2022 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AGU. 22–25, 2022, XI'AN, CHINA

FISH FEEDING INTENSITY ASSESSMENT IN AQUACULTURE: A NEW AUDIO DATASET AFFIA3K AND A DEEP LEARNING ALGORITHM

Meng Cui[1,*], *Xubo Liu*[1,*], *Jinzheng Zhao*[1], *Jianyuan Sun*[1], *Guoping Lian*[2], *Tao Chen*[2], *Mark D. Plumbley*[1], *Daoliang Li*[3], *Wenwu Wang*[1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
[2]Department of Chemical and Process Engineering, University of Surrey, UK
[3]National Innovation Center for Digital Fishery, China Agricultural University, China

ABSTRACT

Fish feeding intensity assessment (FFIA) aims to evaluate the change of fish appetite during the feeding process, which is potentially useful in industrial aquaculture. Previous methods are mainly based on computer vision techniques. However, these methods are limited by water refraction and uneven illumination. In this paper, we introduce a new approach for FFIA using audio. We create a new audio dataset for FFIA, namely AFFIA3K, which contains 3000 labelled audio clips of different fish feeding intensity (*None*, *Weak*, *Medium*, *Strong*). We present a deep learning framework for FFIA, where the audio signal is first transformed into acoustic features, i.e. mel spectrogram, then a convolutional neural network (CNN)-based model is used to classify the fish feeding intensity. Experimental results show that our approach achieves an mAP of 0.74 on the test set of AFFIA3K, and considerably outperforms baseline systems. This indicates the potential of our proposed approach in aquaculture applications.

Index Terms— Fish feeding intensity assessment, Aquaculture, Audio classification, Deep learning

1. INTRODUCTION

The fish feeding process is one of the most important aspects in aquaculture. Insufficient feeding leads to slow growth rate of fish or even cannibalism behavior [1], while excessive feeding increases the cost of aquaculture [2]. Fish feeding intensity assessment (FFIA) aims to evaluate the intensity change of fish appetite during feeding procedure, which potentially improves the farming efficiency and saves the feed cost in industrial aquaculture [3, 4].

The early work on FFIA mainly focused on analyzing the hunger status of farmed fish based on human observation [5]. This method is highly dependent on the observer's experience, and manual observations increase time and labor costs. Recently, machine vision-based methods have been proposed for FFIA. These methods use fish action [6] to describe the intensities of fish appetite such as "*None*", "*Weak*", "*Medium*" and "*Strong*", as shown in Table 1. They adopt machine vision methods to capture the visual features (e.g., spatial distribution of fish school) and then perform classification of the feeding intensity in terms of fish feeding images. Convolutional Neural Network (CNN) based methods [7, 8] achieved better performance than other classic methods (such as an SVM or BPNN) on FFIA with the classification accuracy over 90% in similar datasets [9, 8, 7]. However, machine vision-based methods have some limitations in practical aquaculture, for example, visual measurement is prone to distortions by poor lighting conditions and water surface reflection noise [10, 11]. In addition, when the density of fish in the fish tank is high, it is difficult to capture the spatial distribution of fish from the visual measurements.

Acoustic measurement could be an alternative to the visual measurement for FFIA, as acoustic information could be used effectively in environments with low and uneven illumination conditions. Recently, acoustic sensors have been used for fish behaviour analysis. For example, acoustic acceleration transmitter tags are implanted in fish to measure the activity level of fish [12]. However, this method may cause irreversible harm to the fish, which may affect the experimental results. The dual-frequency recognition sonar (DIDSON) has been used to analyze the fish trajectory (e.g., velocity), but the sonar equipment is often too expensive to be applied in aquaculture factories [13, 10].

Table 1: Descriptions of fish feeding intensities (FFI).

FFI	**Description**
None	Fish do not respond to food
Weak	Fish eat only pellets that fall directly in front of them but not move to take food
Medium	Fish move to take the food, but return to their original positions
Strong	Fish move freely between food items and consume all the available food

* The first two authors contributed equally to this work.

