Journal of Network Intelligence ©2021 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **6**, Number **1**, February **2021**

A multi-stage data augmentation approach for imbalanced samples in image recognition

Ruo-Bin Wang* [1,2], Zhi-Wei An [1], Wei-Feng Wang [1],
Shuo Yin[1] , Lin Xu* [3,4]

[1]North China University of Technology, Beijing. 100043, China
[2]Key Laboratory of Nondestructive Testing, Fuqing Branch of Fujian Normal University, Fuqing 350300,China
[3]University of South Australia, Adelaide 5095, Australia
[4]Fuzhou Institute of Technology, Fuzhou 350506, China

Received November 2020; revised January 2021

Abstract. *Processing imbalanced samples has been a challenging and on-going issue since deep learning is employed as a kind popular technology in Computer Vision community. One of the effective solutions to this issue is applying data augmentation approaches before processing image recognition. In this paper, a multi-stage data augmentation approach is proposed*[1]*. In the first, the object in an image is outlined by the original image and the mask image. In the second, the outlined object is placed into different backgrounds with Alpha Blending to generate new samples. In the third, Gaussian Fusion is employed to make the foreground objects fused into the background smoothly. In model fitting, transfer learning is employed to achieve a higher accuracy. Experimental results have proved the approach's effectiveness, achieving a better performance compared with benchmark models.*

Keywords: Data augmentation, Alpha Blending, Gaussian Fusion, Image recognition, Transfer learning.

1. **Introduction.** Deep learning is widely used in many different domains [1–3]. And implementing image recognition with deep learning technology is an emerging topic in artificial intelligence [4]. However, deep learning algorithms require big volume of data with balanced categories for model fitting, it cannot perform well when processing imbalanced samples. And the results are prone to overfit, which leads to a bad performance. In fact, imbalanced samples are frequently involved when processing image recognition. In some cases, collecting data may be difficult and costly, which leads to inevitable imbalance. However, imbalanced samples will usually spoil the effectiveness of model fitting. Therefore, how to achieve a good recognition accuracy based on imbalanced samples has become a challenging and on-going research issue in the community of Computer Vision.

At present, the primary approaches to address the issues caused by imbalanced samples are data-oriented approaches, algorithm-oriented approaches, and hybrid approaches. Data-oriented approaches [5–7] improve the accuracy of classification by directly altering the training data to reduce the bias to majority classes. They usually copy, delete samples, or generate more samples manually to achieve sample balance. Algorithm-oriented approaches [8, 9] improve the accuracy of classification by modifying losses or weights for different categories, or by changing decision thresholds through reducing bias

[1] https://github.com/jkjkiiiii/IGD

to negative categories. Hybrid approaches [10, 11] refer to both the improvement of data level and algorithm level.

In addition, k-folds cross-validation is commonly used in model fitting image classification based on deep learning. However, in the dataset we used named Fish4-Knowledge (F4K), some images are collected from successive video frames with very similar image contents. Furthermore, there is a high possibility to allocate batches of similar images into the training set and the testing set respectively, which will probably cause overfitting. Therefore, in this paper, we tested on three different datasets for comparison, namely randomly allocated one, manually screened one and the one with data augmentation by the approach we proposed. Through comparing the first two allocations, we found that the fitted model cannot effectively extract the characteristics of fish, though sometimes the accuracy performed very high, and the model visualization also revealed it clearly. And then, the experiments also showed an improved effectiveness of classification when the samples were augmented.

Our key contributions can be summarized as follows: (1) The data augmentation approach we proposed can reduce the overfitting caused by batches of similar images in training set and test set, such as the ones extracted from successive video frames. (2) Through the proposed approach, we improve the recognition effectiveness and obtain a higher accuracy than using traditional data augmentation approaches. (3) By integrating transfer learning, the model is improved to achieve a better robustness on recognition.

This paper is organized as follows. Section 2 describes the related works of data augmentation. Section 3 details the materials and methods. The results and discussions are presented in Section 4, while Section 5 concludes our work.

2. **Related Work.**

To deal with the issue caused by imbalanced samples, approaches employed in previous researches can be divided into the following three categories: data-oriented approaches, algorithm-oriented approaches and hybrid approaches.

2.1. **Data-oriented Approaches.**

Data-oriented approaches mainly reduce the bias to the majority class by changing the training set, thereby reducing the degree of imbalance. Usually, these approaches realize data augmentation by copying, deleting, or manually generating samples. These traditional approaches are employed in situations of dealing with imbalanced samples, such as Random Over Sampling (ROS), which means randomly sampling from the majority classes and Random Under Sampling (RUS), which means randomly copying samples from the minority classes.

