2.5) in Mask R-CNN and Mask Scoring R-CNN models.

We ran a number of ablation experiments to analyze the effect of various backbone networks. We applied ResNet50, ResNet101, MobileNet, ShuffleNet on Mask Scoring R-CNN, ResNet50, ResNet101 on Mask R-CNN, and ResNet101, MobileNet, DRN on DeepLabV3+ respectively. The quantitative and visual results are shown in Tab. 1 and Fig. 2 respectively. The order of pictures (from left to right, from top to bottom) in Fig. 2 follows the order of Tab. 1. The used backbone networks were pre-trained in advance.

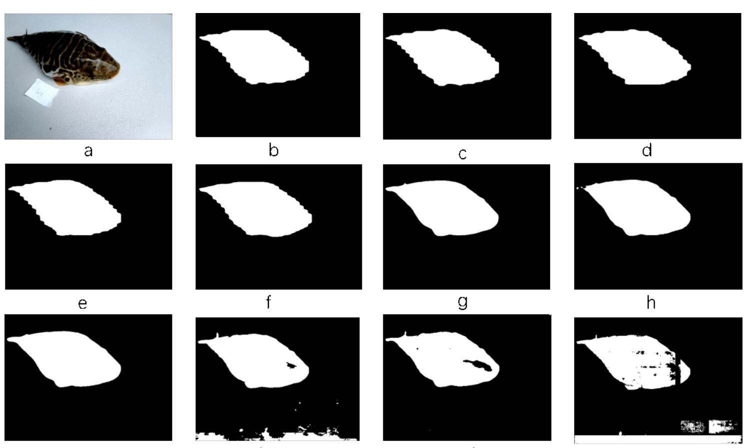


Fig. 2. Segmentation of various image segmentation models with various backbone networks.

The quantitative results show that comparing with Mask R-CNN, Mask Scoring R-CNN only demonstrates a marginal performance improvement on all backbone networks, most probably because single object segmentation task is not sensitive to a more accurate scoring function. On the other hand, Mask Scoring R-CNN has yielded larger model size (in terms of parameters) due to the extra prediction branch. Among semantic segmentation models, DeepLabV3+ based on ResNet101 backbone achieved the highest MPA and MIoU, 99.49% and 98.56%, respectively. Moreover, its average speed is 0.034s, which is efficient. On the contrary, although MobileNet achieved the fastest processing speed and smallest model size due to its compact model architecture, it sacrifices performance in terms of MPA and MIoU. PSPNet and FCN8 are not efficient in terms of speed and model size, although these two approaches can yield high MPA and MIoU. Based on the quantitative results, DeepLabV3+ achieved the best performance, and it was thus chosen as our default segmentation model.

Table 1. Experimental results of various image segmentation models with various backbone networks

Model	Backbone Network	MPA	MIoU	Time(s)	Model Size
Mask R-CNN	ResNet50	0.9884	0.9775	0.0659	335M
	ResNet101	0.9894	0.9772	0.0869	480M
	MobileNet	-	-	-	-
	ShuffleNet	-	-	-	-
Mask Scoring R-CNN	ResNet50	0.9883	0.9773	0.0886	459M
	ResNet101	0.9889	0.9774	0.0662	604M
	MobileNet	0.7518	0.5276	0.0442	295M
	ShuffleNet	0.9840	0.9699	0.0662	309M
DeepLabV3+	ResNet101	0.9949	0.9856	0.0344	453M
	MobileNet	0.9827	0.9761	0.0196	45M
	DRN	0.9886	0.9838	0.0211	311M
PSPNet	-	0.9927	0.9850	19.8835	393M
FCN8	-	0.9915	0.9822	11.9410	1.1G
SegNet	-	0.9872	0.9783	15.6391	125M

4.2 Feature Extraction Experiments

For feature extraction experiments, the dataset contains 4,158 pairs of sample images. As mentioned earlier, we use LFW-style dataset [39] for all the feature

Table 2. Experimental results of various backbone networks with various loss functions

Model	Loss function	ACC	Threshold	Time(s)	Model Size
ResNet50	AAML	0.9997	1.1124	0.2111	250M
	LMCL	0.9985	1.2450	0.2243	250M
	A-Softmax	0.9987	1.0900	0.2207	250M
ResNet101	AAML	0.9997	1.2974	0.3883	323M
	LMCL	0.9990	1.2025	0.4152	323M
IR_50	AAML	0.9992	1.1475	0.2726	251M
	LMCL	0.9963	1.4049	0.3066	251M
IRSE_50	AAML	0.9987	1.2450	0.3592	252M
	LMCL	0.9997	1.2050	0.3596	252M

extraction experiments. First, the testing dataset was divided into ten folds for cross-validation. The final results were obtained by averaging the accuracy of ten experiments. In each experiment, nine random folds were used as training folds, while the remaining fold was used as test fold to calculate the accuracy. Second, the Euclidean distance value on the pairs for each experiment were computed to determine whether a given pair is similar or not, based on a certain threshold s. The best threshold from training folds was used as the threshold for the testing fold. Accuracy was therefore defined as sum of TP and TN divided by total pairs, as shown in following equation.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Ablation Study For each experiment, we considered a unified architecture as the feature embedding backbone network with the same pre-processing and best settings and only change the loss function in order to better investigate performance of different loss functions.