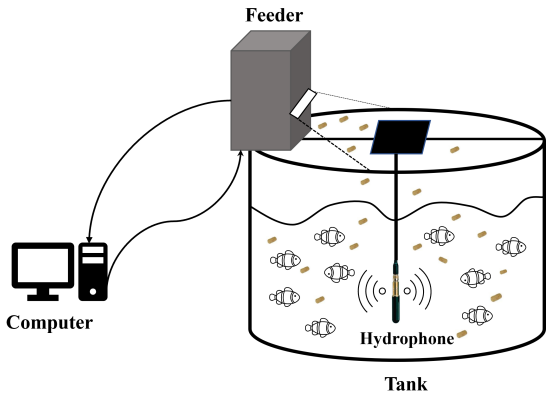

Fig. 1: Experimental systems for AFFIA3K data collection.

Different from the above work, in this paper, we consider the use of hydrophones for capturing fish feeding sound and predict the fish feeding intensity based on such sound data. Our idea is motivated by previous studies [14, 15] which show that the frequency of fish feeding sound generally belongs to a specific range, e.g., from 7-10 kHz and the amplitude of the variations in sound pressure level during feeding is 15 dB to 20 dB for turbots. The structure of fish feeding sound offers a great potential of using audio data for the assessment of fish feeding intensity. To predict the feeding intensity from sound, we use deep learning methods, which have been widely used in literature for a variety of tasks, including bioacoustic sound classification [16, 17, 18]. However, to our knowledge, deep learning has not been used for audio-based FFIA. Furthermore, there is no public fish feeding audio dataset, which poses challenges to audio-based FFIA.

In this work, we introduce AFFIA3K, a new audio dataset comprised of 3000 labelled fish feeding sound clips for FFIA. We present an audio-based deep learning framework for FFIA. In our proposed framework, the audio signal is first transformed into mel spectrogram features, which are used for the first time in the context of fish feeding. Then, they are fed into a CNN, followed by a classification layer, to obtain the probability of the feeding intensity category. We also perform extensive experiments to compare the proposed model with two baseline systems e.g., ResNet [19] and MobileNet [20]. The results show that our proposed method considerably outperforms the baseline methods in classification accuracy.

This paper is organized as follows: Section 2 introduces the AFFIA3K dataset. Section 3 describes the audio-based deep learning framework we proposed for FFIA. Experiments and results are shown in Section 4. Section 5 concludes the contribution of this paper and discusses the future direction.

2. DATASET

We use Oplegnathus punctatus (a kind of marine fish) as experimental subjects, which are farmed in a recirculating tank with the diameter in 3 meters and the depth in 0.75 meters, located in Yantai, Shandong Province, China. Each fish weighs about 150 g and the number of fish is 60. As shown in Fig. 1, a digital hydrophone is used to capture fish feeding audio. In this experiment, we used a high frequency hydrophone, LST-DH01, for audio data collection, which has an acquisition range of 10 kHz - 512 kHz. We used a sampling frequency of 256 kHz to collect the fish feeding sounds during the experiment. In the process of data collection, we followed the feeding rules in the real aquaculture production environment to ensure that the fish adapts to the laboratory environment as soon as possible and reduces the appetite loss caused by environmental changes. We feed the fish twice a day at 8 am and 5 pm. The audio data collection duration is 10 minutes, where the feeding begins in the third minute, and the feeding process lasts about 3 minutes.

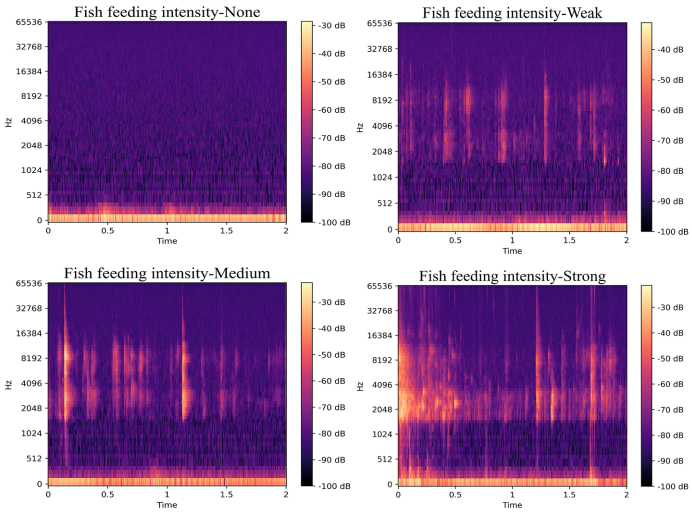

Fig. 2: Mel spectrogram visualizations of four different fish feeding intensity: "*None*", "*Weak*", "*Medium*" and "*Strong*".