Hensman and Masko [5] proposed ROS to balance training data and improve the performance of classification when facing imbalanced image samples. However, because ROS only adds generated samples into minority classes, it will easily lead to overfitting in model training especially when features extracted from the generated data are closely similar to the original data. Lee et al. [6] used RUS and transfer learning to classify the highly imbalanced plankton dataset. The research adopted a two-stage learning strategy, that is, training the threshold by CNN firstly, and then fine-tuning to reduce data imbalance between classes. However, the samples leading to underfitting are the ones of small volume, therefore, excessive deletion of images will easily lead to training failure.

Buda et al. [7] compared the effects of RUS, ROS, and the two-stage learning approach on several imbalanced image datasets based on the MNIST, CIFAR-10 and ImageNet. It is proved that ROS performs well in the whole to solve category imbalance under the same experimental data, however, RUS usually performs poorly. Based on this conclusion, the data augmentation approach we proposed employs a multi-stage approach to increase the data to alleviate the problem of data imbalance between classes. But one of the challenges we discussed in this paper is that experimental datasets such as MNIST are simple in terms of complexity and data scale compared to real datasets. It deserves further exploration for the effectiveness of ROS when processing datasets with more diversity or rare categories.

In recent years, a novel idea has been employed to solve the problem caused by imbalanced samples. Generative Adversarial Network (GAN) is an approach that generating batches of images automatically, synthesizing more training data for minority classes that contain only a small amount of data to reduce the extent of data imbalance [12, 13]. However, because GAN needs a big volume of parameters and the debugger is complex, it is unfit for imbalanced samples. Besides, GAN needs sufficient data for the model training.

2.2. **Algorithm-oriented Approaches.** From the perspective of algorithm, solutions usually focused on the modification of the loss or weight of different classes, or on the reduction of bias from negative classes to change the threshold [14]. In cost-sensitive learning, penalties are assigned to each category through a cost matrix. Increasing the weight of a minority class is equivalent to increase its importance, thereby reducing the possibility that the class is mistakenly classified [15]. One of the challenges from cost-sensitive learning is the allocation of effective cost matrices. Generally, cost matrices are determined by the experience, they can also be determined by experts in the field.

Ding et al. [8] proved through experiments that using a deeper network can achieve better accuracy in the situation of imbalanced data and smaller networks. It shows that a larger network contains more local minima with good performance, and an acceptable solution can be found faster through gradient descent. Different experiments have shown that a deeper network achieves a higher accuracy on the feature extraction of complex images. However, a dataset with imbalanced samples means that the deeper the network, the more difficult to be trained. Besides, Lin et al. [9] proposed the loss function of Focal Loss, which effectively solved the extreme class imbalance issues often encountered in object detection, that is, the number of foreground samples greatly exceeds the number of negative background samples. By adjusting the learning speed of each class, the issue caused by imbalanced samples is improved. However, the experimental results show that the improvement of accuracy of loss function varies with different datasets. Therefore, more comparative experiments are needed to verify the effectiveness of Focal Loss.

2.3. **Hybrid Approaches.** Hybrid approaches refer to combining data-oriented and algorithm-oriented approaches in various ways and applying them to address the issues caused by imbalanced samples [16]. Usually, data sampling is employed to reduce noise and alleviate imbalance, and then a cost-sensitive learning threshold is employed to further reduce the bias against the majority classes.

Ando and Huang [10] introduced the Deep Over Sampling (DOS) approach, which learns embedding layers that can produce more discriminative features, and then supplements a few categories by oversampling in the deep feature space. Tested on different datasets, DOS proved its advantage in imbalanced classification. However, this approach is more complicated than the data-level and algorithm-level approaches, making it difficult to be applied in practical use. Dong et al. [11] proposed an end-to-end approach to process large-scale image classification, which combines hard sample mining with class rectification loss (CRL) to solve the problem of class imbalance. CRL regularization imposes an unbalanced adaptive learning mechanism that applies greater weights to more highly unbalanced tags, while reducing the weights of less unbalanced tags, thereby enhancing the ability to classify minority classes. The disadvantage of this approach is that it's performance on imbalanced samples is not good.

In summary, for the data-oriented approaches, ROS is prone to overfitting by simply generating samples; the disadvantage of RUS is that it cannot perform well on a small dataset, it is easy to fail on model training if the samples are over deleted. For the algorithm-oriented approaches, constructing a larger network can contain more local minima with good performance, and a deeper network performs better on feature extraction, but for imbalanced samples, it is difficult to extract effective features in some categories with a small amount of data. The advantage of modifying the loss function is that it can greatly improve the capacity of feature extraction without increasing the training time. However, because the optimal loss function is usually different for different datasets, it takes a lot of time to find the optimal loss function. For hybrid approaches, they take into account of the advantages of data-oriented and algorithm-oriented approaches. However, the hybrid approaches usually are complex, which makes them difficult to be applied in practice.