We tested the following loss functions: (1) Angular Softmax Loss [31], which introduces an angular margin (A-Softmax) between the ground truth class and other classes to encourage larger inter-class variance; (2) the Large Margin Cosine Loss [33] that maximizes inter-class variance and minimize intra-class variance by reformulating the softmax loss as a cosine loss by L2 normalizing both features and weight vectors; and (3) Additive Angular Margin Loss [32] that adopts an additive angular margin loss, which has a clear geometric interpretation to compare their performance. Four different backbone networks were selected: ResNet50, ResNet101, IR_50 and $IRSE\_50$. IR_50 and $IRSE\_50$ only represent a slight modification of ResNet50.

The quantitative results of the ablation study are shown in Tab. 2, indicating that most state of the art convolutional neural networks are robust to perform feature extraction task. Note that the threshold in Tab. 2 is the aforementioned Euclidean distance value. In addition, we made the following observations: First, the ResNet50 backbone network using AAML loss function is the most efficient in terms of speed to learn representative features, whereas ResNet101 using LMCL loss function requires the longest time. Second, deeper and larger neural networks in general yield better performance, i.e., ResNet101 achieved high ACC performance with a value that exceeds 99.9% for different loss functions. Last while not least, the evaluation of different loss functions shows that AAML is outperformed than LMCL for most backbone networks, except for $IRSE\_50$.

Convolutional Kernels Observing the nature of pufferfish skin pattern, which has in general a rectangular shape, we are inspired by a new hypothesis that a rectangular convolutional kernel might learn feature representation in a more efficient and accurate manner than the traditional square sized convolutional kernel. In this study, we apply 7×3 rectangular convolutional kernel with padding 3×1 and stride equals to 2 at different stages of ResNet50 backbone network and evaluate the performance accordingly. The architecture of $ResNet50\_rc$ is shown in below Tab. 3. The result with 7×3 rectangular convolutional kernel applied at stages 1 to 4 of ResNet50 backbone network is shown in Tab. 3. Surprisingly, the result shows that only marginal performance gain or even performance degradation in terms of accuracy is achieved by applying rectangular convolutional kernel.

Table 3. The result with 7×3 rectangular convolutional kernel applied at stages 1 to 4 of ResNet50 backbone network

model	stage1	stage2	stage3	stage4	ACC
RecNet50_rc1	x	x	x	x	0.9992
RecNet50_rc2		x	x	x	0.9995
RecNet50_rc3			x	x	0.9997
RecNet50_rc4				x	0.9997

4.3 FishIR: a deep learning based pufferfish recognition architecture

In this Section, we present FishIR that utilizes the concept and building blocks of deep face recognition to identify pufferfish individuals. The setup of FishIR is the same as in the previous ablation study. Fig. 3 shows the architecture of the proposed FishIR.

We applied confusion matrix TPR (true positive rate) and FPR (false positive rate) to measure the model performance. As explained in the equations below, TPR is also known as sensitivity, recall or probability of detection, and FPR is

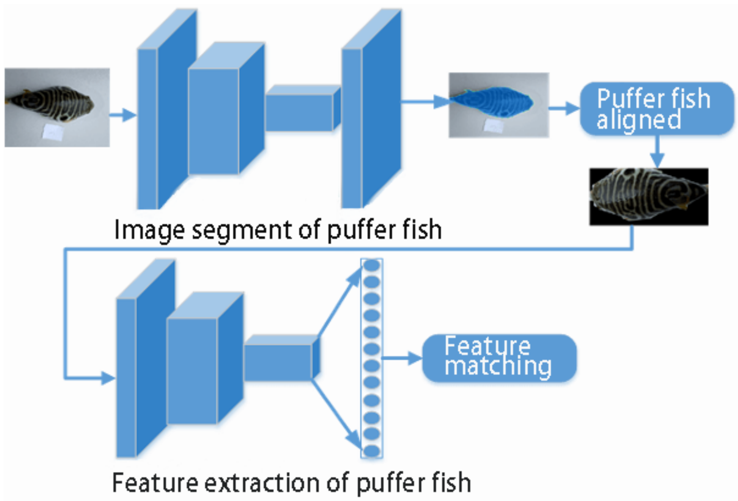


Fig. 3. Architecture of individual pufferfish recognition system.