During the experiment, a total of 25 files of audio data, each of four minutes, from the beginning of feeding (the time for fish to swim to food is short, so it is ignored), were captured using a hydrophone according to the experience of farmers and technicians. Under the guidance of fish-feeding technician and assistance of video recordings, we annotated the feeding audio data as "*Strong*", "*Medium*", "*Weak*" and "*None*". We further divide each one-minute audio clip into 30 segments with each segment in two seconds. As a result, 3000 two-second audio clips are obtained, with 750 audio clips for each category of fish feeding intensity. For each class of fish feeding intensity, we create the training and testing set by randomly choosing the audio clips, and in total, 700 clips are used for training and 50 clips for testing. Finally, we obtained 2800 audio clips for training and 200 audio clips for testing, referred to as AFFIA3K-Train and AFFIA3K-Test, respectively. To encourage further research, we have also released the AFFIA3K dataset and the pre-trained model at: https://github.com/FishMaster93/AFFIA3K.

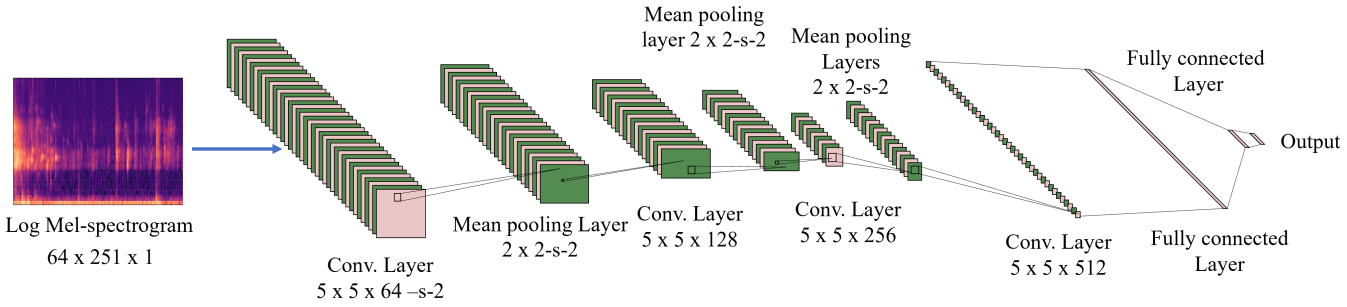

Fig. 3: Architecture of our proposed model for audio-based FFIA.

3. PROPOSED APPROACH

To predict fish feeding intensities, we proposed a deep learning framework which includes three parts, namely, mel spectrogram feature extraction, fish feeding intensity classification based on CNNs, and data augmentation, discussed next.

3.1. Feature Extraction

We use mel spectrogram as acoustic features, which has been widely investigated for audio classification [21]. Specifically, we re-sample audio signals to 128 kHz for reducing computational complexity, then use a Hanning window of 2048 samples as the filters to obtain Short-Time Fourier Transform (STFT), with a hop size of 1024 samples, and the mel filter banks with 64 bins. Therefore, for a 2-second audio signal, we have a mel spectrogram with the shape of 64×251. We visualize the mel spectrogram of different fish feeding intensity in Fig. 2. From this figure, we can observe that the mel spectrogram of the audio signals with different fish feeding intensity has significant difference in energy distribution, therefore, it can be used for predicting the fish feeding intensity.

3.2. Model

As one of the most common methods in deep learning, CNN has been widely used in bioacoustic cases, i.e., bird species classification and mosquito detection [17, 18]. CNN has the ability to analyze spatially invariant features with a small number of parameters, therefore, we choose the CNN as the classification backbone. In this experiment, we proposed a CNN6 model as the basic fish feeding intensity classification model. The CNN6 model structure is shown in Fig. 3. The CNN6 consists of 4 convolutional layers each containing several kernels with a kernel size of 5×5. The convolutional layer is used to extract features from log mel spectrogram. The pooling layer is used to reduce the dimensionality of the subsequent layers. We used the average pooling of size 2×2 after each convolutional block for downsampling. The ReLU activation function is used after each pooling layer to ensure the nonlinearity and to eliminate the gradient vanishing problem. For the output layer, we use softmax activation function for classification. The function scales a tensor into a range of $(0, 1)$ that all the values add up to 1 with each value corresponding to a probability for class category. The cross-entropy loss is used in the model as follows,