Therefore, we provide a multi-stage data augmentation approach to address the issues caused by imbalanced samples.

3. Procedures and Methods.

3.1. **Procedures.** The method we proposed is used to train on the target domain. In the first, the object in an image is abstract from the mask image, details are described in 3.2.1. Then, this object will be placed in different backgrounds with Alpha Blending to generate new samples, details in 3.2.2. Finally, Gaussian Fusion is employed to make the foreground objects fused into the background smoothly, which is detailed in 3.2.3. And transfer learning is employed to train the model in source domain.

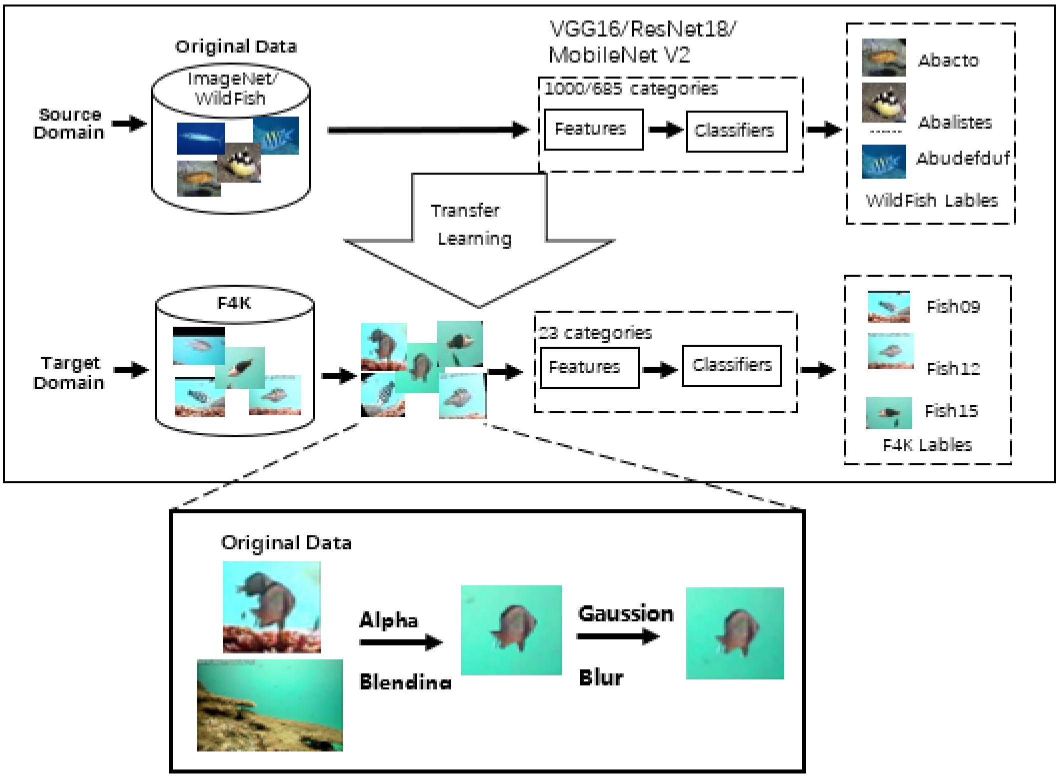


FIGURE 1. Experimental Framework

The source domain is trained with WildFish and ImageNet respectively in order to pre-train the parameters in VGG16, ResNet18 and MobileNet V2, which are the models we used to process image classification, and then the target domain is trained by F4K dataset. The more specific process is shown in 3.3.

3.2. **Data Augmentation.** There are three stages in the procedure of image preprocessing. Firstly, the object in the original image will be outlined by a mask image. Secondly, we employ Alpha blending method to replace the background to generate more images. Thirdly, we use Gaussian filtering to smooth the contour of the subject in an image.

3.2.1. *Object Extracted.* Image segmentation has been wildly used in different domains [17,18]. We use mask image to extract the main object in the image, which is calculated by the following formula:

$$p_{x,y} = \begin{cases} p_{x,y} & if \quad p_{x,y} > threshold \\ 0 & else \end{cases} \tag{1}$$

In formula (1), means one pixel with coordinate in the mask image, when $p_{x,y} > threshold$, the pixel should be kept; otherwise, the pixel will be deleted.