also known as probability of false alarm in machine learning domain.

$$FPR = \frac{FP}{TN + FP}; TPR = \frac{TP}{TP + FN} \tag{3}$$

Based on the results of the aforementioned ablation study, we chose DeepLabV3+ as our segmentation model, and ResNet50 as feature extraction backbone network. We applied three different loss functions for evaluation throughout the experiments. Note that we can use the pre-trained model parameters for weight initialization without re-training.

We can draw various conclusions from the results. First, different network architectures yield similar accuracy, whereas speed can vary greatly for different model structures. Intuitively, extracted features are discriminative if their intra-class compactness and inter-class separability are well maximized. Therefore, in individual object recognition, efficiency of different loss functions can vary when applied to learn representative features. Among the loss functions we have tested, AAML performs the best for individual pufferfish recognition tasks. The main experimental results show that FishIR can successfully recognize pufferfish individuals based on their back skin patterns.

5 Conclusion

In this work, we present FishIR, a comprehensive method based on deep face recognition techniques, to identify takifugu bimaculatus, one type of pufferfish that inhabits in East China sea. Our experiments indicates that, by training a similar facial recognition architecture on images of pufferfish, we enable accurate identification of each individual without physical intervention. The encouraging

Table 4. Recognition results on the recognition dataset

Segmentation model	Loss function	TPR	FPR	Time(s)
MobileNet	AAML	0.9432	0.0010	0.4644
DRN	AAML	0.9492	0.0010	0.6689
ResNet101	AAML	0.9485	0.0011	0.7439
MobileNet	LMCL	0.9423	0.0067	0.4993
DRN	LMCL	0.9412	0.0065	0.8865
ResNet101	LMCL	0.9429	0.0066	0.7438
MobileNet	A-Softmax	0.9312	0.0053	0.4994
DRN	A-Softmax	0.9374	0.0049	0.8859
ResNet101	A-Softmax	0.9361	0.0044	0.4436

results will attract more research efforts to apply deep learning and facial recognition mechanisms into the field of food industry and fishery in the future.

As future work we would like to investigate the model's ability to recognize individuals from wild environment, i.e., video streams are captured in underwater habitat. We would also like to test if we can improve performance by employing other critical variants. Finally, we would like to test if we can generalize the concept by extending current recognition task from pufferfish to other species. Individual pufferfish identification and tracking could also enable new research areas that require monitoring of individuals over time such as feeding behavior, detection of diseases and social behavior. FishIR can facilitate such research through offering a non-invasive and efficient approach for identifying individuals, thus contribute to the massive growth of smart farming and fishery, as well as introducing potential benefits in food quality and safety inspection.

References

1. Lei Zhou, Chu Zhang, Fei Liu, Zhengjun Qiu , and Yong He. *Application of Deep Learning in Food: A Review.* Comprehensive Reviews in Food Science and Food Safety. vol.18, pp.1793-1811, 2019
2. A. Regattieri et al. *Traceability of food products: General framework and experimental evidence.* Journal of Food Engineering. vol.81, pp.347-356, 2007
3. Aleksander Børresen Eilertsen. *Identifikasjon av lakseindivider - biometri fase 1 (salmid).* Publikasjoner fra CRIStin - SINTEF Ocean (2017).
4. SHIAU Yi-Haur et al. *Fish observation, detection, recognition and verification in the real world.* In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition(IPCV), pp.1, 2012.
5. BLANC Katy et al. *Fish species recognition from video using SVM classifier.* In Proceedings of the 3rd ACM International Workshop on Multimedia Analysis for Ecological Data. pp.1-6. 2014.