$$l = -\sum_{n=1}^{N} \left(y_n \cdot \ln f\left(x_n\right) + \left(1 - y_n\right) \cdot \ln\left(1 - f\left(x_n\right)\right)\right), \tag{1}$$

where N is the number of training clips in AFFIA3K, x_n is the waveform of an audio clip, n is the index of audio clips, and $f(x_n) \in [0, 1]^K$ is the output of a CNN model representing the presence probabilities of K classes of fish feeding intensity. The label of x_n is denoted as $y_n \in [0, 1]^K$.

3.3. Data Augmentation

Deep learning-based methods generally need a large amount of data. However, AFFIA3K data we created is of relatively small size, as collecting fish feeding intensity sound data is labour-intensive and costly. Daily feeding times are limited and data need to be manually screened. To make up for our lack of fish feeding intensity data, we adopted the method of data augmentation to expand our training samples. With SpecAugment, the spectrogram can be modified by warping it in the time direction, masking blocks of consecutive frequency channels, and masking blocks of time steps [22]. Frequency masking is applied such that f consecutive mel frequency bins $[f_0, f_0 + f]$ are masked, where f is chosen from 0 to a frequency mask parameter f' in terms of a uniform distribution, and f_0 is chosen from $[0, F - f]$, where F is the number of mel frequency bins [12]. In addition, SpecAugment prevents network from overfitting by providing intentionally distorted data to the network.

4. EXPERIMENTS AND RESULTS

4.1. Baseline Methods

We evaluate our proposed CNN6 model on AFFIA3K, and compare it with several baseline models, including ResNet22

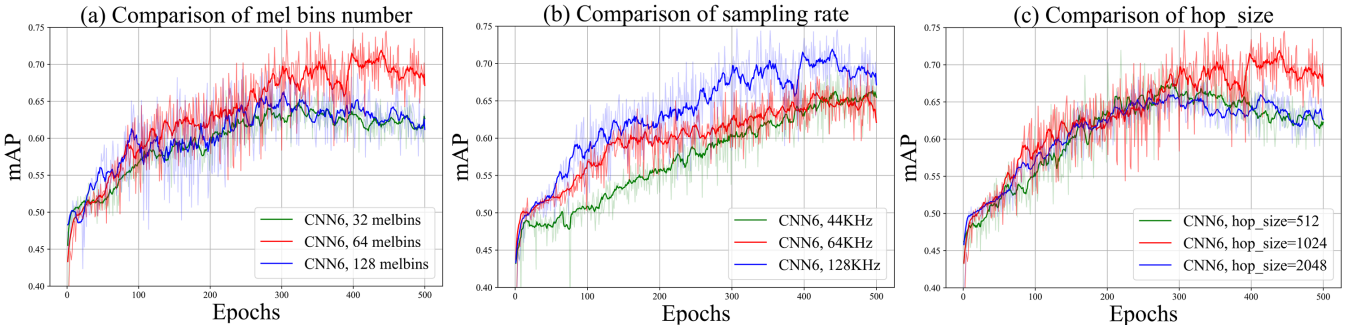

Fig. 4: Results of the CNN6 model on AFFIA3K. The transparent and solid curves are un-smoothed and smoothed mAP results, respectively. The three plots show the results of (a) comparison of different number of mel bins; (b) comparison of different sampling rate; (c) comparison of different hop size.

[23], ResNet38 [24], ResNet54 [25], MobileNetv1 [26], and MobileNetV2 [27]. ResNets (i.e. the conventional Residual Networks) is shown to offer better performance than shallower CNN networks in audio classification, which aims to solve the degeneration problem of the network, i.e. the optimization of the model becomes harder with the increase in the depth of the network. The design of residual blocks makes it easy to learn identity mapping. Even if excessive blocks are stacked, ResNet can make redundant blocks learn identity mapping without performance degradation [19]. Compared to CNNs and ResNets, MobileNets uses a new convolution method (depth-wise separable convolutions) to reduce the number of parameters and improve the speed of operation [20]. By comparing the different depths of network and parameters, we can find a model with a good trade-off in size and performance.