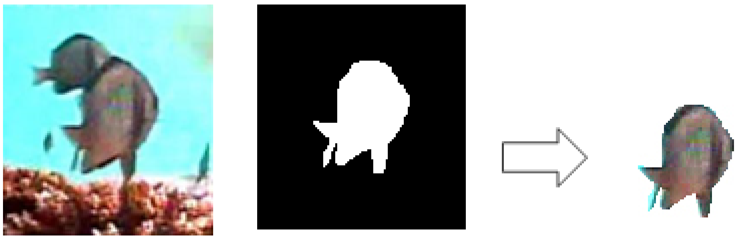


FIGURE 2. To exact subject by mask image

3.2.2. *Alpha Blending.* Alpha Blending [19] was employed to fuse the foreground into different backgrounds to achieve data augmentation. Transparency is usually regarded as the fourth channel of an image, but it can also be separated into a single image. This transparent mask is usually called an alpha mask. Figure 3 is the image processed by Alpha blending, which is calculated by the following formula:

$$OutputImage = \alpha * \ forgroundImage + (1-\alpha) * \ backgroundImage \tag{2}$$

In formula (2), α is the scale factor, when $\alpha = 0$, the output pixel value belongs to the background, when $\alpha = 1$, the output pixel value belongs to the foreground.

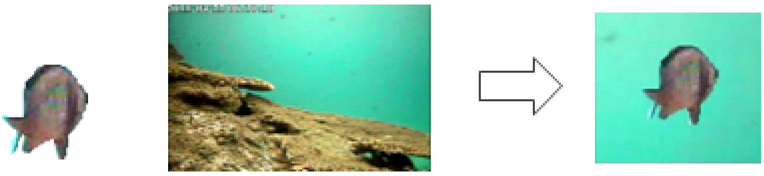


FIGURE 3. The image processed by Alpha Blending

3.2.3. *Gaussian Filtering.* To make the edges smooth, Gaussian filtering [20] was performed on the edges of the mask. Gaussian filtering is a linear smoothing filter suitable for eliminating Gaussian noise and is widely employed in the denoising process of images. Gaussian filtering links the image frequency domain processing with the time domain processing and is used as a low-pass filter to filter out low-frequency energy and therefore smooth images. Gaussian filtering is a process of weighted average of the entire image. The value of each pixel is obtained by weighted average of itself and other pixel values in the neighborhood. The formula is as follows:

$$G\left(x,y\right)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

In formula (3), (x, y) is the point coordinates, which can be considered as integers in image processing; σ is the standard deviation, which represents the degree of dispersion of the data. When σ is small, the center coefficient of the generated template is large, and the surrounding coefficients are small, so that the edge is not too sharp; on the contrary, when σ is large, the coefficients of the generated template have little difference, and the effect of smoothing will be better. Figure 4 shows the image smoothing procedure. The formula is as follows:

$$pic_{augument} = GaussianBlur(OutputImage(pic, \alpha), ksize) \tag{4}$$

In formula (4), α is the scale factor, means Gaussian kernel size.

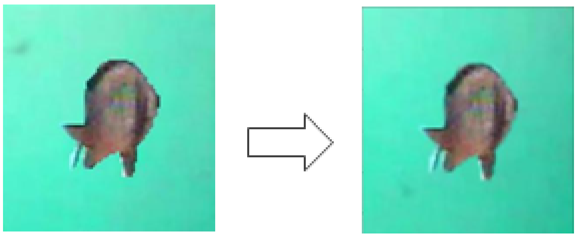


FIGURE 4. Image processed by Gaussian Fusion

3.3. **The Model based on Transfer Learning.** Imbalanced samples are commonly encountered in actual classification tasks, and usually, the number of samples for each class does not exceed 1000. F4K for fish recognition is a typical dataset with imbalanced samples. The samples for some categories are not sufficient for effective training. Therefore, we integrated transfer learning into our model to achieve a better performance. The reason we integrated transfer learning into the model is the previously learned knowledge can be reused to achieve a better performance.

The main assumption of most deep learning tasks is that training data and testing data must have the same distribution and in the same feature space [21]. A domain D is mainly composed of the feature space x and the marginal probability distribution $P(X)$, where $X = \{x_1, x_2, \ldots, x_n\} \in x$, for a given specific domain, a task T consists of two parts: label space Y and prediction function $f(\bullet)$, namely $T = Y, f(\bullet)$. Given a source domain D_s and a corresponding task T_s, and a target domain D_t and a corresponding task T_t. Under the circumstances $D_s \neq D_t, T_s \neq T_t$, the information learned in D_s, T_s is applied to D_t, T_t , so that the prediction function $f(\bullet)$ in T_t is improved. When the target domain sample is small, the pre-training model on the source domain data can be applied to the target domain through the transfer learning algorithm, and the new training after transfer learning can implement the target task better. Algorithm 1 [20] illustrate the process of transfer learning.