6. ROVA Andrew et al. *One Fish, Two Fish, Butter fish, Trumpeter: Recognizing Fish in Underwater Video*. In Machine Vision Applications. pp.404-407. 2007.
7. ZHU Qiang et al. *Fast human detection using a cascade of histograms of oriented gradients*. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp.1491-1498. 2006.
8. R. Mehdi, et al. *Automated Fish Detection in Underwater Images Using Shape-Based Level Sets*. The Photogrammetric Record. vol 30, issue 149, pp. 46-62, 2015.
9. A. Voulodimos, et al. *Deep Learning for Computer Vision: A Brief Review*. Special issue: Recent Developments in Deep Learning for Engineering Applications. In Computational Intelligence and Neuroscience, vol.2018.
10. Md. Moniruzzaman, et al. *Deep Learning on Underwater Marine Object Detection: A Survey*. International Conference on Advanced Concepts for Intelligent Vision Systems, pp.150-160, 2017.
11. *https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring*. Kaggle Competition. The Nature Conservancy Fisheries Monitoring. 2017.
12. Shang-Hung Lin. *An Introduction to Face Recognition Technology*. Informing Science: The International Journal of an Emerging Transdiscipline. vol.3, pp.1-7, 2000.
13. Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. *Deepface: Closing the gap to human-level performance in face verification*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.1701–1708, 2014.
14. Yi Sun, et al. *Deep Learning Face Representation from Predicting 10,000 Classes*. IEEE Conference on Computer Vision and Pattern Recognition, 2014.
15. F. Schroff, et al. *Facenet: A unified embedding for face recognition and clustering*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.815–823, 2015.
16. Mei Wang, Weihong Deng. *Deep Face Recognition: A Survey*. Neurocomputing. vol. 429, pp 215-244, 2021.
17. B. M. Mathisen, et al. *FishNet: A Unified Embedding for Salmon Recognition*. The 9th International Conference on Prestigious Applications of Intelligent Systems, 2020.
18. Lucas D. and Helen N. *Facial Recognition Technology*. A Survey of Policy and Implementation Issues. Lancaster University, UK. Centre for the Study of Technology and Organization. 2009.
19. J. Long, E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3431-3440, 2015.
20. G. Wang, W. Li, S. Ourselin, and T. Vercauteren. *Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks*. International MICCAI Brainlesion Workshop. Springer, pp.178-190, 2017.
21. N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan. *Accurate iris segmentation in non-cooperative environments using fully convolutional networks*. 2016 IEEE International Conference on Biometrics (ICB), pp.1-8, 2016.
22. T. Y. Lin, et al. *Feature pyramid networks for object detection*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2117-2125, 2017.
23. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. *Pyramid scene parsing network*. Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2881-2890, 2017.
24. V. Badrinarayana, et al. *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*. IEEE transactions on pattern analysis and machine intelligence, vol.39, no.12, pp.2481-2495, 2017.
25. K. He, G. Gkioxari, P. Dollar, and R. Girshick. *Mask R-CNN*. Proceedings of the IEEE international conference on computer vision, pp.2961-2969, 2017.

26. Z Huang, et al. *Mask Scoring R-CNN.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019
27. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. *Semantic image segmentation with deep convolutional nets and fully connected crfs.* The International Conference on Learning Representations (ICLR), 2015.
28. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L.Yuille. *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.* IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834-848, 2017.
29. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. *Rethinking atrous convolution for semantic image segmentation.* arXiv: Computer Vision and Pattern Recognition, 2017.
30. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. *Encoder-decoder with atrous separable convolution for semantic image segmentation.* Proceedings of the European conference on computer vision (ECCV), pp. 801-818, 2018.
31. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. *Sphereface: Deep hypersphere embedding for face recognition.* Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 212-220, 2017.
32. Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface. *ArcFace: Additive angular margin loss for deep face recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690-4699, 2019.
33. H. Wang, et al. *CosFace: Large Margin Cosine Loss for Deep Face Recognition.* IEEE Conference on Computer Vision and Pattern Recognition, 2018.
34. Hu, G., et al. *When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition.* IEEE International Conference on Computer Vision Workshop (ICCVW), 2015.
35. M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A.Zisserman. *The PASCAL Visual Object Classes (VOC) Challenge.* International Journal of Computer Vision, Vol. 88, Issue 2, pp. 303–338, 2010. https://doi.org/10.1007/s11263-009-0275-4.
36. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. *Microsoft coco: Common objects in context.* Proceedings of the European Conference on Computer Vision. Springer-Verlag, pp.740-755, 2014.
37. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M. *Imagenet large scale visual recognition challenge.* International Journal of Computer Vision, vol.115, issue 3, pp.211-252, 2015.
38. Nech, A. and I. Kemelmacher. *Level Playing Field for Million Scale Face Recognition.* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
39. M. Parkhi, O., A. Vedaldi, and A. Zisserman. *Deep Face Recognition.* Proceedings of the British Machine Vision Conference (BMVC). 2015.
40. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. *Automatic differentiation in pytorch.* NIPS Workshop, 2017.
41. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez *A review on deep learning techniques applied to semantic segmentation.* arXiv preprint arXiv:1704.06857, 2017
42. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. *You only look once: Unified, real-time object detection.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016.
43. Lloyd, Stuart P. *Least squares quantization in PCM.* Information Theory, IEEE Transactions on 28.2 (1982): 129-137.