4.2. Training Procedure

We train the models on AFFIA3K-Train and evaluate them on AFFIA3K-Test. The Adam optimizer [28] with a learning rate of 0.001 is used for training the model. The batch size is set to 300 and the number of epochs is 500. The training and evaluation are performed on a Nvidia-RTX-3090-24GB GPU.

4.3. Evaluation Metrics

Mean average precision (mAP) and Accuarcy were used as the performance metrics in our evaluations. Accuracy refers to the number of correct predictions divided by total number of predictions. Compared with Accuracy which does not consider whether the predicted sample is positive or negative, mAP is better at reflecting the false positive and true positive results. We also used Precision, Recall and F1-Score as the metrics to further analyze the classification performance of the trained model in each class.

4.4. Results

4.4.1. Comparison with baseline methods

Table 2 shows the performance comparison of our proposed CNN6 model with other common CNN-based model. Our proposed CNN6 model achieves an mAP of 0.74 and Accuracy of 0.65, which outperforms other five CNN-based models. We also analyze the computational complexity of different models. Our proposed CNN6 model contains 4.5 million parameters which is much smaller than ResNet models. Compared with small models, such as MobileNetV1 and MobileNetV2 system, with only 4.2 million and 3.5 million parameters, respectively, our proposed model has a similar size, but offers performance that is 9%- 15% higher. Table 3 shows the results in Pr (Precision), Re (Recall) and F1 (F1-score) metrics. We can see that the classification precision for the "*Strong*" and "*None*" feeding intensity is higher, as compared with the classification precision for "*Weak*" and "*Medium*" feeding intensity. This could be due to the fact that when the fish is in a state of high hungriness, the feeding process proceeds very quickly. As a result, fish will quickly eat up all food in a short time, and this can lead to a sharp drop in their hunger intensity. In addition, the difference in the patterns corresponding to the "*Weak*" and "*Medium*" feeding intensity is relatively small, and thus difficult to capture in the original data collection process.

Table 2: The results of the different methods on the AFFIA3K.

Model	mAP	Accuracy	Parameters
ResNet22 [23]	68.75	60.00	62,603,460
ResNet38 [24]	65.55	49.50	72,711,620
ResNet54 [25]	61.96	53.50	103,246,532
MobileNetV1 [26]	65.06	49.50	4,260,228
MobileNetV2 [27]	59.39	54.00	3,539,268
CNN6 (Proposed)	**74.63**	**65.50**	**4,569,156**

4.4.2. Comparison of different mel bins

Fig. 4 (a) shows the performance of the CNN6 model trained with different mel bins. The CNN6 model can achieve an mAP of 0.74 with 64 mel bins, as compared to 0.65 with 32 mel bins and 0.64 with 128 mel bins. Our result shows that the CNN6 model can achieve better performance with the increase in the number of mel bins. However, the computation

Table 3: Results of each fish feeding intensities under different models on AFFIA3K-Test.

	ResNet22 [23]			ResNet38 [24]			ResNet54 [25]			MobileNetV1 [26]			MobileNetV2 [27]			CNN6 (Proposed)		
Class	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
None	0.76	0.44	0.56	0.51	**0.82**	0.63	0.63	0.44	0.52	0.56	0.64	0.60	0.51	0.52	0.51	**0.83**	0.60	**0.70**
Strong	0.83	0.80	**0.82**	**0.92**	0.68	0.78	0.86	0.76	0.81	0.62	**0.92**	0.71	0.84	0.74	0.79	0.90	0.74	0.81
Medium	0.54	0.40	0.46	0.50	0.44	0.47	0.41	0.64	0.50	0.42	0.22	0.39	0.42	0.42	0.42	**0.54**	**0.66**	**0.59**
Weak	0.42	**0.72**	0.53	**0.56**	0.44	0.49	0.35	0.30	0.32	0.44	0.38	0.41	0.44	0.48	0.46	0.50	0.62	**0.55**

complexity will also increase linearly with the number of mel bins. Considering the balance between computational complexity and system performance, we adopt an intermediate value of 64 mel bins to extract log mel-spectrogram.