3.4. **Experiment Implementation.**

Algorithm 1 Transfer Learning Execution Process

Input: Source domain training set D_s, target domain training set D_t, source domain training number m target domain training number n
Output: The classification results of T_s and T_t of the source domain and target domain.

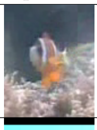
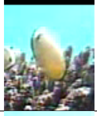
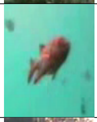
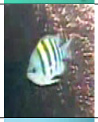
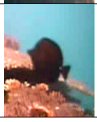
```
Random initialization of network parameters $M_s$ used in source domain
Repeat
for $i = 1$ to $m$ do
  Calculate $Loss$ through the results between predictions and real labels
  Reverse propagation updates the parameters of the model $M_s$
end for
Obtain model $M_s$ with source domain
Use model $M_s$ to initialize the parameters in model $M_t$
Repeat
for $i = 1$ to $n$ do
  Calculate $Loss$ through the results between predictions and real labels
  Reverse propagation updates the parameters of the weight $W_t$
end for
Complete the training of the final model $M_t$
```

3.4.1. *Dataset.* The datasets used in the experiment are ImageNet, WildFish, and Fish4-Knowledge(F4K). The WildFish dataset is collected by Zhuang et al. [22]. It contains images about 684 fish species, including 54,459 images, which are used as the source domain for transfer learning. The F4K dataset [23] is constructed by Taiwan Electric Power Company, Taiwan Institute of Oceanography and Kenting National Park. They share their image data collected in Taiwan's Nanwan and other places from October 2010 to September 2013. The F4K dataset contains images of 23 species of fish, with a total of 27,370 fish images, with the largest category containing 12,112 images, and the smallest one containing only 16 images, with a median of 181 and a mean of 1190. After screening, the largest category containing 3,713 pictures, the smallest one containing 13 images, the median is 155, the mean is 500.

3.4.2. *Data Preprocessing.* For the dataset of WildFish, since the number of images in each type of fish provided is roughly the same, only the F4K dataset has been preprocessed online. Firstly, we resized the images to the scale of 150*150 pixels, and then randomly flipped the images horizontally and vertically, rotated with 30° randomly, zoomed in and out, and adjusted the brightness and saturation. Finally, the 128*128 scaled images are randomly cut out as the training set.

Some of typical samples are listed in Table 1.

Table 1. F4K.Dataset and statistical information of data augmentation

Species No.	Sample image	Total number of images	Training set	Training set after data augmentation	Test set
04		4049	1495	1495	183
05		2534	1605	1605	231
06		190	165	495	25
07		450	420	1260	30
11		98	80	480	18
14		90	81	486	9

For data augmentation, the small sample is defined as the dataset with images less than 1000. For the dataset with images between 100 and 1000, the sample will be enlarged three times, for the one with images less than 100, it will be enlarged six times. Processed by data augmentation, the training set we employed contains the largest category with 3713 images, the smallest one with 78 images. For the training set, the median is 486, the average is 731.

In Table 1, data augmentation is not necessary because the number of images in category No.04 and No.05 is greater than 1000, while the number of images in No.06 and No.07 is three times higher because the number of images is between 100 and 1000, and the number of images in No.11 and No.14 is increased to six times because the number of pictures is less than 100.

3.4.3. *Model fitting.* The computer used in the experiment is equipped with Intel Core i7 CPU, GTX1660Ti GPU and 16GB RAM. The platform for model training is Pytorch CUDA V1.5.0.

In the training of transfer learning on WildFish, the model structure before the classification layers remains unchanged, only the number of fully connected layers is changed to ensure the output of 684 fish classifications.

When training the F4K dataset, we employed ImageNet and WildFish as the source domain respectively to implement transfer learning. In this research, we got three datasets through randomly selected, selected by sequence, and selected by sequence with data augmentation to train three different network models for further comparisons.

The loss function used in the three models is Cross Entropy Error Function, and the optimizer is Adam. The learning rate of 0.001 is used in the top 20 cycles to freeze the top n-1 layers parameters, and use the learning rate of 0.001 in the last 10 cycles to unfreeze the parameters of the bottom layer, and the batch size for each epoch is 32. The experiment results showed that the loss function and accuracy rate tend to stabilize when the training lasted to the 30 epochs.

Table 2. F-score values of VGG16, ResNet18 and MobileNet V2 models

Cross Validation results	Model		
Experiment Result	VGG16	ResNet19	MobileNet V2
Training set F-score	0.95	0.95	0.94
Test set F-score	0.90	0.88	0.88
Average training time(hour)	7	6	5.5

Table 3. Precision, Recall, F-score, Accuracy of the MobileNet V2 model on F4K

	ImageNet		WildFish					
Sample Type	P	R	F	A	P	R	F	A
Data Randomly Selected	0.96	0.96	0.96	0.99	0.95	0.97	0.95	0.99
Data Picked	0.76	0.77	0.77	0.93	0.76	0.76	0.77	0.93
Ours	0.79	0.80	0.82	0.94	0.81	0.82	0.81	0.94

For different datasets, the preprocessing methods employed are as follows: firstly, in order to reduce the computational complexity, all images were adjusted to 150*150 pixels, and then they were flipped, 30° rotated and cropped into 128*128 pixels randomly. During the training, 10-fold cross-validation is employed with 9/10 for training and 1/10 for testing.

3.4.4. *Assessments.* To evaluate the effectiveness of the approach, we use classification metrics, which are precision, recall, F-score and accuracy:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F_{\text{score}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Accuracy} = \frac{\mathrm{TP} + TN}{TP + FP + FN + TN} \tag{8}$$

In the formulas, TP stands for true positives, FP for false positives, FN for false negatives, and TN for true negatives. The values are all between 0 to 1, and the larger the data, the better the prediction effect of the model fitting.

4. **Experimental Results and Analysis.**

4.1. **Model Comparison.** We used VGG16 [24], ResNet19 [25] and MobileNet V2 [26] to train the transfer learning models based on WildFish. When facing extremely unbalanced samples, the indicator of Precision, Recall and Accuracy performed not well, so F-score is taken as the indicator for assessment. As shown in Table 2, there is no significant differences of F-scores between the training set and the test set. On the one hand, the performances of the three models are close to each other; on the other hand, comparing to other data set, F4K is a small one, therefore, we took MobileNetV2 as the model for transfer learning because of fewer parameters, which is helpful for preventing overfit.

The experimental results are shown in Table 3. In Table 3, The symbol of Precision, Recall, F-score and accuracy are P, R, F and A, respectively. The results show that when data is randomly selected, the F-score is even higher than the one on data picked. According to our experience, regardless of whether the images are randomly selected or assigned, the performance should be close to each other. However, the data randomly selected and data picked obtain a significate different performance. Therefore, we investigated the data samples further. We found some images in the F4K dataset are sequential frames with high similarity. Under the condition of random selection, the sequential images appear in both the training set and the test set, which will lead to an unreal high F-score. Based on the analysis above, we picked data to avoid getting images in consecutive frames, and therefore avoid similar images appearing in both the training set and the test set, and then a lower F-score was obtained.

The results show that the second experiment effectively divided the dataset so that the capacity of model classification reached a normal level. And the last experiment shows that the proposed approach is effective for imbalanced samples. And, the F-score is improved from 0.77 to 0.82 in ImageNet as source domain, 0.77 to 0.81 in WildFish as source domain. The approach also prevents the overfit caused by randomly selecting sequential images.

4.2. **Confusion Matrix.** In order to validate the effectiveness of data augmentation in the condition of data imbalance, we deliver the confusion matrix of different datasets. In the confusion matrix, the number of misclassified fish is marked by the cells with different darkness. The darker the cell, the more misclassifications in the corresponding fish species. Where the horizontal and ordinate coordinates of each graph represent the number of the 23 categories of fish corresponding to the classification. Figure 5 shows the results of training on the dataset with images selected randomly. The matrix reveals that the accuracy of the small categories (No.11, 14, 15, 17, 18, 19, 20, 21, 22, 23) is high, however, it is caused by overfitting. Figure 6 shows the real accuracy when we picked some images inconsecutively from sequential frames then put them into the training and testing dataset. Figure 7 shows the accuracy after the data augmentation processed by our approach. The matrix illustrates that the accuracy of some samples, especially the accuracy of small samples, has been improved. It shows that the data augmentation approach proposed is effective in preventing overfitting and improving accuracy.