4.4.3. Comparison of different sampling rate

Fig. 4 (b) shows the performance of CNN6 model at different sampling rate. The CNN6 model can achieve an mAP of 0.69 with the 44 kHz audio sampling rate, which is close to the result at 256 kHz sampling rate (0.69). The CNN6 model achieved the best performance at sampling rates of 64 KHz and 128 KHz, with mAP values of 0.71 and 0.74, respectively. This indicates that audio tagging of fish feeding intensity is more effective for the sampling rate in the range of 64 kHz to 128 kHz. In this paper, we adopt the sampling rate of 128 KHz when training our model.

4.4.4. Comparison of different hop size

We also compared the performance of the CNN6 model at different hop size. The hop size is the number of samples between adjacent frames. The smaller the hop size, the higher the timing resolution and the higher the computational cost. We investigate hop sizes of 512, 1024 and 2048 samples, which correspond to time domain resolutions of 4.00 ms, 8.00 ms and 16.00 ms between adjacent frames, respectively. As shown in Fig. 4 (c), the mAP achieved is 0.68, 0.74, 0.64, for hop size of 512, 1024 and 2048 samples, respectively. Therefore, we adopt the hop size of 1024 samples when training our model.

5. CONCLUSION

We have created a new audio dataset AFFIA3K for fish feeding intensity classification under the guidance of professional farmers. We also developed a CNN6 model for FFIA and compared it with several baseline models. The experimental results show that our proposed CNN6 model achieves an mAP up to 74.63%, which is about 5% higher than other baseline models on AFFIA3K. Our proposed model also provides an mAP that is 15%-20% higher than traditional support vector machine and clustering algorithms. Although the performance is lower than the fish feeding intensity classification based on video (mAP: 91%), the experimental results prove that the use of audio data to assess fish feeding intensity is feasible and could be used in practical aquaculture applications. However, due to the limited size of the AFFIA3K data, the classification performance for the "*Weak*" and "*Medium*" fish feeding intensity is limited, further improvement of the model (such as expanding data and improving the features) is required. We also found that the energy of fish feeding activities was concentrated in the middle to high frequency range and dispersed evenly. Mel spectrogram has a higher frequency resolution at low frequencies than high frequencies. In this paper, we use mel spectrogram as acoustic features, which achieved good results, however, there may be a better alternative for extracting acoustic features. In the next step, we will try to further improve the methods by extracting features from the high frequency regions using other methods such as band pass filters. In addition, audio and video based measurements could be fused to improve the accuracy of fish feeding intensity classification, e.g. via knowledge distillation and mutual learning.

6. ACKNOWLEDGMENT

This work was supported by the UK-China collaborative research program "Advancing digital precision aquaculture in China (ADPAC)" [UK BBSRC grant BB/S020896/1], and UK EPSRC Grant EP/T019751/1 "AI for Sound", and a PhD scholarship from the University of Surrey, and a Research Scholarship from the China Scholarship Council. Ethics approval for this study was obtained from the Welfare and Ethical Committee of China Agricultural University (Ref: AW30901202-5-1).

7. REFERENCES

[1] N. Saiz, M. Gomez-Boronat, N. De Pedro, M. J. Delgado, and E. Isorna, "The lack of light-dark and feeding-fasting cycles alters temporal events in the goldfish (Carassius auratus) stress axis," *Animals*, vol. 11, no. 3, p. 669, 2021.

[2] C. Chang, W. Fang, R.-C. Jao, C. Shyu, and I.-C. Liao, "Development of an intelligent feeding controller for indoor intensive culturing of eel," *Aquacultural Engineering*, vol. 32, no. 2, pp. 343–353, 2005.

[3] J. Zhao, W. J. Bao, F. D. Zhang, Z. Y. Ye, Y. Liu, M. W. Shen, and S. M. Zhu, "Assessing appetite of the swim-

ming fish based on spontaneous collective behaviors in a recirculating aquaculture system," *Aquacultural Engineering*, vol. 78, pp. 196–204, 2017.

[4] Y. Atoum, S. Srivastava, and X. Liu, "Automatic feeding control for dense aquaculture fish tanks," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1089–1093, 2014.

[5] O. Overli, C. Sorensen, and G. E. Nilsson, "Behavioral indicators of stress-coping style in rainbow trout: do males and females react differently to novelty?" *Physiology & Behavior*, vol. 87, no. 3, pp. 506–512, 2006.

[6] L. Yang, H. Yu, Y. Cheng, S. Mei, Y. Duan, D. Li, and Y. Chen, "A dual attention network based on EfficientNet-B2 for short-term fish school feeding behavior analysis in aquaculture," *Computers and Electronics in Agriculture*, vol. 187, pp. 106–316, 2021.

[7] C. Zhou, D. Xu, L. Chen, S. Zhang, C. Sun, X. Yang, and Y. Wang, "Evaluation of fish feeding intensity in aquaculture using a convolutional neural network and machine vision," *Aquaculture*, vol. 507, pp. 457–465, 2019.

[8] N. Ubina, S.-C. Cheng, C.-C. Chang, and H.-Y. Chen, "Evaluating fish feeding intensity in aquaculture with convolutional neural networks," *Aquacultural Engineering*, vol. 94, pp. 102–178, 2021.

[9] D. Wei, E. Bao, Y. Wen, S. Zhu, Z. Ye, and J. Zhao, "Behavioral spatial-temporal characteristics-based appetite assessment for fish school in recirculating aquaculture systems," *Aquaculture*, vol. 545, p. 737215, 2021.

[10] D. Li, Z. Wang, S. Wu, Z. Miao, L. Du, and Y. Duan, "Automatic recognition methods of fish feeding behavior in aquaculture: a review," *Aquaculture*, vol. 528, p. 735508, 2020.

[11] C. Zhou, K. Lin, D. Xu, L. Chen, Q. Guo, C. Sun, and X. Yang, "Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture," *Computers and Electronics in Agriculture*, vol. 146, pp. 114–124, 2018.

[12] J. Kolarevic, O. Aas-Hansen, A. Espmark, G. Baeverfjord, B. F. Terjesen, and B. Damsgard, "The use of acoustic acceleration transmitter tags for monitoring of atlantic salmon swimming activity in recirculating aquaculture systems (RAS)," *Aquacultural Engineering*, vol. 72, pp. 30–39, 2016.

[13] G. Rakowitz, M. Tuser, M. Riha, T. Juza, H. Balk, and J. Kubecka, "Use of high-frequency imaging sonar (DIDSON) to observe fish behaviour towards a surface trawl," *Fisheries Research*, vol. 123, pp. 37–48, 2012.

[14] J. Lagardere and R. Mallekh, "Feeding sounds of turbot (Scophthalmus maximus) and their potential use in the control of food supply in aquaculture: I. spectrum analysis of the feeding sounds," *Aquaculture*, vol. 189, no. 3-4, pp. 251–258, 2000.

[15] T. C. Tricas and K. S. Boyle, "Acoustic behaviors in hawaiian coral reef fish communities," *Marine Ecology Progress Series*, vol. 511, pp. 1–16, 2014.

[16] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *arXiv:2112.06725*, 2021.

[17] J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 141–145.

[18] I. Kiskin, D. Zilli, Y. Li, M. Sinka, K. Willis, and S. Roberts, "Bioacoustic detection with wavelet-conditioned convolutional neural networks," *Neural Computing and Applications*, vol. 32, no. 4, pp. 915–927, 2020.

[19] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv:1603.08029*, 2016.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[21] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[22] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv:1904.08779*, 2019.

[23] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *arXiv:2106.01621*, 2021.

[24] S. Okazaki, K. Quan, and T. Yoshinaga, "Ldslvision submissions to dcase: A multi-modal fusion approach for audio-visual scene classification enhanced by clip variants," DCASE Challenge, Tech. Rep., 2021.

[25] Q. Guo, Z. Yu, Y. Wu, D. Liang, H. Qin, and J. Yan, "Dynamic recursive neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5147–5156.

[26] S. Singh, A. Pankajakshan, and E. Benetos, "Audio tagging using linear noise modelling layer," 2019.

[27] K. Dong, C. Zhou, Y. Ruan, and Y. Li, "MobileNetV2 model for image classification," in *International Conference on Information Technology and Computer Application (ITCA)*. IEEE, 2020, pp. 476–480.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.