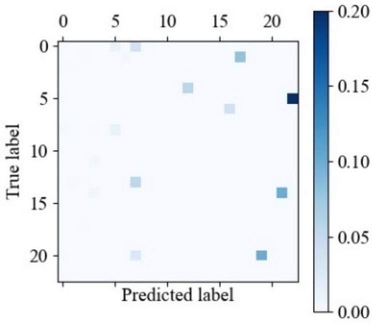


FIGURE 5
Random picked

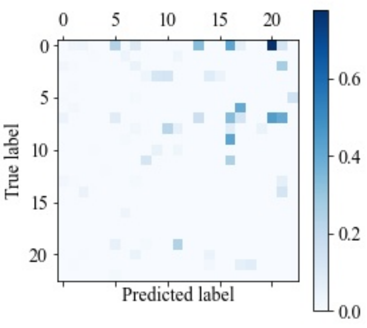


FIGURE 6
Inconsecutively picked

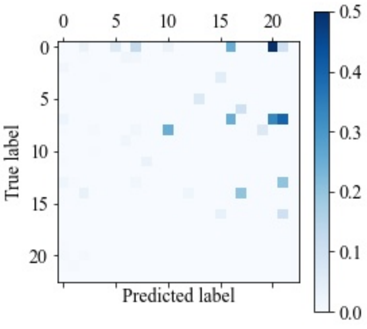


FIGURE 7
Data augmentation

4.3. **Model Visualization.** In order to further illustrate the effectiveness of the proposed data augmentation approach, model visualization technologies are employed to verify whether the model will focus on the areas of species-related characters but not unrelated areas, such as the background. Therefore, we employed grad-CAM [27] to calculate the fish species output and the gradient of the last convolutional layer relative to the input image, which provides a visualized explanation for the classification.

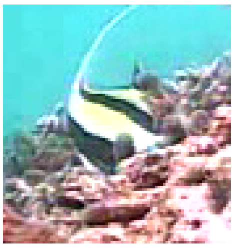


FIGURE 8
Original image

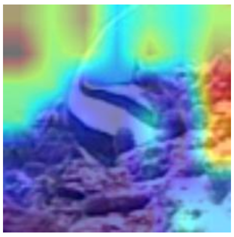


FIGURE 9
Results of randomly picked samples

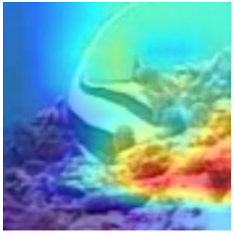


FIGURE 10
Results of inconsecutively picked samples

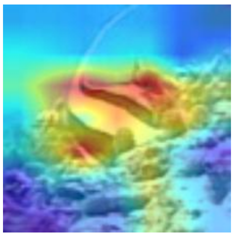


FIGURE 11
Results of data augmentation samples

Figure 8 to Figure 11 illustrate the corresponding gradient maps on the same image through grad-CAM on different training datasets. On a visualized picture, the red areas represent the ones focused by the model, and the blue areas represent the ones not focused. The focused areas (red areas) divided by grad-CAM indicate that models trained on different datasets will extract different features of fish species. In the case of using a sample consisted of randomly selected images very close to each other, shown as Figure 9, the focused areas are located in the background areas but not the ones of the fish itself, which shows that the model is inclined to ”remember” the images instead of extracting features from the similar images of a small sample. When we fitting the dataset consisted of selected images to avoid close similarity, as shown in Figure 10. The model still does not focus on the fish itself in the case of the small sample. And when we employed the multi-stage data augmentation approach, as shown in Figure 11, when we replaced different backgrounds to form more augmented images, the model can identify the fish from the background. Therefore, the proposed approach can improve the learning capability with better robustness of recognition.

5. **Conclusions.** In this paper, we propose a multi-stage approach for data augmentation to address the issues caused by imbalanced samples. Our approach is composed of three stages: Extract object, Alpha Blending, and Gaussian Filtering. After data augmentation, transfer learning is employed to acquire better robustness. Experiments were carried out on the image dataset of ImageNet, WildFish as source domain, and F4K as the target domain. The index concluding precision, recall, F-score and accuracy show that the approach we proposed improves the ability of the model recognition. The experimental results on confusion matrix and model visualization also revealed that the performance of fish recognition on imbalanced samples is improved significantly.

Since a potential limitation of this research is lacking further validation on different models, future research will explore the model integrated GAN to pursue further improvements.

Acknowledgment: This work is supported by Beijing College Students Innovation and Entrepreneurship Training Project 2020-Fish Detection Mobile System Based on Deep Learning and the Key Laboratory of Nondestructive Testing in Fuqing Branch of Fujian Normal University, Fujian Province under Grants No. # S2-KF1901.

REFERENCES

[1] E. K. Wang, S. P. Xu, C. M. Chen, and N. Kumar Neural Architecture Search Based Multiobjective Cognitive Automation System,*IEEE Systems Journal*,https://doi.org/10.1109/JSYST.2020.3002428 , 2020.

[2] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, Human Motion Recognition Based on SVM in VR Art Media Interaction Environment, *Human-centric Computing and Information Sciences*,9: 40, 2019.

[3] E. K. Wang, X. Zhang, F. Wang, T.-Y. Wu, and C.-M. Chen, Multilayer Dense Attention Model for Image Caption,*IEEE Access*, vol. 7, pp. 66358-66368, 2019.

[4] Y. Cui, M. Jia, T. Lin, Y. Song and S. Belongie, Class-Balanced Loss Based on Effective Number of Samples,*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,Long Beach, CA, USA, pp. 9260-9269, 2019.

[5] P. Hensman and D. Masko, The Impact of Imbalanced Training Data for Convolutional Neural Networks,*Ph.D. Thesis, KTH Royal Institute of Technology*, 2015.

[6] H. Lee, M. Park, and J. Kim, Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning,*IEEE International Conference on Image Processing (ICIP)*, pp. 3713–3717, 2016.

[7] M. Buda, A. Maki, and M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks,*Neural Networks*,vol. 106, pp. 249-259, 2018.

[8] W. Ding, D.-Y. Huang, Z. Chen, X. Yu, and W. Lin, Facial action recognition using very deep networks for highly imbalanced class distribution, *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, pp. 1368–1372, 2017.

[9] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, Focal Loss for Dense Object Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol. 42, no. 2, pp. 318-327, 2020.

[10] S. Ando, and C. Y. Huang, Deep Over-sampling Framework for Classifying Imbalanced Data,*Machine Learning and Knowledge Discovery in Databases*,vol 10534, Springer, pp. 770-785, 2017.

[11] Q. Dong, S. Gong, and X. Zhu, Imbalanced Deep Learning by Minority Class Incremental Rectification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*,vol. 41, no. 6, pp. 1367-1381, 2019.

[12] A. Ali-Gombe and E. Elyan, MFC-GAN: Class-imbalanced dataset classification using Multiple Fake Class Generative Adversarial Network,*Neurocomputing*,vol. 361, pp. 212–221, 2019.

[13] E. K. Wang, J. Yu, C.-M. Chen, S. Kumari, and J. J. P. C. Rodrigues, Data Augmentation for Internet of Things Dialog System,*Mobile Networks and Applications*,https://doi.org/10.1007/s11036-020-01638-9, 2020.

[14] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations,*arXiv:1707.03237*, 2017.

[15] J. M. Ren, W. Zeng, B. Yang, and R. Urtasun, Learning to Reweight Examples for Robust Deep Learning,*arXiv preprint arXiv:1803.09050*, 2018.

[16] J. B. Krawczyk, Learning from imbalanced data: open challenges and future directions,*Progress in Artificial Intelligence*,vol. 5, No. 4, pp. 221–232, 2016.

[17] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, A deep learning based medical image segmentation technique in Internet-of- Medical-Things domain, *Future Generation Computer Systems*,vol. 108, pp. 135-144, 2020.

[18] J. Wang, W. Song, X. Sun, L. Tang, and J,-H. Yeh, Annotation Method to Improve the Mapping Between Image Features and High Level Semantic Expression,*Journal of Network Intelligence*, vol. 5, No. 4, pp. 211-217, 2020.

[19] P. Baudisch, and C. Gutwin, Multiblending: Displaying overlapping windows simultaneously without the drawbacks of alpha blending,*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*,pp. 367-374, 2004.

[20] J. Steinbring, and U. D. Hanebeck, Progressive Gaussian Filtering Using Explicit Likelihoods,*Proceedings of the 17th International Conference on Information Fusion (Fusion 2014)*,Salamanca, pp. 1-8, 2014.

[21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, A Comprehensive Survey on Transfer Learning.*Proceedings of the IEEE*,vol. 109, no. 1, pp. 43-76, 2021.

[22] J. P. Q. Zhuang, Y. L. Wang, and Y. Qiao. WildFish: A Large Benchmark for Fish Recognition in the Wild,*2018 ACM international conference on Multimedia*,Seoul, pp. 1301-1309, 2018.

[23] J.B. J. Boom, P. X. Huang, J. He, and R. B. Fisher, Supporting Ground-Truth annotation of image datasets using clustering,*Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*,Tsukuba, pp. 1542-1545, 2012.

[24] K. Simonyan, and A. Zisserman, Very Deep Convolutional Networks for Largescale Image Recognition,*International Conference on Learning Representations(ICLR)*,San Diego, 2015.

[25] K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition,*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,Las Vegas, NV, pp. 770-778, 2016.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks,*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,Salt Lake City, UT, pp. 4510-4520, 2018.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,*International Journal of Computer*,pp. 336-359, 2020.